LEARNING OBSERVATION-ROBUST CONTENT FOR UNBALANCED INCOMPLETE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering is commonly studied under the assumption that every sample is observed in all views. In practice, however, some views are missing, and the amount of observed data can vary substantially across views, producing strong views with dense observations and weak views with sparse ones. Existing incomplete multi-view clustering methods mainly recover missing information through completion or alignment, while unbalanced methods adjust the contributions of strong and weak views. These strategies do not directly constrain the consensus representation of the same sample when its observed-view set changes. As a result, the representation may vary across different missing patterns and may still correlate with frequently occurring observation patterns. We propose COIN, a content–observation invariance framework for unbalanced incomplete multi-view clustering. COIN uses an observation factor to encode the missing pattern, while the content factor for clustering is computed from mask-gated view features. Nested content invariance encourages the same sample to maintain a consistent representation under coarser missing patterns, and rebalanced decorrelation reduces linear correlation between the content factor and the missing pattern after rare patterns are upweighted. Experiments on five unbalanced incomplete multi-view clustering benchmarks show that COIN consistently outperforms representative incomplete and unbalanced methods.

1 INTRODUCTION

Multi-view clustering (MVC) seeks a common partition from multiple descriptions of the same instances, where each view may correspond to a different descriptor, language, or sensing modality (Gao et al., 2015; Peng et al., 2019; Zhang et al., 2017; Nie et al., 2016; Zhan et al., 2018; Chen et al., 2020; Kang et al., 2020; Wang et al., 2018b; 2015; Li et al., 2015). When all views are available, consensus fusion and cross-view agreement can produce semantically meaningful clusters (Li et al., 2021; Xu et al., 2022b; Pan & Kang, 2021; Zhan et al., 2019; Nie et al., 2017a;b; Zhou et al., 2026). Real data rarely satisfy this ideal setting: sensors may fail, records may be corrupted, and some modalities may be too costly or intrusive to collect for every subject. Incomplete multi-view clustering (IMVC) therefore learns a shared representation from the subset of views that is actually observed (Wen et al., 2023b; Lin et al., 2021; Liu et al., 2021; Yang et al., 2023; Liu et al., 2019; Wen et al., 2019; Liu et al., 2020). The amount of available data may also differ substantially across views, producing densely observed strong views and sparsely observed weak views. Unbalanced incomplete multi-view clustering (UIMVC) studies this unequal-view setting (Sun et al., 2023; Jin et al., 2026). For example, one modality may be routinely collected for most samples, while a more costly or difficult-to-acquire modality is available only for a smaller subset, leading to markedly different observation rates across views (Jin et al., 2026).

Existing IMVC methods mainly recover or align missing information before constructing the final representation. Missing-view completion reconstructs unobserved measurements from kernel, graph, or tensor structure, or predicts them using deep autoencoders and prototype-level imputation (Wen et al., 2019; Liu et al., 2019; 2020; Wen et al., 2020a; 2021b; Xia et al., 2022a; Wen et al., 2020b;c; Lin et al., 2021; 2023; Xu et al., 2022a; Li et al., 2023; Jin et al., 2023; Wen et al., 2021a; Liu et al., 2021; Wang et al., 2022; Wen et al., 2023a; Yu et al., 2025). Distribution and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

semantic alignment methods, including imputation-free projections, reduce discrepancies between complete and incomplete samples without explicitly synthesizing raw views (Xu et al., 2023; 2024; Deng et al., 2024; Dai et al., 2025; Wang et al., 2026; Zhao et al., 2026). Other approaches transfer graph structure across partially observed views (Wen et al., 2020a; 2019; Zhao et al., 2026), while UIMVC-specific methods rebalance the influence of strong and weak views (Sun et al., 2023). These strategies determine how available or recovered views contribute to the representation, but do not directly require the representation of an individual sample to remain stable when its observed-view set changes. If one recorded view is removed, the remaining views form a nested subset of the original observations; we refer to the corresponding mask as a coarser missing pattern. The semantic identity of the sample is unchanged, yet its consensus representation may still shift substantially.

This issue is more consequential in UIMVC because missing different views is not equally informative. Frequently observed views appear in more training instances, whereas sparse views receive much less exposure. Removing a rarely observed but discriminative view can therefore induce a different representation shift from removing a dense view. Unequal availability also creates uneven frequencies of joint missing patterns, so a learner trained directly on the empirical distribution may become biased toward common observation conditions. The representation of a fixed instance should therefore remain stable when one observed view is removed, while the representation space should avoid being linearly organized around frequent missing patterns.

Based on these observations, we introduce COIN, a content–observation invariance framework for unbalanced incomplete multi-view clustering. COIN represents each sample with a mask-gated content factor for clustering and an observation factor that explicitly encodes the missing pattern. This parameterization does not imply statistical independence or full information-theoretic disentanglement. Nested content invariance constructs a coarser observation by dropping one recorded view, with scarce views sampled more often, and aligns its content with that of the original observation. Rebalanced decorrelation compensates for unequal missing-pattern frequencies before suppressing linear correlation between the content factor and the observation mask. The two constraints act at different levels: nested invariance controls how the same sample changes across related observation patterns, while decorrelation reduces the tendency of the representation space to follow dominant missing patterns.

The main contributions of this work are summarized as follows.

- We identify two representation-level issues not addressed by view-contribution rebalancing alone: instability of the consensus representation of the same instance across missing patterns, and residual correlation between that representation and the missing pattern under observation bias.
- We develop COIN, which combines a mask-gated content factor, an explicit observation factor, nested content invariance, and rebalanced decorrelation to reduce sensitivity to unbalanced observation patterns.
- Experiments on five unbalanced IMVC benchmarks verify the effectiveness of the proposed framework and the complementary roles of nested invariance and rebalanced decorrelation.

2 RELATED WORK

2.1 INCOMPLETE MULTI-VIEW CLUSTERING

Incomplete multi-view clustering aims to identify a common partition when only a subset of views is available for each sample (Wen et al., 2023b). Early approaches mainly rely on kernel completion, graph propagation, or tensor structure to recover information from the observed samples (Wen et al., 2019; Liu et al., 2019; 2020; Wen et al., 2020a; 2021b; Xia et al., 2022a). More recent deep methods reconstruct missing measurements, predict unavailable views, transfer neighborhood information across views, or align incomplete representations without explicitly filling the raw features (Peng et al., 2018; Wen et al., 2020c;b; Lin et al., 2021; 2023; Zhang et al., 2019b; Wen et al., 2021a; Liu et al., 2021; Wang et al., 2022; Wen et al., 2023a; Yu et al., 2025; Huang et al., 2019; Cheng et al., 2020; Xia et al., 2022b; Jin et al., 2023; Li et al., 2023; Dai et al., 2025; Xu et al., 2022a; Lu et al., 2024; Yang et al., 2023; Zeng et al., 2024; Wang et al., 2018a; 2021a; Zhang et al., 2019a; Wang et al., 2023; Xu et al., 2023; 2024; Deng et al., 2024; Wang et al., 2026; Zhao et al., 2026). Although these methods are effective at recovering a shared partition from partial views, the learned consensus

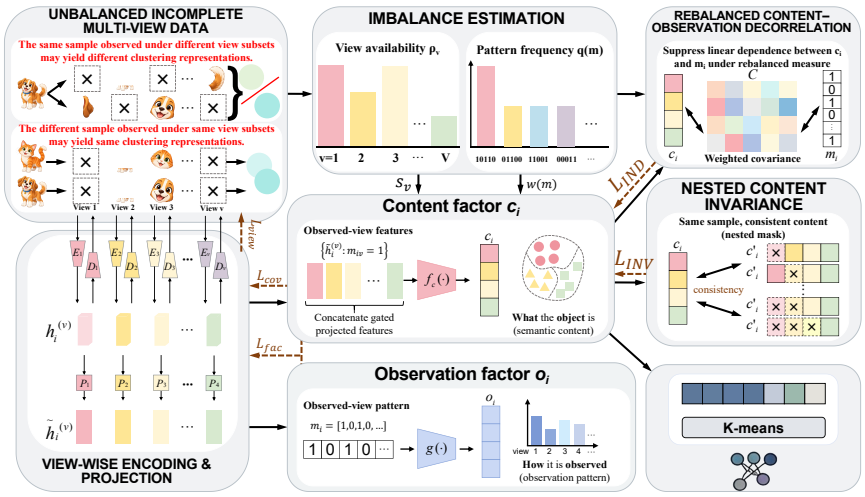


Figure 1: COIN. Mask-gated view features form a content factor, and the mask forms an observation factor. Nested invariance keeps the same sample consistent under coarser missing patterns; rebalanced decorrelation suppresses linear correlation with the missing pattern; clustering uses the content factor.

representation is still typically formed from the views that happen to be available for a sample. As a result, the representation of that sample may change when one of its recorded views is removed.

2.2 Unbalanced Incomplete Multi-View Clustering

Unbalanced incomplete multi-view clustering (UIMVC) further considers the case in which different views are observed at substantially different frequencies, producing densely observed strong views and sparsely observed weak views (Sun et al., 2023; Jin et al., 2026). Existing work has mainly handled this imbalance by changing how information is recovered or weighted. View-evolution models and balance-guided spectral clustering adjust the relative influence of strong and weak views (Sun et al., 2023), while CIMLN uses causal-invariant meta-learning to complete missing views under observation bias (Jin et al., 2026). BURG instead reconstructs unavailable views through neighborhood- and prototype-level distributional consistency, explicitly accounting for changes across missing patterns (Jin et al., 2025). These methods address unequal view contribution, missing-view recovery, or distribution shift across observation patterns, but they do not directly require the recorded consensus representation of the same sample to remain stable when its observed-view set becomes coarser. COIN focuses on this representation-level behavior rather than introducing another completion mechanism: it constrains the content representation of a fixed instance across nested missing patterns and reduces its linear correlation with the missing pattern after rare patterns are reweighted.

3 Methodology

3.1 Unbalanced Observations and Criteria

A dataset of unbalanced incomplete multi-view data comprises N samples and V views, $\mathcal{X} = \{\mathbf{X}^{(v)} \in \mathbb{R}^{N\times d_v}\}_{v=1}^{V}$. The indicator $\mathbf{M} \in \{0,1\}^{N\times V}$ satisfies $m_{iv} = 1$ if and only if view v of sample i is recorded. Write $\mathbf{m}_i$ for the i-th row and $\mathcal{O}_i = \{v : m_{iv} = 1\}$, with $|\mathcal{O}_i| \geq 1$. The learner receives $(\{\mathbf{x}_i^{(v)}\}_{v\in\mathcal{O}_i}, \mathbf{m}_i)$ and must partition the N samples into K clusters. Labels are not used to build masks or representations. Let $\mathbf{x}_i$ denote the complete measurement of instance i, which is not fully revealed, and write q for the empirical distribution of recorded masks.

The observation pattern is produced by an operator on the instance, not by a property of the instance.

Definition 1 (Observation operator). *For a mask $m \in \{0,1\}^V$, the observation operator is*

$$A_m(\mathbf{x}_i) = \big(\{m_v\, \mathbf{x}_i^{(v)}\}_{v=1}^V,\, m\big). \tag{1}$$

The learner observes only $A_{\mathbf{m}_i}(\mathbf{x}_i)$. Write $m' \prec m$ when $\mathrm{supp}(m') \subset \mathrm{supp}(m)$ and $m' \neq 0$. The nested family of sample i is $\mathcal{M}_i = \{m : \mathrm{supp}(m) \subseteq \mathcal{O}_i,\, m \neq 0\}$.

A nested mask of sample i is therefore any non-empty restriction of the recorded mask, equivalently a coarser missing pattern supported by $\mathcal{O}_i$.

Any fused map of $A_m(\mathbf{x})$ depends on both the instance and the operator, and frequent missing patterns can dominate the fused geometry. Full invariance $f_c(A_m(\mathbf{x})) = c(\mathbf{x})$ is not observable from recorded views; we therefore enforce nested invariance of the same sample and linear uncorrelation of content with the missing pattern after rare patterns are reweighted.

Imbalance estimation, as in Figure 1, computes view availability and pattern frequency,

$$\rho_v = \frac{1}{N}\sum_{i=1}^N m_{iv}, \qquad q(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\mathbf{m}_i = \mathbf{m}\}. \tag{2}$$

Let P_N be the empirical law of recorded pairs (X, M). The rebalanced law P_μ reweights P_N by $w(M)$; write μ for the induced mask marginal and Cov_μ for covariance under P_μ,

$$\mu(m) \propto w(m)\, q(m), \qquad w(m) = \min\big\{w_{\max},\, \max\big(1, (q(m)+\varepsilon)^{-\gamma}\big)\big\}, \tag{3}$$

a truncated inverse-frequency reweighting of recorded patterns. The weight w depends on masks alone. Let $\mathcal{I} = \{i : |\mathcal{O}_i| \geq 2\}$. The nested kernel $K(\cdot \mid i)$ of sample $i \in \mathcal{I}$ is supported exactly on one-view restrictions of $\mathbf{m}_i$, with masses given by Eq. (17). This eligibility condition is per sample and does not require $V \geq 3$: for a two-view mask $\mathbf{m}_i = [1,1]$, either $[1,0]$ or $[0,1]$ is a valid non-empty restriction. A sample with $|\mathcal{O}_i| = 1$ is excluded only from $\mathcal{L}_{\mathrm{INV}}$, because dropping its sole recorded view would leave an empty mask. The law π_μ on nested pairs is

$$\pi_\mu(i, m') \propto \mathbf{1}_{i\in\mathcal{I}}\, w(\mathbf{m}_i)\, K(m' \mid i). \tag{4}$$

Definition 2 (Content–observation criterion). *For a content map f_c,*

$$J_{\mathrm{INV}}(f_c) = \mathbb{E}_{(i,m')\sim\pi_\mu} \big\| f_c(A_{m'}(\mathbf{x}_i)) - f_c(A_{\mathbf{m}_i}(\mathbf{x}_i)) \big\|_2^2, \tag{5}$$

$$J_{\mathrm{IND}}(f_c) = \big\| \mathrm{Cov}_\mu\big(f_c(A_M(X)), M\big) \big\|_F^2, \tag{6}$$

$$\mathcal{J}_\mu(f_c) = J_{\mathrm{INV}}(f_c) + J_{\mathrm{IND}}(f_c). \tag{7}$$

Condition (C1) is $J_{\mathrm{INV}}(f_c) = 0$: nested invariance of a fixed instance in mean square under π_μ. Condition (C2) is $J_{\mathrm{IND}}(f_c) = 0$: zero linear cross-covariance between content and the recorded mask under P_μ. Thus $\mathcal{J}_\mu(f_c) = 0$ if and only if both hold. Condition (C2) is a linear Hilbert–Schmidt criterion, not statistical independence $C \perp M$.

Proposition 1. *Neither (C1) nor (C2) implies the other.*

Proposition 2. *The losses $\mathcal{L}_{\mathrm{INV}}$ and $\mathcal{L}_{\mathrm{IND}}$ below are surrogates for J_{INV} and J_{IND}, not empirical risks of those functionals. They are not interchangeable: enforcing only one leaves the other population term uncontrolled.*

Proofs are given in Appendix A. View-wise encoding and reconstruction, introduced next, fit recorded views through a pair $(\mathbf{c}, \mathbf{o})$ but do not enforce (C1) or (C2); those conditions are imposed by $\mathcal{J}_\mu$.

3.2 View-wise Encoding and Factorization

View-wise encoding and reconstruction assign an encoder–decoder pair to each view, which is standard in incomplete multi-view representation learning (Peng et al., 2018; Wen et al., 2020c; Lin et al., 2021; Wang et al., 2026). For a recorded view $v \in \mathcal{O}_i$,

$$\mathbf{h}_i^{(v)} = E_v(\mathbf{x}_i^{(v)}), \qquad \widehat{\mathbf{x}}_i^{(v)} = D_v(\mathbf{h}_i^{(v)}). \tag{8}$$

Observed-view features are gated by a view-specific projection P_v and concatenated to form the content factor, while the mask is mapped to the observation factor,

$$\widetilde{\mathbf{h}}_i^{(v)} = m_{iv}\, P_v(\mathbf{h}_i^{(v)}), \tag{9}$$

$$\mathbf{c}_i = f_c\Big(\big[\widetilde{\mathbf{h}}_i^{(1)}; \ldots; \widetilde{\mathbf{h}}_i^{(V)}\big]\Big), \tag{10}$$

$$\mathbf{o}_i = g(\mathbf{m}_i), \qquad \widehat{\mathbf{m}}_i = \sigma(W\mathbf{o}_i). \tag{11}$$

The content factor $\mathbf{c}_i$ is the representation used for clustering and is computed from mask-gated view features; it can therefore retain observation dependence at the input level. The observation factor $\mathbf{o}_i = g(\mathbf{m}_i)$ explicitly records the observation mask. The product in Eq. (9) prevents an unobserved view from being read as a measured zero, and $\mathbf{o}_i$ is discarded at clustering time. This is a factorization of the parameterization, not a guarantee of statistical independence or complete information-theoretic disentanglement. Let R_v reconstruct view v from $[\mathbf{c}_i; \mathbf{o}_i]$ as $\widetilde{\mathbf{x}}_i^{(v)} = R_v([\mathbf{c}_i; \mathbf{o}_i])$. With a reconstruction discrepancy ℓ, view-wise reconstruction, factor reconstruction, and a covariance penalty on $\mathbf{c}$ define

$$\mathcal{L}_{\text{view}} = \frac{1}{V}\sum_{v=1}^{V} \frac{\sum_i m_{iv}\, \ell(\widehat{\mathbf{x}}_i^{(v)}, \mathbf{x}_i^{(v)})}{\sum_i m_{iv}}, \tag{12}$$

$$\mathcal{L}_{\text{fac}} = \frac{1}{V}\sum_{v=1}^{V} \frac{\sum_i m_{iv}\, \ell(\widetilde{\mathbf{x}}_i^{(v)}, \mathbf{x}_i^{(v)})}{\sum_i m_{iv}} + \lambda_{\text{obs}}\, r(\mathbf{o}, \mathbf{m}), \tag{13}$$

$$\mathcal{L}_{\text{cov}} = \frac{1}{d}\sum_{j=1}^{d}(\Sigma_{jj} - 1)^2 + \frac{1}{d^2}\sum_{j\neq k}\Sigma_{jk}^2, \tag{14}$$

$$\mathcal{L}_{\text{REC}} = \mathcal{L}_{\text{view}} + \mathcal{L}_{\text{fac}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}}. \tag{15}$$

Here $r(\mathbf{o}, \mathbf{m})$ is the binary cross-entropy of the mask readout in Eq. (11). On a mini-batch of raw content factors with $B = |\mathcal{B}|$, $\mathcal{L}_{\text{cov}} = 0$ if $B < 2$; otherwise Σ in Eq. (14) is

$$\begin{aligned} \mu_j &= \frac{1}{B}\sum_{b\in\mathcal{B}} c_{bj}, \\ s_j &= \max\left\{\sqrt{\frac{1}{B}\sum_{b\in\mathcal{B}}(c_{bj} - \mu_j)^2},\ 10^{-4}\right\}, \\ z_{bj} &= \frac{c_{bj} - \mu_j}{s_j}, \qquad \Sigma = \frac{1}{\max(B-1, 1)}\sum_{b\in\mathcal{B}} \mathbf{z}_b\mathbf{z}_b^\top, \end{aligned} \tag{16}$$

where $\mathbf{z}_b = (z_{b1}, \ldots, z_{bd})^\top$. The batch is not ℓ_2-normalized and is not detached. The scale s_j uses denominator B and Σ uses $\max(B-1, 1)$, so the diagonal term is not an identity of the scaling (Appendix B.1). The coefficients λ_{cov} and λ_{obs} are trade-off parameters, and only recorded views enter the reconstruction sums.

3.3 Nested Content Invariance

The first term of $\mathcal{J}_\mu$ is nested content invariance of a fixed instance. This term targets instance-level consistency: a content map that fits frequent, richly observed patterns can still become inconsistent when the same sample is seen through a coarser missing pattern. We therefore estimate J_{INV} from one-view nested masks supported by recorded views.

For each $i \in \mathcal{I}$, a nested mask $\mathbf{m}_i'$ is drawn from $K(\cdot \mid i)$ by holding out a single recorded view v while retaining at least one other recorded view, with

$$\mathbb{P}(v \mid i) = \frac{m_{iv}s_v}{\sum_{u=1}^{V} m_{iu}s_u}, \qquad s_v = (\rho_v + \varepsilon)^{-\gamma_s}, \tag{17}$$

so scarce views are more likely to be held out. This is the kernel in Eq. (4); it never drops more than one view. The restricted content uses the same map f_c,

$$\mathbf{c}_i' = f_c\Big(\big[m_{i1}'P_1(\mathbf{h}_i^{(1)}); \ldots; m_{iV}'P_V(\mathbf{h}_i^{(V)})\big]\Big). \tag{18}$$

Algorithm 1 COIN for unbalanced IMVC

Require: Incomplete views $\mathcal{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{V}$, mask $\mathbf{M}$, cluster number K, and hyperparameters.
Ensure: A K-way partition of N samples.
1: Estimate $\rho_v, q(\mathbf{m})$ by Eq. (2), $w(\mathbf{m})$ by Eq. (3), and $K(\cdot \mid i)$ by Eq. (17).
2: Initialize $\{E_v, D_v\}$, f_c, and g.
3: **for** epoch $t = 1, \ldots, T$ **do**
4: Compute $(\mathbf{c}_i, \mathbf{o}_i)$ and $\mathcal{L}_{\mathrm{REC}}$ by Eqs. (8)–(15).
5: **if** $t \le T_w$ **then**
6: Minimize $\mathcal{L}_{\mathrm{REC}}$.
7: **else**
8: Sample $\mathbf{m}'_i \sim K(\cdot \mid i)$ for $i \in \mathcal{I}$, compute $\mathcal{L}_{\mathrm{INV}}$, $\mathcal{L}_{\mathrm{IND}}$, and minimize $\mathcal{L}_{\mathrm{REC}} + \alpha\mathcal{L}_{\mathrm{INV}} + \beta\mathcal{L}_{\mathrm{IND}}$.
9: **end if**
10: **end for**
11: Normalize $\{\mathbf{c}_i\}$ by Eq. (23) and apply K-means.

Write $\widehat{\mathbf{c}}_i = \mathbf{c}_i/\|\mathbf{c}_i\|_2$ and $\widehat{\mathbf{c}}'_i = \mathbf{c}'_i/\|\mathbf{c}'_i\|_2$. The nested invariance surrogate is

$$\mathcal{L}_{\mathrm{INV}} = \frac{\sum_{i\in\mathcal{I}} w(\mathbf{m}'_i)\|\widehat{\mathbf{c}}'_i - \mathrm{sg}(\widehat{\mathbf{c}}_i)\|_2^2}{\sum_{i\in\mathcal{I}} w(\mathbf{m}'_i)}. \tag{19}$$

Here $\mathrm{sg}(\cdot)$ treats its argument as constant, so the nested state is aligned to the recorded-state factor. The term is a surrogate of J_{INV} on one-view nested pairs of the same sample; it is not identical to J_{INV} (Appendix A).

3.4 Rebalanced Content–Observation Decorrelation

The second term of $\mathcal{J}_\mu$ is rebalanced content–observation decorrelation. The implemented criterion $\mathcal{L}_{\mathrm{IND}}$ suppresses linear content–observation correlation under the rebalanced measure, not statistical independence $C \perp M$. This term targets observation bias: even if content is consistent along nested masks, it may still correlate with $\mathbf{m}$ across the population, so clustering of $\mathbf{c}$ can recover dominant missing patterns. We therefore estimate J_{IND} under μ.

Let $\widetilde{w}_i = w(\mathbf{m}_i)/\sum_j w(\mathbf{m}_j)$ be the normalized weights from Eq. (3). The rebalanced, centered cross-covariance is the weighted covariance of Figure 1,

$$\widetilde{\mathbf{c}}_i = \mathbf{c}_i - \sum_j \widetilde{w}_j \mathbf{c}_j, \qquad \widetilde{\mathbf{m}}_i = \mathbf{m}_i - \sum_j \widetilde{w}_j \mathbf{m}_j, \tag{20}$$

$$\mathbf{C} = \sum_{i=1}^{N} \widetilde{w}_i\, \widetilde{\mathbf{c}}_i\, \widetilde{\mathbf{m}}_i^{\top}, \qquad \mathcal{L}_{\mathrm{IND}} = \frac{1}{dV}\|\mathbf{C}\|_F^2. \tag{21}$$

Eq. (21) suppresses linear dependence between $\mathbf{c}_i$ and $\mathbf{m}_i$ under the rebalanced measure; it does not imply $C \perp M$ or remove nonlinear observation information. If $\gamma = 0$, then $w \equiv 1$ and $\mathcal{L}_{\mathrm{IND}}$ reduces to ordinary empirical linear uncorrelation, which is dominated by frequent patterns; no class label enters $w(\mathbf{m})$.

By Propositions 1 and 2, neither condition controls the other, and $\mathcal{L}_{\mathrm{INV}}$ and $\mathcal{L}_{\mathrm{IND}}$ are surrogates for the two summands of $\mathcal{J}_\mu$. $\mathcal{L}_{\mathrm{INV}}$ without $\mathcal{L}_{\mathrm{IND}}$ can still leave consensus representations linearly separable by missing pattern. $\mathcal{L}_{\mathrm{IND}}$ without $\mathcal{L}_{\mathrm{INV}}$ can still allow a sample to move when a recorded view is dropped.

3.5 Objective and Inference

The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{REC}} + \alpha\mathcal{L}_{\mathrm{INV}} + \beta\mathcal{L}_{\mathrm{IND}}. \tag{22}$$

Here α and β are trade-off parameters, and the invariance and decorrelation terms are introduced after T_w epochs of reconstruction. The remaining coefficients $\{\lambda_{\mathrm{cov}}, \lambda_{\mathrm{obs}}, \gamma, \gamma_s, w_{\max}, T_w\}$ are structural hyperparameters. The 10^{-4} covariance scale floor is a fixed implementation constant, not a tuned hyperparameter. The training and inference procedure is summarized in Algorithm 1.

Table 1: Dataset sizes and view dimensions used by COIN.

Dataset	N	K	V	Input dimensions
MSRCV1	210	7	6	1302, 48, 512, 100, 256, 210
Washington	230	5	4	230, 1703, 230, 230
Synthetic3d	600	3	3	3, 3, 3
HandWritten	2000	10	5	240, 76, 216, 47, 64
TinyImageNet	6000	20	4	576, 944, 512, 640

Nested masks and the observation factor are discarded at clustering time. The clustering representation is the ℓ_2-normalized content factor,

$$\widehat{\mathbf{c}}_i = \frac{\mathbf{c}_i}{\|\mathbf{c}_i\|_2}, \tag{23}$$

and the partition is obtained by K-means on $\{\widehat{\mathbf{c}}_i\}_{i=1}^N$.

4 EXPERIMENTS

We evaluate COIN on five unbalanced incomplete multi-view datasets by ACC, NMI, and ARI.

4.1 DATASETS

We evaluate COIN on MSRCV1, Washington, Synthetic3d, HandWritten, and TinyImageNet; dataset statistics and view dimensions are listed in Table 8. Incomplete observations are generated using the view-wise quota protocol in Appendix C.1 with target mean missing rates $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Here, r is the average missing rate across views, with view-specific rates $r_v = \alpha_v r$. The generated masks are independent of features and labels and are kept fixed during training; realized view availabilities ρ_v are reported in Tables 4 and 5. Labels are used only for evaluation and the post-training coefficient selection in Appendix C.5.

4.2 COMPARED METHODS AND RESULTS

We compare COIN with twelve IMVC methods spanning imputation-free alignment (APADC), reconstruction and prototype completion (COMPLETER, CPSPAN, RecFormer, ProImp), imputation-free deep clustering (DIMVC, SMILE, SURE, SIMVC-SA), and unbalanced or missing-pattern methods (BGIMVSC, BURG, CIMLN) (Xu et al., 2023; Lin et al., 2021; Jin et al., 2023; Xu et al., 2022a; Liu et al., 2024; Li et al., 2023; Zeng et al., 2024; Yang et al., 2023; Sun et al., 2023; Wen et al., 2023c; Jin et al., 2025; 2026). Table 2 reports ACC, NMI, and ARI at unbalanced missing rates $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

COIN attains the highest ACC, NMI, and ARI in all 75 cells of Table 2. BURG and CIMLN are the closest recent baselines, but they do not rank first on any cell. At $r = 0.1$, the ACC margins over the second-best method are 2.43, 7.61, 0.67, 1.58, and 0.62 points on MSRCV1, Washington, Synthetic3d, HandWritten, and TinyImageNet; at $r = 0.5$ they are 26.81, 6.09, 10.45, 12.67, and 1.37 points. On TinyImageNet, ACC remains below 16% for every method, so the ranking is stable but the absolute gap is small.

4.3 ABLATION OF INVARIANCE AND DECORRELATION

Table 3 isolates nested content invariance and rebalanced decorrelation at $r = 0.5$ with reconstruction always enabled.

The full model attains the best accuracy on all four datasets. Removing nested invariance drops accuracy by 9.5 points on average, removing rebalanced decorrelation drops it by 2.4 points, and removing both drops it by 19.6 points, matching Propositions 1 and 2. The two modules target instance-level consistency and observation bias, respectively; Appendix E measures these roles on frozen content.

Table 2: Clustering performance (%) at unbalanced missing rates $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The best and second-best results are marked in **bold** and underline. Ours denotes COIN.

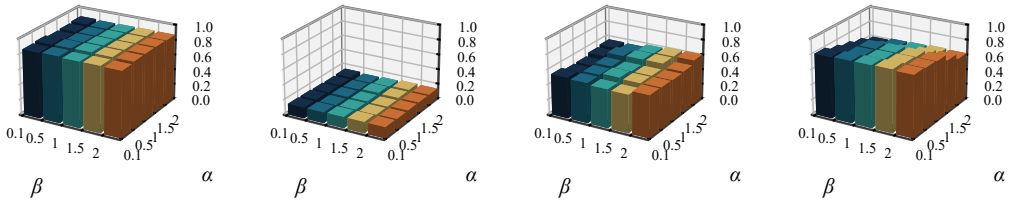
r	Method	MSRCV1			Washington			Synthetic3d			HandWritten			TinyImageNet		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
0.1	APADC	41.43	24.79	14.91	39.13	7.65	6.15	88.83	66.85	70.33	71.55	68.05	57.04	13.02	8.34	2.85
	COMPLETER	23.81	18.08	0.58	47.83	8.92	2.55	91.00	72.89	75.44	62.55	73.42	45.61	11.65	6.71	2.16
	CPSPAN	67.62	57.80	47.65	59.13	21.41	26.00	88.67	64.31	69.43	72.00	67.04	57.59	11.98	6.77	2.09
	DIMVC	52.38	40.87	31.63	49.13	19.33	18.12	84.17	54.89	59.58	70.65	63.86	53.42	13.60	8.28	2.68
	RecFormer	59.05	57.16	38.16	50.00	17.03	17.56	78.50	49.34	50.76	86.50	77.35	73.09	13.08	8.07	2.88
	ProImp	67.62	57.57	44.94	31.30	11.28	8.78	83.50	56.99	58.93	81.75	79.95	72.93	14.17	9.39	3.16
	SMILE	76.19	61.33	55.10	42.36	19.20	17.06	80.54	49.09	51.86	85.94	83.55	79.68	14.15	9.39	3.40
	SURE	54.29	41.82	27.34	31.74	8.57	0.07	88.67	67.44	68.95	69.60	68.50	58.42	13.67	8.98	3.02
	BGIMVSC	69.05	65.49	51.97	53.91	16.22	22.04	38.83	1.85	1.42	72.35	74.14	61.89	12.73	6.99	2.06
	SIMVC-SA	70.48	58.60	51.59	58.26	25.89	32.30	58.33	14.37	15.38	52.70	41.34	31.24	11.62	5.39	1.57
	BURG	87.24	79.20	74.34	60.61	24.85	30.85	89.00	67.62	69.44	92.71	86.15	84.74	12.55	8.66	2.45
	CIMLN	53.90	42.15	32.74	60.43	30.47	34.55	97.33	88.24	92.19	79.64	78.28	71.08	14.66	9.98	3.34
	Ours	**89.67**	**80.96**	**77.34**	**68.22**	**32.19**	**40.17**	**98.00**	**91.24**	**94.12**	**94.29**	**87.65**	**87.75**	**15.28**	**10.06**	**3.47**
0.2	APADC	35.71	24.41	5.40	46.52	7.59	5.42	85.17	58.18	62.17	74.80	70.83	61.95	13.27	8.18	2.66
	COMPLETER	22.38	12.55	0.30	44.35	6.36	-2.12	61.67	40.09	34.47	54.65	58.42	28.84	10.87	6.40	2.42
	CPSPAN	53.81	40.73	26.24	46.52	13.51	15.60	81.17	49.89	52.72	65.00	56.85	37.41	10.28	5.19	1.28
	DIMVC	28.57	20.01	9.47	45.22	21.47	15.13	62.50	40.67	38.52	55.55	53.53	34.11	10.33	5.80	1.55
	RecFormer	35.24	25.42	12.39	51.74	23.27	26.03	72.00	38.16	37.81	57.90	53.39	36.76	12.13	7.54	2.45
	ProImp	67.14	62.34	52.18	33.91	9.86	5.86	80.00	48.54	51.59	74.50	72.38	60.90	13.73	9.13	3.17
	SMILE	41.15	33.74	22.64	41.65	16.77	14.49	78.56	44.86	48.10	75.57	73.46	67.08	12.57	8.30	2.90
	SURE	37.14	22.65	11.45	31.74	4.70	0.49	78.67	43.61	46.70	65.75	64.80	53.94	12.65	8.18	2.47
	BGIMVSC	55.71	55.60	37.58	50.00	19.78	17.48	53.83	10.63	11.78	70.40	75.34	61.72	8.98	3.61	0.45
	SIMVC-SA	55.24	42.43	30.32	46.09	11.99	15.11	48.33	9.45	8.94	68.15	58.42	50.47	10.93	4.24	1.18
	BURG	76.76	68.84	60.96	59.09	23.77	28.54	82.75	56.30	54.08	72.13	70.02	57.79	11.18	6.96	1.85
	CIMLN	27.29	16.95	7.56	44.78	25.63	18.77	93.55	77.10	81.87	90.94	82.89	81.27	14.21	9.78	3.16
	Ours	**92.67**	**86.36**	**83.55**	**61.87**	**26.19**	**33.69**	**95.33**	**81.78**	**86.49**	**91.98**	**84.06**	**83.08**	**14.81**	**9.90**	**3.37**
0.3	APADC	36.67	19.71	6.27	46.52	6.43	9.04	91.00	69.48	74.77	78.00	76.15	67.99	13.25	7.60	2.58
	COMPLETER	19.52	9.96	0.17	50.43	11.00	5.96	63.33	53.75	45.28	66.35	67.71	48.00	10.78	5.65	2.06
	CPSPAN	39.05	26.57	11.41	44.78	19.69	15.37	56.83	16.43	11.61	52.10	51.04	30.49	9.60	4.70	1.22
	DIMVC	25.24	9.95	2.41	51.30	26.64	23.09	52.33	13.81	11.29	31.05	29.05	9.34	9.72	4.65	1.11
	RecFormer	32.86	21.08	6.51	36.52	15.06	8.87	78.17	54.28	49.50	53.95	48.89	30.88	11.12	6.75	2.02
	ProImp	55.24	44.50	30.25	34.35	8.65	6.61	57.17	29.61	26.35	68.75	70.14	58.13	12.85	9.20	2.92
	SMILE	51.20	40.17	30.32	32.28	10.39	6.78	58.46	28.46	28.37	77.71	75.28	69.52	13.14	8.15	2.87
	SURE	33.81	18.08	9.72	35.65	5.97	3.06	61.33	37.22	34.53	70.55	66.92	58.15	12.50	7.98	2.55
	BGIMVSC	52.38	52.17	36.32	45.22	16.21	12.90	39.33	3.06	2.33	73.05	76.33	65.51	10.58	5.44	1.51
	SIMVC-SA	62.86	49.23	39.93	34.35	10.42	6.79	47.83	6.99	7.50	63.50	48.29	40.25	10.92	4.40	1.21
	BURG	51.67	41.45	24.91	58.70	26.41	28.94	77.33	49.05	47.12	53.20	50.10	31.04	9.94	5.34	1.26
	CIMLN	20.48	3.96	-1.06	44.87	28.74	21.49	91.83	72.96	77.54	89.91	82.13	79.22	13.87	9.71	3.17
	Ours	**87.14**	**77.85**	**71.75**	**63.61**	**33.99**	**37.33**	**93.67**	**77.72**	**81.95**	**93.65**	**86.74**	**86.45**	**14.92**	**9.90**	**3.34**
0.4	APADC	36.19	21.34	3.80	42.17	3.54	-0.20	88.00	63.12	68.07	74.60	67.78	56.94	11.40	6.35	1.97
	COMPLETER	20.48	12.11	0.11	48.26	5.94	2.07	85.50	59.71	62.18	52.15	53.92	27.19	10.20	5.32	1.74
	CPSPAN	39.52	29.94	12.28	42.61	20.50	13.21	66.50	37.03	33.82	32.75	30.92	13.38	8.63	3.18	0.62
	DIMVC	23.33	5.97	0.62	49.13	22.45	17.24	63.17	29.05	27.90	30.60	29.29	11.09	10.18	5.22	1.33
	RecFormer	29.05	18.23	6.07	36.09	10.35	7.71	59.17	34.26	31.64	18.80	12.79	4.40	11.15	6.69	1.94
	ProImp	61.90	47.50	36.81	30.87	10.39	5.25	53.50	26.40	19.78	82.45	77.26	72.54	12.40	8.05	2.39
	SMILE	27.27	13.77	6.33	38.52	17.56	10.00	66.91	33.24	34.81	78.44	71.66	66.95	12.09	7.26	2.42
	SURE	36.67	18.55	9.35	29.13	4.48	-0.08	58.83	36.18	29.29	42.75	38.05	25.80	12.13	7.64	2.38
	BGIMVSC	60.00	59.65	45.25	46.96	17.25	12.46	45.00	3.68	3.69	69.50	74.85	62.98	11.32	5.95	1.61
	SIMVC-SA	44.76	28.79	17.49	49.57	11.89	17.46	41.00	6.97	6.43	49.70	38.74	27.99	9.78	3.31	0.81
	BURG	33.00	21.05	7.36	57.91	22.13	26.17	66.50	38.39	36.40	31.65	26.71	14.91	9.69	5.35	1.23
	CIMLN	21.95	4.45	-0.33	42.35	16.74	15.10	61.33	43.55	38.05	75.14	68.86	61.04	12.54	8.46	2.56
	Ours	**76.24**	**68.64**	**59.47**	**67.17**	**28.06**	**35.14**	**94.67**	**80.36**	**84.84**	**90.77**	**81.88**	**80.72**	**14.78**	**9.60**	**3.22**
0.5	APADC	31.90	15.29	3.69	43.91	3.79	3.74	78.17	43.39	46.44	74.05	68.87	59.46	12.10	6.98	2.16
	COMPLETER	21.90	13.22	0.32	38.26	4.57	-1.96	40.00	16.15	3.90	42.30	49.55	29.62	10.68	5.73	1.66
	CPSPAN	30.48	19.15	4.11	33.91	20.09	5.68	50.50	12.33	7.07	46.50	39.51	19.37	8.52	3.76	0.62
	DIMVC	22.86	5.96	0.37	41.30	17.08	9.94	49.33	9.49	8.31	32.10	26.91	8.17	9.87	4.79	1.12
	RecFormer	36.67	25.76	12.50	28.70	7.19	2.77	57.83	28.72	12.51	13.15	1.00	0.05	11.00	6.70	1.77
	ProImp	41.43	32.81	19.98	42.61	14.84	14.75	74.17	42.55	42.38	77.95	73.05	66.29	12.32	8.67	2.82
	SMILE	34.83	16.38	8.25	37.94	10.18	8.42	65.78	30.59	32.11	75.91	69.66	63.68	11.93	7.49	2.46
	SURE	28.10	10.25	2.63	41.74	7.77	3.35	51.00	18.65	16.67	48.40	37.51	30.01	11.22	6.82	2.00
	BGIMVSC	52.86	44.36	27.73	41.74	18.70	14.31	42.50	3.51	3.78	62.65	71.33	56.48	10.43	5.87	1.40
	SIMVC-SA	33.81	21.03	9.30	38.26	12.25	9.76	39.17	2.66	2.34	42.00	34.58	24.85	8.98	3.30	0.74
	BURG	25.33	8.91	1.34	52.48	14.55	18.35	56.00	19.74	17.62	21.60	11.32	3.59	9.09	4.58	0.86
	CIMLN	24.38	5.73	1.06	26.48	3.44	-0.42	62.87	26.59	27.87	44.04	49.02	31.22	11.24	7.01	1.88
	Ours	**79.67**	**70.20**	**58.15**	**58.57**	**22.94**	**27.60**	**88.62**	**65.40**	**68.64**	**90.62**	**81.53**	**80.36**	**13.69**	**9.03**	**2.84**

4.4 Sensitivity Analysis

Figure 2 reports the joint sensitivity to α and β, with both coefficients varied over $\{0.1, 0.5, 1.0, 1.5, 2.0\}$ at $r = 0.5$. Synthetic3d, TinyImageNet, and Washington remain relatively stable over a broad range of coefficient combinations, with no isolated setting dominating the surrounding region. This indicates that the gains of COIN on these datasets do not rely on a narrowly

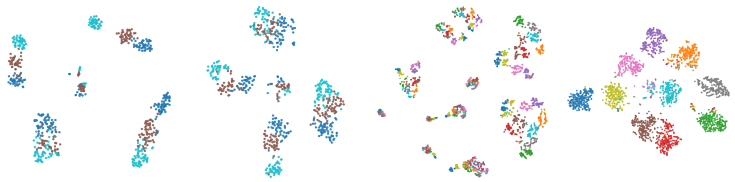
Table 3: Ablation study at unbalanced missing rate $r = 0.5$ on MSRCV1, Synthetic3d, HandWritten, and TinyImageNet. Best results are in **bold**.

			MSRCV1			Synthetic3d			HandWritten			TinyImageNet		
$\mathcal{L}_{\text{REC}}$	$\mathcal{L}_{\text{INV}}$	$\mathcal{L}_{\text{IND}}$	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
✓			69.86	58.58	48.29	36.17	0.47	0.10	75.89	69.18	59.22	12.30	7.45	2.20
✓	✓		75.43	67.37	57.86	87.37	63.52	65.70	87.54	77.74	74.26	12.71	8.90	2.57
✓		✓	74.24	66.39	56.61	58.85	27.31	20.73	88.62	77.72	76.58	12.86	7.65	2.34
✓	✓	✓	**79.67**	**70.20**	**58.15**	**88.62**	**65.40**	**68.64**	**90.62**	**81.53**	**80.36**	**13.69**	**9.03**	**2.84**



(a) Synthetic3d (b) TinyImageNet (c) Washington (d) MSRCV1

Figure 2: Sensitivity analysis of hyperparameters α and β on four datasets.



(a) Synthetic3d raw (b) Synthetic3d learned (c) HandWritten raw (d) HandWritten learned

Figure 3: t-SNE visualization of raw and learned features on Synthetic3d and HandWritten.

selected balance between the invariance and decorrelation terms. MSRCV1 shows a more noticeable response to the two coefficients, suggesting that the relative strength of the two constraints matters more on this dataset. Even in this case, the useful region extends across multiple settings rather than being confined to a single parameter pair.

4.5 Visualization

Figure 3 visualizes the raw features and the learned COIN representations on Synthetic3d and HandWritten at $r = 0.5$. The raw representations contain substantial overlap between several groups, particularly in regions where samples from different clusters are intermingled. After training, the content representations form more compact and visually separated groups on both datasets. The change is especially clear on HandWritten, where the learned space exhibits more distinct cluster structure than the raw input space.

5 Conclusion

We introduced COIN for unbalanced incomplete multi-view clustering. COIN combines a mask-gated content representation with nested invariance and rebalanced decorrelation to reduce sensitivity to changing and imbalanced observation patterns. Experiments on five benchmarks show that both constraints improve clustering beyond reconstruction alone. Nested invariance is applied only to samples with at least two observed views, ensuring that the restricted observation remains non-empty.

AI USE STATEMENT

Generative AI tools were used for manuscript editing, LaTeX formatting, and implementation assistance. They were not used to generate experimental numbers or to replace the evaluation protocol. The authors reviewed all AI-assisted text, code, and figures, and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies clustering on public multi-view benchmarks and does not involve human subjects or sensitive personal data.

REPRODUCIBILITY STATEMENT

The datasets and experimental protocol are described in Section 4.1 and Appendix C.1. The model objectives and training procedure are given in Section 3 and Algorithm 1, with proofs, implementation details, and parameter settings provided in the appendices. Anonymous source code is included in the supplementary material.

REFERENCES

Min-Sheng Chen, Ling Huang, Chang-Dong Wang, and Dong Huang. Multi-view clustering in latent embedding space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3513–3520, 2020.

Jiajia Cheng, Qianqian Wang, Zhiqiang Tao, Dihong Xie, and Quanxue Gao. Multi-view attribute graph convolution networks for clustering. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pp. 2973–2979, 2020.

Yuzhuo Dai, Jiaqi Jin, Zhibin Dong, Siwei Wang, Xinwang Liu, En Zhu, Xihong Yang, Xinbiao Gan, and Yu Feng. Imputation-free and alignment-free: Incomplete multi-view clustering driven by consensus semantic learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Shijie Deng, Jie Wen, Chengliang Liu, Ke Yan, Gehui Xu, and Yong Xu. Projective incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8): 10539–10551, 2024.

Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):913–927, 2022.

Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4238–4246, 2015.

Shudong Huang, Zhao Kang, Ivor W. Tsang, and Zenglin Xu. Auto-weighted multi-view clustering via kernelized graph learning. *Pattern Recognition*, 88:174–184, 2019.

Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11600–11609, 2023.

Jiaqi Jin, Siwei Wang, Zhibin Dong, Xihong Yang, Xinwang Liu, En Zhu, and Kunlun He. Deep incomplete multi-view clustering with distribution dual-consistency recovery guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1016–1026, 2025.

Jiaqi Jin, Siwei Wang, Taichun Zhou, Zhibin Dong, Siqi Wang, Miaomiao Li, Xinwang Liu, and En Zhu. Alleviating observation bias via causal-invariant meta-learning for unbalanced incomplete multi-view clustering. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Zhao Kang, Xue Zhao, Xi Peng, Hongmin Zhu, Joey Tianyi Zhou, Wenwen Chen, and Zenglin Xu. Partition level multiview subspace clustering. *Neural Networks*, 122:279–288, 2020.

Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

Yiqun Li, Feiping Nie, Heng Huang, and Jun Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8547–8555, 2021.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11183, 2021.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2023.

Chengliang Liu, Jie Wen, Zhihao Wu, Xiaoling Luo, Chao Huang, and Yong Xu. Information recovery-driven deep incomplete multiview clustering network. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):15442–15452, 2024.

Xinwang Liu, Xinbo Zhu, Miaomiao Li, Lijun Wang, Chang Tang, Jianping Yin, Dinggang Shen, Han Wang, and Wen Gao. Late fusion incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2410–2423, 2019.

Xinwang Liu, Xinbo Zhu, Miaomiao Li, Lijun Wang, En Zhu, Ting Liu, Marius Kloft, Dinggang Shen, Jianping Yin, and Wen Gao. Multiple kernel k-means with incomplete kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1191–1204, 2020.

Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2634–2646, 2021.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14193–14201, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Feiping Nie, Jing Li, and Xuelong Li. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 1881–1887, 2016.

Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017a.

Feiping Nie, Jing Li, and Xuelong Li. Self-weighted multiview clustering with multiple graphs. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2564–2570, 2017b.

Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Xi Peng, Jiashi Feng, Shijie Xiao, Wei-Yun Yau, Joey Tianyi Zhou, and Songfan Yang. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 27(10):5076–5086, 2018.

Xi Peng, Zhiqiang Huang, Jiancheng Lv, Hongmin Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5092–5101, 2019.

Lilei Sun, Jie Wen, Chengliang Liu, Lunke Fei, and Lusi Li. Balance guided incomplete multi-view spectral clustering. *Neural Networks*, 166:260–272, 2023.

Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Partial multi-view clustering via consistent gan. In *2018 IEEE International Conference on Data Mining*, pp. 1290–1295, 2018a.

Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30:1771–1783, 2021a.

Qianqian Wang, Huan Lian, Guangzhi Sun, Quanxue Gao, and Licheng Jiao. icmsc: Incomplete cross-modal subspace clustering. *IEEE Transactions on Image Processing*, 30:305–317, 2021b.

Qianqian Wang, Zhiqiang Tao, Wei Xia, Quanxue Gao, Xian Cao, and Licheng Jiao. Adversarial multiview clustering networks with adaptive fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10):7635–7647, 2023.

Qianqian Wang, Xurui Liao, Wei Feng, and Quanxue Gao. Debiased and denoised representation learning for incomplete multi-view clustering. In *International Conference on Learning Representations*, 2026.

Siwei Wang, Xinwang Liu, Li Liu, Wenxuan Tu, Xinzhong Zhu, Jiyuan Liu, Sihang Zhou, and En Zhu. Highly-efficient incomplete large-scale multi-view clustering with consensus bipartite graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9776–9785, 2022.

Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, Qing Zhang, and Xiaodi Huang. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing*, 24(11):3939–3949, 2015.

Yang Wang, Lin Wu, Xuemin Lin, and Junbin Gao. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (10):4833–4843, 2018b.

Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Hong Liu. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5393–5400, 2019.

Jie Wen, Yong Xu, and Hong Liu. Incomplete multiview spectral clustering with adaptive graph learning. *IEEE Transactions on Cybernetics*, 50(4):1418–1429, 2020a.

Jie Wen, Zheng Zhang, Yong Xu, Bob Zhang, Lunke Fei, and Guo-Sen Xie. CDIMC-net: Cognitive deep incomplete multi-view clustering network. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3230–3236, 2020b.

Jie Wen, Zheng Zhang, Zhao Zhang, Zhihao Wu, Lunke Fei, Yong Xu, and Bob Zhang. DIMC-net: Deep incomplete multi-view clustering network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3753–3761, 2020c.

Jie Wen, Ke Yan, Zheng Zhang, Yong Xu, Junqian Wang, Lunke Fei, and Bob Zhang. Adaptive graph completion based incomplete multi-view clustering. *IEEE Transactions on Multimedia*, 23:2493–2504, 2021a.

Jie Wen, Zheng Zhang, Zhao Zhang, Lunke Fei, and Meng Wang. Generalized incomplete multiview clustering with flexible locality structure diffusion. *IEEE Transactions on Cybernetics*, 51(1): 101–114, 2021b.

Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, and Yong Xu. Highly confident local structure based consensus graph learning for incomplete multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2023b.

Yi Wen, Siwei Wang, Ke Liang, Weixuan Liang, Xinhang Wan, Xinwang Liu, Suyuan Liu, Jiyuan Liu, and En Zhu. Scalable incomplete multi-view clustering with structure alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3031–3040, 2023c.

Wei Xia, Quanxue Gao, Qianqian Wang, and Xinbo Gao. Tensor completion-based incomplete multiview clustering. *IEEE Transactions on Cybernetics*, 52(12):13635–13644, 2022a.

Wei Xia, Qianqian Wang, Quanxue Gao, Xinbo Zhang, and Xinbo Gao. Self-supervised graph convolutional network for multi-view clustering. *IEEE Transactions on Multimedia*, 24:3182–3192, 2022b.

Gehui Xu, Jie Wen, Chengliang Liu, Bing Hu, Yicheng Liu, Lunke Fei, and Wei Wang. Deep variational incomplete multi-view clustering: Exploring shared clustering structures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16147–16155, 2024.

Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8761–8769, 2022a.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16051–16060, 2022b.

Jie Xu, Chao Li, Liang Peng, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 32:1354–1366, 2023.

Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2023.

Shengju Yu, Zhibin Dong, Siwei Wang, Pei Zhang, Yi Zhang, Xinwang Liu, Naiyang Guan, Tiejun Li, and Yiu-ming Cheung. Simple yet effective incomplete multi-view clustering: Similarity-level imputation and intra-view hybrid-group prototype construction. In *International Conference on Learning Representations*, 2025.

Pengxin Zeng, Mouxing Yang, Yiding Lu, Changqing Zhang, Peng Hu, and Xi Peng. Semantic invariant multi-view clustering with fully incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2139–2150, 2024.

Kun Zhan, Changqing Zhang, Junpeng Guan, and Junsheng Wang. Graph learning for multiview clustering. *IEEE Transactions on Cybernetics*, 48(10):2887–2895, 2018.

Kun Zhan, Feiping Nie, Jing Wang, and Yi Yang. Multiview consensus graph clustering. *IEEE Transactions on Image Processing*, 28(3):1261–1270, 2019.

Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4279–4287, 2017.

Changqing Zhang, Zongbo Han, Yajie Cui, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. CPM-Nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 557–567, 2019a.

Changqing Zhang, Yeqing Liu, and Huazhu Fu. AE2-Nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2577–2585, 2019b.

Shuping Zhao, Yulong Chen, Jie Wen, Lunke Fei, Jinrong Cui, and Tingting Chai. Confident block diagonal structure-aware invariable graph completion for incomplete multi-view clustering. In *International Conference on Learning Representations*, 2026.

Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. Uncover underlying correspondence for robust multi-view clustering. In *International Conference on Learning Representations*, 2026.

A PROOFS

A.1 STANDING NOTATION

The functionals of Definition 2 are written in the main text on the empirical pair (P_N, P_μ). The same formulae extend verbatim to any law P of (X, M) with mask marginal μ and any nested kernel K. Propositions 1 and 2 are statements about those functionals. The training losses replace them by finite-sample surrogates on P_N. No class label enters w, K, or either functional.

A content map is a measurable function of the operator output $A_m(\mathbf{x})$. When a map is specified by its values on nested pairs of a finite collection of instances, we assume those operator outputs are distinct, so the specification is well-defined. This holds generically for continuous view features that are not duplicated across samples; write (A0) for this injectivity.

Throughout, $\mathcal{I} = \{i : |\mathcal{O}_i| \geq 2\}$. Samples outside $\mathcal{I}$ are omitted only from $\mathcal{L}_{\mathrm{INV}}$; the reconstruction and decorrelation terms retain all samples with at least one observed view. The nested kernel used in training, and in every argument below, is the one-view restriction

$$K(\mathbf{m}_i - e_v \mid i) = \frac{m_{iv} s_v}{\sum_{u=1}^{V} m_{iu} s_u}, \qquad s_v = (\rho_v + \varepsilon)^{-\gamma_s}, \tag{24}$$

and $K(m' \mid i) = 0$ if m' is not obtained by dropping exactly one recorded view. Since $s_v > 0$, the kernel is strictly positive on every admissible one-view restriction. The law π_μ is Eq. (4) with this kernel. Thus (C1) is invariance along the nested pairs that π_μ charges, not invariance to every coarser mask. Write $\Sigma_M = \mathrm{Cov}(M, M)$ under the law of M currently in force. The implemented weight w is finite and strictly positive, so μ and q share a support.

A.2 PROOF OF PROPOSITION 1

It is enough to exhibit one law and one content map with $J_{\mathrm{INV}} = 0$ and $J_{\mathrm{IND}} > 0$, and one law and one content map with $J_{\mathrm{IND}} = 0$ and $J_{\mathrm{INV}} > 0$. The experimental quota protocol is not used: existence of any mask law already separates the two conditions. Neither construction uses class labels.

(C1) does not imply (C2). Assume $\Sigma_M \neq 0$, i.e. μ is not a Dirac mass on a single mask. Under (A0), the assignment $f_c(A_m(\mathbf{x}_i)) = \mathbf{m}_i$ for every nested argument $m \in \mathcal{M}_i$ is a well-defined function of $A_m(\mathbf{x}_i)$. Then $f_c(A_{m'}(\mathbf{x}_i)) = f_c(A_{\mathbf{m}_i}(\mathbf{x}_i)) = \mathbf{m}_i$ for every nested pair, so $J_{\mathrm{INV}}(f_c) = 0$ and (C1) holds. Under the recorded argument $M = \mathbf{m}_i$, one has $f_c(A_M(X)) = M$ almost surely, hence

$$J_{\mathrm{IND}}(f_c) = \left\|\Sigma_M\right\|_F^2 > 0. \tag{25}$$

Thus (C2) fails. Any empirical mask distribution with at least two distinct recorded patterns satisfies $\Sigma_M \neq 0$, because w does not alter the support of q. The construction is a function of the operator output and does not require dependence between X and M; it is compatible with $M \perp X$.

(C2) does not imply (C1). A map that depends on the current mask m through a nonlinear coordinate can be linearly uncorrelated with M and still move when a recorded view is dropped. The following lemma records this on product Bernoulli laws, which remain product Bernoulli under untruncated power reweighting.

Lemma 1. *Let $V \geq 2$ and let $M \in \{0,1\}^V$ be product Bernoulli with rates $\rho_v \in (0,1)$. Set $\eta(m) = m_1 m_2$ and*

$$f_c(A_m(\mathbf{x})) = \eta(m) - \mathbb{E}[\eta(M)] - b^\top \big(m - \mathbb{E}[M]\big), \qquad b = \Sigma_M^{-1} \mathrm{Cov}\big(\eta(M), M\big). \tag{26}$$

Then $\mathrm{Cov}(f_c(A_M(X)), M) = 0$. If a nested pair drops view 1 from a mask with $m_1 = m_2 = 1$, the nested difference equals $\rho_2 - 1 \neq 0$.

Proof of Lemma 1. View-wise independence of the coordinates of M gives $\Sigma_M = \mathrm{diag}(\rho_v(1-\rho_v))$, which is invertible because $\rho_v \in (0,1)$. Moreover

$$\mathrm{Cov}(\eta(M), M_1) = \rho_1\rho_2(1-\rho_1), \qquad \mathrm{Cov}(\eta(M), M_2) = \rho_1\rho_2(1-\rho_2), \tag{27}$$

and $\mathrm{Cov}(\eta(M), M_j) = 0$ for $j > 2$. Hence $b_1 = \rho_2$, $b_2 = \rho_1$, and $b_j = 0$ for $j > 2$. The normal equation $\Sigma_M b = \mathrm{Cov}(\eta(M), M)$ implies $\mathrm{Cov}(\eta(M) - b^\top M, M) = 0$, and therefore $\mathrm{Cov}(f_c(A_M(X)), M) = 0$. For m with $m_1 = m_2 = 1$ and $m' = m - e_1$,

$$f_c(A_{m'}(\mathbf{x})) - f_c(A_m(\mathbf{x})) = \big(\eta(m') - \eta(m)\big) - b^\top(m' - m) = (0-1) - b^\top(-e_1) = \rho_2 - 1. \tag{28}$$

The coordinate $b_2 = \rho_1$ does not enter this particular difference. Since $\rho_2 \in (0,1)$, the difference is nonzero, so f_c is not nested-invariant. □

Now take the mask law of Lemma 1 as q. If $\gamma = 0$, then $\mu = q$ and the lemma applies directly. If the truncation in Eq. (3) is inactive and $\mu(m) \propto q(m)^{1-\gamma}$, then μ is product Bernoulli with rates

$$\rho_v' = \frac{\rho_v^{1-\gamma}}{\rho_v^{1-\gamma} + (1-\rho_v)^{1-\gamma}} \in (0,1), \tag{29}$$

because $q(m)^{1-\gamma}$ factorizes across coordinates; in that case construct f_c from the μ-moments and apply the lemma with rates ρ_v'. Write ρ_2' for the second-view rate in either case, so $\rho_2' = \rho_2$ when $\gamma = 0$. Then $J_{\mathrm{IND}}(f_c) = 0$. The event $\{M_1 = M_2 = 1\}$ has positive μ-mass, and $K(\mathbf{m}_i - e_1 \mid i) > 0$ on that event, so π_μ charges the corresponding nested pairs and

$$J_{\mathrm{INV}}(f_c) \geq \pi_\mu\big(\{M_1 = M_2 = 1,\, m' = M - e_1\}\big)\,(\rho_2' - 1)^2 > 0. \tag{30}$$

Thus (C1) fails. The implemented truncation of w is a bounded modification of the power weight and is not claimed to preserve the product structure. The implication "(C2)⇒(C1)" is already refuted by these product laws; they are a product-Bernoulli q and its untruncated power reweighting, not the experimental quota protocol.

This second map is a function of the current mask in $A_m(\mathbf{x})$ only, so it likewise does not require dependence between X and M.

A.3 Proof of Proposition 2

By Definition 2, both summands of $\mathcal{J}_\mu$ are nonnegative, and $\mathcal{J}_\mu(f_c) = 0$ if and only if both vanish, i.e. if and only if (C1) and (C2) hold. A nonnegative random variable has mean zero if and only if it vanishes almost surely, so $J_{\mathrm{INV}}(f_c) = 0$ if and only if $f_c(A_{m'}(\mathbf{x}_i)) = f_c(A_{\mathbf{m}_i}(\mathbf{x}_i))$ for π_μ-almost every nested pair. Likewise $\|A\|_F^2 = 0$ if and only if $A = 0$, so $J_{\mathrm{IND}}(f_c) = 0$ if and only if $\mathrm{Cov}_\mu(f_c(A_M(X)), M) = 0$.

The training losses are surrogates, not identities, for those summands. Eq. (21) is a plug-in estimate of $\mathrm{Cov}_\mu(f_c(A_M(X)), M)$ up to the constant $1/(dV)$, so $\mathcal{L}_{\mathrm{IND}} = 0$ if and only if the empirical (C2) holds on the displayed full-sample average. Eq. (19) differs from the π_μ-empirical J_{INV} in the following respects: it uses ℓ_2-normalized content rather than f_c; it applies a stop-gradient to the recorded factor (which does not change the zero set, but changes the gradient); and it weights pairs by $w(\mathbf{m}_i')$ rather than by the π_μ-weight $w(\mathbf{m}_i)K(m' \mid i)$. In addition, training draws a single $m_i' \sim K(\cdot \mid i)$ per sample, whereas J_{INV} is an expectation under K. Vanishing $\mathcal{L}_{\mathrm{INV}}$ therefore forces equality of normalized contents on the sampled pairs, not equality of f_c. Minimizing only one surrogate therefore does not control the other population term, which is the content of Proposition 1.

Remark 1. *$\mathcal{J}_\mu(f_c) = 0$ does not uniquely determine f_c. Any nested-invariant map that is linearly uncorrelated with M has vanishing $\mathcal{J}_\mu$, including nonzero maps that retain a nonlinear function of the mask. Consequently, zero linear cross-covariance does not establish $C \perp M$ and does not provide a complete information-theoretic disentanglement guarantee. The training objective therefore uses $\mathcal{L}_{\mathrm{REC}}$ together with the two surrogates.*

B Further Specifications

This appendix records specifications that are not required to follow the method in Section 3. The main text remains self-contained: the content–observation criterion, the factorization, nested invariance, rebalanced decorrelation, the inference rule, and the training procedure are fully specified by Eqs. (1)–(23) and Algorithm 1.

B.1 Covariance Penalty

The penalty $\mathcal{L}_{\mathrm{cov}}$ is evaluated on the raw mini-batch of content factors produced by f_c. It is not evaluated on ℓ_2-normalized codes, not on a detached copy, and not on the full dataset. If $B < 2$, the implementation returns zero. If $B \geq 2$, it forms Σ by Eq. (16) and $\mathcal{L}_{\mathrm{cov}}$ by Eq. (14). The off-diagonal sum in Eq. (14) equals $\|\Sigma - \mathrm{diag}(\Sigma)\|_F^2/d^2$: the zero diagonal is included in the d^2 denominator.

The scale s_j is the population standard deviation (denominator B), matching `unbiased=False`, then floored at 10^{-4}. The product for Σ uses $\max(B-1, 1)$. These denominators are not the same. When the floor is inactive, $\sum_b z_{bj}^2 = B$ and $\Sigma_{jj} = B/(B-1)$, so

$$\frac{1}{d}\sum_{j=1}^{d}(\Sigma_{jj} - 1)^2 = \frac{1}{(B-1)^2} \tag{31}$$

is constant in the content batch and contributes no gradient. When $s_j = 10^{-4}$, the floor is constant and the diagonal term has a gradient that penalizes a collapsed coordinate. The off-diagonal term is the active decorrelation penalty in both regimes.

B.2 Nested Observation Pairs

A nested pair for sample i is admissible only if it is supported by the observed topology. Write $\mathbf{m}_i$ and $\mathbf{m}_i'$ for two binary masks of the same sample. The pair is used in $\mathcal{L}_{\mathrm{INV}}$ when

$$\mathrm{supp}(\mathbf{m}_i') \subset \mathrm{supp}(\mathbf{m}_i) \quad \text{and} \quad |\,\mathrm{supp}(\mathbf{m}_i)| \geq 2, \tag{32}$$

so every restricted coordinate was recorded and at least one view remains. The eligibility condition is per sample: if $|\mathcal{O}_i| = 1$, dropping the sole recorded view would produce the empty mask and no pair, whereas a two-view mask $[1, 1]$ admits either non-empty one-view restriction $[1, 0]$ or $[0, 1]$. The implemented kernel $K(\cdot \mid i)$ holds out exactly one recorded view; (C1) is invariance along those pairs. Nested masks are used only while estimating J_{INV} and are discarded thereafter; clustering uses the recorded mask $\mathbf{m}_i$ alone.

B.3 Pattern Weights

The empirical pattern distribution $q(\mathbf{m})$ is estimated from the observed training mask, which in the transductive IMVC protocol is the full matrix $\mathbf{M}$. No test label and no class prior is used. The weight $w(\mathbf{m})$ in Eq. (3) is a truncated inverse-frequency transform of the recorded pattern distribution. Truncation avoids a single rare pattern dominating the estimate of J_{IND}; the exponent γ interpolates between the empirical measure ($\gamma = 0$) and a stronger rebalancing. The population functional J_{INV} reweights recorded masks by $w(\mathbf{m}_i)$; the surrogate $\mathcal{L}_{\mathrm{INV}}$ uses $w(\mathbf{m}_i')$ on the nested mask.

B.4 Complexity

Let Ω be the number of observed view entries and let d be the content dimension. One training iteration encodes and reconstructs the observed views, which scales with Ω, and computes a second

Algorithm 2 View-wise unbalanced missingness mask

Require: Sample count N, view count V, target mean missing rate $r \in [0, 1)$, seed $s = 42$.
Ensure: Mask $\mathbf{M} \in \{0, 1\}^{N \times V}$ with $|\mathcal{O}_i| \geq 1$.
$\varepsilon \leftarrow 10^{-6}$; base $\leftarrow$ `RandomState`(s)
$\pi \leftarrow$ random permutation of $\{1, \ldots, V\}$ from base
$w \leftarrow$ increasing sort of V draws $\mathrm{Uniform}[0.5, 2.0)$ from base
$g \leftarrow V$ draws $\mathrm{Uniform}[0.2, 0.8)$ from base
$\tilde{\alpha}_{\pi(i)} \leftarrow w_i(1 + rg_i)$ for $i = 1, \ldots, V$
if $\max_v \tilde{\alpha}_v > (1 - \varepsilon)/r$ **then**
    $\tilde{\alpha} \leftarrow \tilde{\alpha} \cdot ((1 - \varepsilon)/r) / \max_v \tilde{\alpha}_v$
end if
$\alpha \leftarrow V\tilde{\alpha} / \sum_u \tilde{\alpha}_u$; $r_v \leftarrow \alpha_v r$
rng $\leftarrow$ `RandomState`$(s + \mathrm{int}(10r))$
$\mathrm{miss}_v \leftarrow \mathrm{clip}(\mathrm{stochround}(Nr_v), 0, N - 1)$; $\mathrm{exist}_v \leftarrow N - \mathrm{miss}_v$
if $\sum_v \mathrm{exist}_v < N$ **then**
    reduce miss (increase exist) on views with largest miss until $\sum_v \mathrm{exist}_v \geq N$
end if
$e_v \leftarrow N(1 - r_v)$; $n_v \leftarrow \min(\mathrm{exist}_v, \lfloor e_v \rfloor)$; $L \leftarrow N - \sum_v n_v$
fill leftover L by largest fractional parts of e_v, without exceeding exist_v, so that $\sum_v n_v = N$
$\mathbf{M} \leftarrow 0$; shuffle row indices with rng; assign the n_v first-pass ones on disjoint rows
for $v = 1$ to V **do**
    add $\mathrm{exist}_v - n_v$ further ones, sampled uniformly without replacement among currently unobserved entries of view v
end for
return $\mathbf{M}$

content map of nested masks for samples in $\mathcal{I}$. The cross-covariance in Eq. (21) is a $d \times V$ matrix and costs $O(NdV)$. Inference applies Eq. (23) and K-means on $\{\widehat{\mathbf{c}}_i\}$. The parameterization consists of the view-specific autoencoders, the content map f_c, and the observation map g.

C Experimental Protocol Notes

C.1 Unbalanced Missingness Protocol

The comparison in Section 4 uses one generated mask per dataset and each target mean missing rate $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The scalar r is the target mean missing rate across views, not a per-view Bernoulli parameter and not the fraction of fully unobserved samples. Write $\alpha \in \mathbb{R}^V$ for a view-weight vector of mean one. The intended per-view missing probabilities are

$$r_v = \alpha_v r, \qquad \frac{1}{V} \sum_{v=1}^{V} r_v = r. \tag{33}$$

Because α is drawn once and is not uniform, some views are systematically scarcer; this is the unbalanced missingness. Integer column quotas then realize r_v, a first-pass assignment enforces $|\mathcal{O}_i| \geq 1$, and no column is allowed to be entirely missing. The mask depends only on (N, V, r) and seed 42; features and labels are not used. Samples are first permuted by NumPy `RandomState`(42). Algorithm 2 is then applied in that shuffled order. The resulting matrix $\mathbf{M}$ is generated once per dataset and rate by Algorithm 2. Occupancy statistics in this appendix are computed from these canonical COIN masks. Compared methods are evaluated under this view-wise quota protocol rather than by drawing independent missingness realizations for each method.

Seed 42 has two roles. It draws the view permutation and the weights α, which decide which views are scarce. Quota sampling and entry assignment use `RandomState`$(42 + \mathrm{int}(10r))$. No other seed enters mask generation. Let $\mathrm{stochround}(x)$ denote coordinate-wise stochastic rounding, $\mathrm{stochround}(x)_v = \lfloor x_v \rfloor + \mathrm{Bernoulli}(x_v - \lfloor x_v \rfloor)$, with independent Bernoulli draws. Algorithm 2 uses NumPy `RandomState` (not `Generator`). Realized availabilities ρ_v are reported in Tables 4 and 5: mean availability satisfies $\bar{\rho} \approx 1 - r$ up to integer rounding, while $\Delta_\rho = \max_v \rho_v - \min_v \rho_v$ grows with r.

Table 4: Realized per-view availability $\rho_v = \frac{1}{N}\sum_{i=1}^{N} m_{iv}$ on the three-view and four-view datasets, generated by Algorithm 2 with seed 42. Mean availability is $\bar{\rho} = \frac{1}{V}\sum_v \rho_v$. The spread $\Delta_\rho = \max_v \rho_v - \min_v \rho_v$ is the view-wise imbalance. A dash marks a view that the dataset does not contain.

Dataset	r	ρ_1	ρ_2	ρ_3	ρ_4	$\bar{\rho}$	Δ_ρ
Synthetic3d	0.1	0.915	0.903	0.883	–	0.901	0.032
	0.2	0.830	0.805	0.767	–	0.801	0.063
	0.3	0.742	0.707	0.652	–	0.700	0.090
	0.4	0.657	0.607	0.537	–	0.600	0.120
	0.5	0.570	0.507	0.423	–	0.500	0.147
Washington	0.1	0.874	0.935	0.857	0.935	0.900	0.078
	0.2	0.752	0.874	0.713	0.865	0.801	0.161
	0.3	0.626	0.817	0.561	0.796	0.700	0.257
	0.4	0.496	0.765	0.413	0.726	0.600	0.352
	0.5	0.374	0.709	0.265	0.652	0.500	0.443
TinyImageNet	0.1	0.875	0.936	0.856	0.933	0.900	0.081
	0.2	0.749	0.876	0.710	0.865	0.800	0.166
	0.3	0.624	0.819	0.563	0.794	0.700	0.256
	0.4	0.498	0.764	0.414	0.723	0.600	0.350
	0.5	0.372	0.712	0.265	0.650	0.500	0.448

Table 5: Realized per-view availability on the five-view and six-view datasets, generated by the same seed-42 instance of Algorithm 2.

Dataset	r	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6	$\bar{\rho}$	Δ_ρ
HandWritten	0.1	0.887	0.947	0.900	0.860	0.905	–	0.900	0.086
	0.2	0.779	0.894	0.798	0.716	0.813	–	0.800	0.178
	0.3	0.674	0.841	0.695	0.567	0.723	–	0.700	0.275
	0.4	0.573	0.789	0.589	0.414	0.635	–	0.600	0.374
	0.5	0.474	0.737	0.480	0.259	0.549	–	0.500	0.478
MSRCV1	0.1	0.948	0.938	0.871	0.838	0.876	0.929	0.900	0.110
	0.2	0.890	0.871	0.738	0.686	0.752	0.857	0.799	0.205
	0.3	0.838	0.814	0.600	0.533	0.633	0.786	0.701	0.305
	0.4	0.776	0.757	0.457	0.386	0.519	0.710	0.601	0.390
	0.5	0.719	0.700	0.310	0.238	0.400	0.624	0.498	0.481

C.2 Nested-Invariance Eligibility

The nested-invariance term $\mathcal{L}_{\text{INV}}$ is defined only for samples with at least two recorded views, because a one-view-drop restriction must remain non-empty. Eligibility is therefore per sample: Table 6 reports the fraction of samples with $|\mathcal{O}_i| \geq 2$ under each canonical mask. A sample with $|\mathcal{O}_i| = 1$ is excluded only from $\mathcal{L}_{\text{INV}}$ and remains in $\mathcal{L}_{\text{REC}}$ and $\mathcal{L}_{\text{IND}}$. Nested invariance is not applied to every sample.

C.3 Observation-Pattern Frequencies

The empirical pattern distribution $q(m)$ in Eq. (2) is the occupancy of recorded joint masks. Table 7 and Figure 4 summarize that occupancy on the same canonical masks, without labels and without resampling.

The generated masks exhibit not only unequal marginal view availability but also uneven frequencies across joint observation patterns. Consequently, empirical content–observation statistics can be dominated by frequently occurring masks. This supports the inverse-frequency reweighting in Eq. (3), which reduces the dominance of common observation conditions when estimating the decorrelation objective. The statistics demonstrate uneven pattern occupancy; they do not establish statistical independence, causal bias, or a clustering-performance guarantee.

Table 6: Nested-invariance eligibility on the canonical seed-42 masks. A sample is eligible for $\mathcal{L}_{\mathrm{INV}}$ iff $|\mathcal{O}_i| \geq 2$. Single-view samples remain in $\mathcal{L}_{\mathrm{REC}}$ and $\mathcal{L}_{\mathrm{IND}}$.

Dataset	r	Single-view (%)	INV-eligible (%)	Mean $\lvert\mathcal{O}_i\rvert$
MSRCV1	0.1	0.00	100.00	5.40
	0.2	0.00	100.00	4.80
	0.3	0.95	99.05	4.20
	0.4	0.95	99.05	3.60
	0.5	6.19	93.81	2.99
Washington	0.1	0.43	99.57	3.60
	0.2	3.04	96.96	3.20
	0.3	6.09	93.91	2.80
	0.4	11.74	88.26	2.40
	0.5	22.17	77.83	2.00
Synthetic3d	0.1	1.83	98.17	2.70
	0.2	7.50	92.50	2.40
	0.3	17.00	83.00	2.10
	0.4	32.00	68.00	1.80
	0.5	52.33	47.67	1.50
HandWritten	0.1	0.00	100.00	4.50
	0.2	0.20	99.80	4.00
	0.3	1.00	99.00	3.50
	0.4	5.00	95.00	3.00
	0.5	13.45	86.55	2.50
TinyImageNet	0.1	0.13	99.87	3.60
	0.2	0.88	99.12	3.20
	0.3	4.95	95.05	2.80
	0.4	11.55	88.45	2.40
	0.5	28.62	71.38	2.00

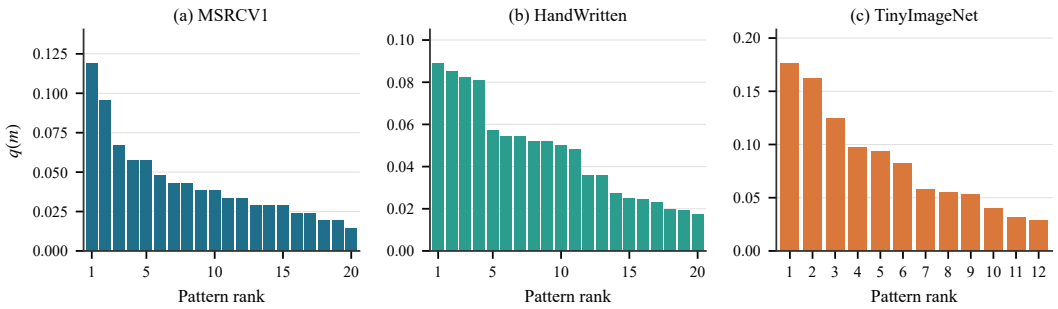


Figure 4: Descending observation-pattern frequencies $q(m)$ at $r = 0.5$ on the canonical masks. Bars are ordered by rank; binary pattern strings are omitted. MSRCV1 and HandWritten show the twenty most frequent patterns ($K_m = 37$ and 24); TinyImageNet shows all 12 recorded patterns.

Labels are used after training to select per-cell coefficients for the reported COIN cells, as specified in Appendix C.5. We report ACC, NMI, and ARI averaged over ten K-means initializations. CIMLN and BURG are run from their official code with official recommended hyperparameters. For BURG, we use $\min(512, N)$ with $\mathrm{drop_last} = (N > B)$ for the dual-consistency loader, so $N < 512$ does not yield an empty iterator, and recover losses on batches with no complete samples are skipped rather than back-propagated. Clustering follows Eq. (23) on every run: ℓ_2-normalized content followed by K-means. Training uses a reconstruction warm-up followed by a subsequent joint-optimization stage; the epoch schedule is documented in Appendix C, while sensitivity ranges are reported in the main text and appendix. Ablations isolate $\mathcal{L}_{\mathrm{REC}}$, $\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{INV}}$, $\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{IND}}$, and $\mathcal{L}_{\mathrm{REC}} + \alpha\mathcal{L}_{\mathrm{INV}} + \beta\mathcal{L}_{\mathrm{IND}}$ rather than reporting every auxiliary coefficient as a separate method. The clustering rule is held fixed; α and β remain scalar trade-offs. The α–β grid used in Section 4.4 holds the remaining coefficients fixed as described in Appendix C.

Table 7: Observation-pattern frequencies $q(m)$ on the same canonical masks. K_m counts distinct recorded patterns; $q_{\max}/q_{\min}$ is the ratio of the most to the least frequent non-empty pattern. Normalized entropy is $H(M)/\log K_m$.

Dataset	r	K_m	$q_{\max}$	$q_{\min}$	$q_{\max}/q_{\min}$	H_{norm}
MSRCV1	0.1	20	0.510	0.0048	107.0	0.607
	0.2	32	0.219	0.0048	46.0	0.776
	0.3	43	0.124	0.0048	26.0	0.860
	0.4	36	0.138	0.0048	29.0	0.888
	0.5	37	0.119	0.0048	25.0	0.899
Washington	0.1	9	0.657	0.0043	151.0	0.562
	0.2	12	0.383	0.0087	44.0	0.764
	0.3	12	0.200	0.0043	46.0	0.885
	0.4	12	0.217	0.0130	16.7	0.922
	0.5	11	0.304	0.0043	70.0	0.832
Synthetic3d	0.1	6	0.720	0.0083	86.4	0.534
	0.2	6	0.477	0.0283	16.8	0.807
	0.3	6	0.270	0.0700	3.9	0.942
	0.4	6	0.293	0.0733	4.0	0.954
	0.5	6	0.297	0.0233	12.7	0.884
HandWritten	0.1	20	0.576	0.0005	1151.0	0.530
	0.2	24	0.288	0.0005	576.0	0.742
	0.3	24	0.118	0.0040	29.6	0.882
	0.4	24	0.086	0.0170	5.0	0.973
	0.5	24	0.089	0.0155	5.7	0.951
TinyImageNet	0.1	12	0.641	0.0003	1922.5	0.507
	0.2	12	0.349	0.0040	87.2	0.738
	0.3	12	0.189	0.0163	11.6	0.897
	0.4	12	0.214	0.0242	8.9	0.903
	0.5	12	0.176	0.0287	6.2	0.938

Table 8: Dataset sizes and view dimensions used by COIN.

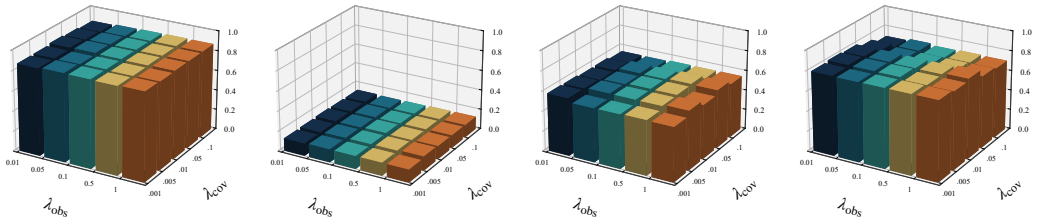
Dataset	N	K	V	Input dimensions
MSRCV1	210	7	6	1302, 48, 512, 100, 256, 210
Washington	230	5	4	230, 1703, 230, 230
Synthetic3d	600	3	3	3, 3, 3
HandWritten	2000	10	5	240, 76, 216, 47, 64
TinyImageNet	6000	20	4	576, 944, 512, 640

C.4 Network Architecture

Widths are shared across datasets. Each view uses an encoder–decoder pair $E_v : \mathbb{R}^{d_v} \to \mathbb{R}^{64}$ and $D_v : \mathbb{R}^{64} \to \mathbb{R}^{d_v}$, implemented as $\text{Linear}(d_v, 256) \to \text{GELU} \to \text{LayerNorm} \to \text{Linear}(256, 64)$ and $\text{Linear}(64, 256) \to \text{GELU} \to \text{Linear}(256, d_v)$. The view-specific map P_v is $\text{Linear}(64, 64) \to \text{LayerNorm}$. Observed tokens are gated by m_{iv} as in Eq. (9) and concatenated, so the content map f_c is $\text{Linear}(64V, 256) \to \text{GELU} \to \text{LayerNorm} \to \text{Linear}(256, 64)$. The observation map g is $\text{Linear}(V, 64) \to \text{GELU} \to \text{LayerNorm} \to \text{Linear}(64, 16)$, with mask readout $\text{Linear}(16, V)$. Each renderer R_v is $\text{Linear}(80, 256) \to \text{GELU} \to \text{Linear}(256, d_v)$, where $80 = 64 + 16$ is the concatenation of $(\mathbf{c}_i, \mathbf{o}_i)$. Gated concatenation and a one-view nested restriction are shared across datasets and rates.

C.5 Training Settings

Optimization uses Adam with encoder learning rate 10^{-4}, remaining-parameter learning rate 10^{-3}, weight decay 10^{-5}, configured batch size 256, and gradient clipping at 5; the final, smaller batch is retained. The covariance penalty is the batch estimator of Eqs. (14)–(16): raw $\mathbf{c}$, per-coordinate



(a) Synthetic3d (b) TinyImageNet (c) Washington (d) MSRCV1

Figure 5: Sensitivity analysis of hyperparameters λ_{obs} and λ_{cov} on four datasets.

mean subtraction, population standard deviation with denominator B (`unbiased=False`) floored at 10^{-4}, and $\Sigma = \sum_b \mathbf{z}_b \mathbf{z}_b^\top / \max(B-1, 1)$. Batches with $B < 2$ contribute zero. There is no ℓ_2 normalization, no stop-gradient, and no full-dataset covariance; the diagonal term is not an identity of the scaling (Appendix B.1). Training runs $T_w = 80$ reconstruction epochs and 100 subsequent epochs with $\mathcal{L}_{\text{INV}}$ and $\mathcal{L}_{\text{IND}}$. Nested restrictions drop exactly one recorded view on samples with at least two observed views. Inference ℓ_2-normalizes $\mathbf{c}_i$ and applies K-means with ten initializations. Table results use the seed-42 mask of Algorithm 2. Shared widths are latent dimension 64, observation dimension 16, and hidden width 256. The coefficients $\gamma = 0.5$, $w_{\max} = 8$, and $\gamma_s = 1$ are shared, and $\mathcal{L}_{\text{INV}}$ has no hinge margin. The 10^{-4} covariance scale floor is fixed by the implementation and is not tuned. Tables 2 and 3 do *not* use one $(\alpha, \beta, \lambda_{\text{cov}}, \lambda_{\text{obs}})$ setting for every dataset and missing rate. Those coefficients, together with implementation multipliers λ_{view} and λ_{render} on $\mathcal{L}_{\text{view}}$ and the recorded-view rendering term, are chosen independently for each dataset–rate cell by maximizing ACC, then NMI, then ARI, on a discrete grid after training. Labels do not enter the training loss; they are used for this post-training selection. Sensitivity sweeps in Section 4.4 and Appendix D are post-hoc and were not used to re-select the reported cells.

D Sensitivity of Structural Coefficients

Section 4.4 varies α and β. This appendix varies λ_{cov}, λ_{obs}, γ, and γ_s at unbalanced missing rate $r = 0.5$ on the same four datasets as Figure 2. Gated concatenation and a one-view nested restriction are held fixed. The coefficients α, β, and any coefficient not on the plotted axis are frozen at the Table 2 per-cell values for $r = 0.5$, not at a single shared default. The truncation $w_{\max}$ and the reconstruction horizon T_w are not re-swept.

Figure 5 reports a 5×5 grid. The axes are $\lambda_{\text{obs}} \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$ and $\lambda_{\text{cov}} \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$. ACC ranges are 86.28–88.62 on Synthetic3d, 12.23–14.21 on TinyImageNet, 53.83–59.22 on Washington, and 70.38–80.67 on MSRCV1. The largest swing is on MSRCV1 (10.29 points).

Figures 6–7 report one-dimensional sweeps with the same freeze. For $\gamma \in \{0, 0.25, 0.5, 0.75, 1.0\}$, ACC ranges are 87.80–88.67 on Synthetic3d, 12.59–13.75 on TinyImageNet, 58.57–60.35 on Washington, and 76.95–80.14 on MSRCV1. For $\gamma_s \in \{0, 0.5, 1.0, 1.5, 2.0\}$, ACC ranges are 87.72–88.62 on Synthetic3d, 13.08–13.88 on TinyImageNet, 53.52–58.74 on Washington, and 71.95–79.67 on MSRCV1. The largest one-dimensional swings are 3.19 points for γ on MSRCV1 and 7.71 points for γ_s on MSRCV1. A hinge $[\|\widehat{\mathbf{c}}_i' - \text{sg}(\widehat{\mathbf{c}}_i)\|_2^2 - \tau]_+$ was considered and is not used in Eq. (19). Figure 8 is a post-hoc negative control of that discarded option: $\tau = 0$ matches the implemented squared-distance objective, while $\tau = 0.2$ drops Synthetic3d ACC from 88.62 to 53.00. These plots isolate scalar coefficients of a single factorization; they do not change the pooling rule or the nested restriction.

E Direct Diagnostics of Observation Dependence

Table 3 shows that nested invariance and rebalanced decorrelation are complementary for clustering accuracy. This appendix measures the two intended mechanisms on frozen content, without changing the architecture, the training objective, or Table 2. The four variants match Table 3: $\mathcal{L}_{\text{REC}}$,

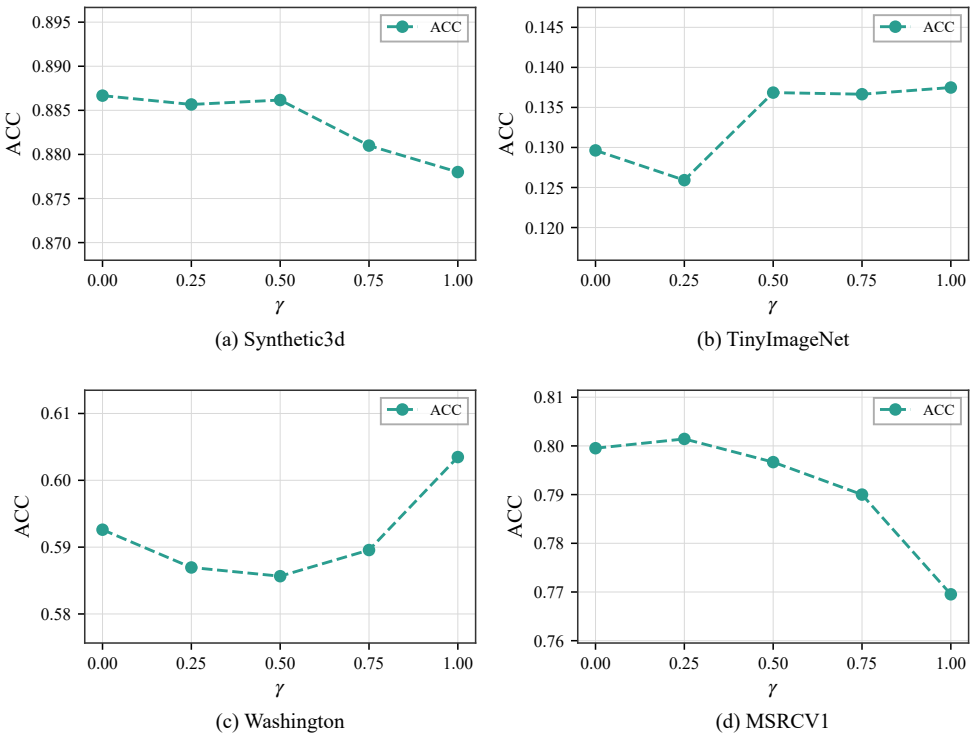


Figure 6: Clustering performance of COIN with different γ on different datasets.

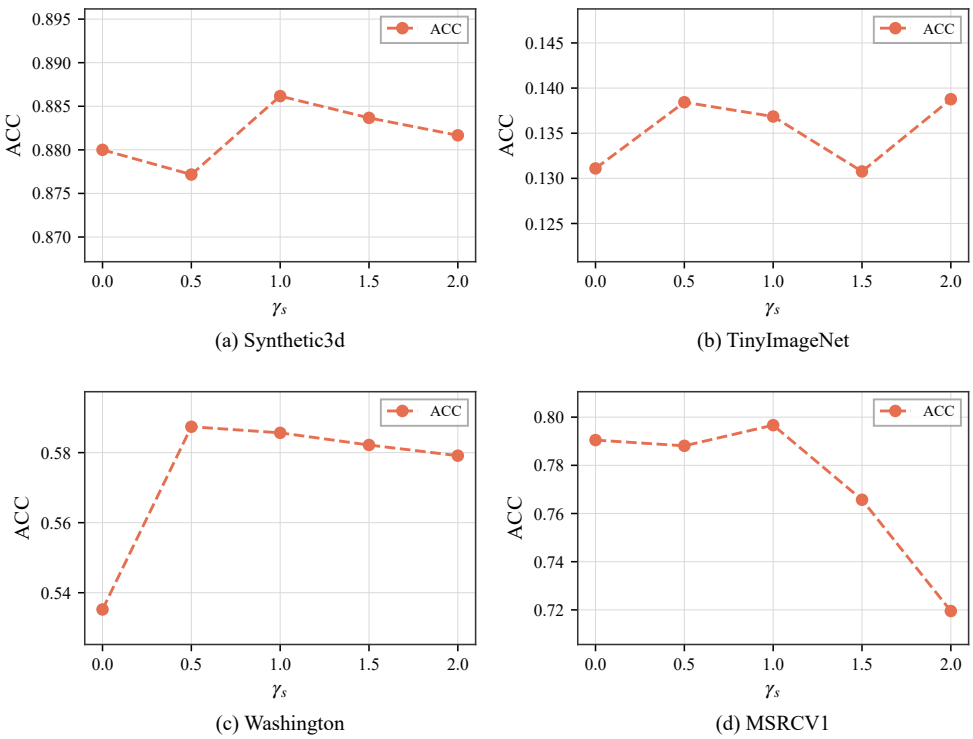
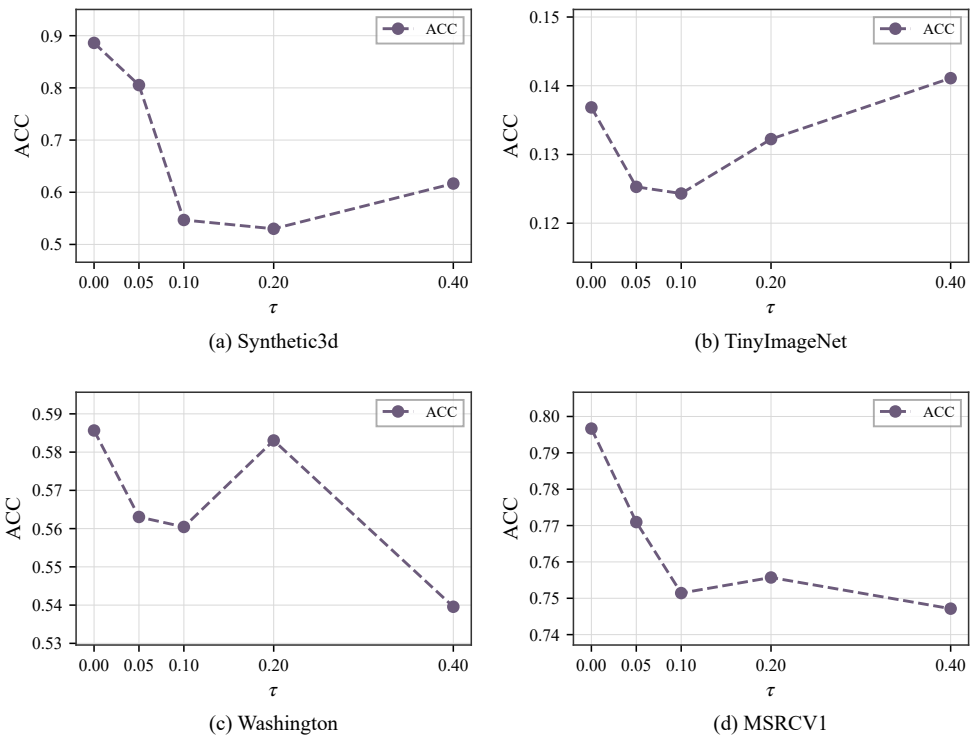


Figure 7: Clustering performance of COIN with different γ_s on different datasets.

$\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{INV}}$, $\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{IND}}$, and the full model. Checkpoints are matched copies of those cells (gated concatenation, one-view nested restriction, child invariance weights, seed 42).



(a) Synthetic3d (b) TinyImageNet

(c) Washington (d) MSRCV1

Figure 8: Post-hoc sweep of a discarded hinge on $\mathcal{L}_{\mathrm{INV}}$. COIN uses $\tau = 0$, i.e. weighted squared distance.

Same-instance representation drift. For each sample with $|\mathcal{O}_i| \geq 2$, let $\mathbf{c}_i = f_c(A_{\mathbf{m}_i}(\mathbf{x}_i))$ be the recorded-mask content and $\mathbf{c}'_i = f_c(A_{\mathbf{m}'_i}(\mathbf{x}_i))$ the content after a one-view nested restriction drawn from Eq. (17) with a shared drop seed. Write $\mathbf{z}_i = \mathbf{c}_i/\|\mathbf{c}_i\|_2$ and $\mathbf{z}'_i = \mathbf{c}'_i/\|\mathbf{c}'_i\|_2$. The drift is

$$D_{\mathrm{obs}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|\mathbf{z}_i - \mathbf{z}'_i\|_2^2. \tag{34}$$

Observation leakage probe. A logistic classifier predicts each recorded bit m_{iv} from frozen $\mathbf{c}_i$ on a fixed 80/20 split with seed 42; the probe does not enter COIN training. We report the mean per-view ROC-AUC and the unweighted $\|\mathrm{Cov}(C, M)\|_F^2$ of ℓ_2-normalized content versus the recorded mask. The probe is a linear readout. A lower AUC is not statistical independence $C \perp M$.

Table 9 supports the intended complementarity. Adding $\mathcal{L}_{\mathrm{INV}}$ reduces D_{obs} by one to three orders of magnitude on every dataset (e.g., $0.5164 \to 0.0026$ on Synthetic3d and $0.5273 \to 0.0020$ on TinyImageNet), while adding only $\mathcal{L}_{\mathrm{IND}}$ changes D_{obs} modestly. $\|\mathrm{Cov}(C, M)\|_F^2$ is smallest for the full model on every dataset. The linear mask probe does not track that covariance: $\mathcal{L}_{\mathrm{REC}} + \mathcal{L}_{\mathrm{IND}}$ often lowers mean ROC-AUC, including values below chance on the smallest datasets, which we report as raw test AUCs rather than as anti-correlation. The full model can retain a high probe AUC (perfect on Synthetic3d) even when $\|\mathrm{Cov}(C, M)\|_F^2$ is of order 10^{-5}. Thus $\mathcal{L}_{\mathrm{IND}}$ is a linear decorrelation penalty, not a test of independence, and the two diagnostics should be read jointly with Table 3 rather than as a substitute for clustering accuracy. The same qualitative pattern holds at $r = 0.3$ in Table 10.

Table 9: Direct diagnostics of observation dependence at unbalanced missing rate $r = 0.5$. D_{obs} is same-instance content drift under a one-view nested restriction. Mask AUC is a frozen linear probe $c_i \to m_{iv}$ (chance $= 0.5$); a lower AUC does not imply $C \perp M$.

Dataset	Method	$D_{\text{obs}} \downarrow$	Mask AUC$\downarrow$	$\|\text{Cov}(C, M)\|_F^2 \downarrow$
MSRCV1	$\mathcal{L}_{\text{REC}}$	0.2728	0.998	0.0207
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0045	0.973	2.48×10^{-4}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.2211	0.342	0.0013
	Full COIN	0.0040	0.846	5.58×10^{-5}
Washington	$\mathcal{L}_{\text{REC}}$	0.4983	0.678	0.0099
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0003	0.991	2.16×10^{-5}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.4743	0.239	0.0018
	Full COIN	0.0003	0.963	1.14×10^{-5}
Synthetic3d	$\mathcal{L}_{\text{REC}}$	0.5164	1.000	0.0595
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0026	1.000	1.98×10^{-4}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.3160	0.989	5.86×10^{-4}
	Full COIN	0.0031	1.000	5.50×10^{-5}
HandWritten	$\mathcal{L}_{\text{REC}}$	0.3855	0.937	0.0098
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0035	0.932	4.58×10^{-5}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.3242	0.528	5.84×10^{-4}
	Full COIN	0.0025	0.919	1.97×10^{-5}
TinyImageNet	$\mathcal{L}_{\text{REC}}$	0.5273	0.881	0.0090
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0020	0.970	1.19×10^{-5}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.4998	0.731	0.0041
	Full COIN	0.0016	0.924	8.52×10^{-6}

Table 10: Direct diagnostics of observation dependence at unbalanced missing rate $r = 0.3$. Metric definitions match Table 9; a lower mask AUC does not imply $C \perp M$.

Dataset	Method	$D_{\text{obs}} \downarrow$	Mask AUC$\downarrow$	$\|\text{Cov}(C, M)\|_F^2 \downarrow$
MSRCV1	$\mathcal{L}_{\text{REC}}$	0.1784	0.999	0.0173
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0024	0.905	1.35×10^{-4}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.1166	0.319	5.60×10^{-4}
	Full COIN	0.0022	0.880	8.16×10^{-5}
Washington	$\mathcal{L}_{\text{REC}}$	0.4083	0.673	0.0069
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0007	0.993	1.58×10^{-5}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.3863	0.166	7.47×10^{-4}
	Full COIN	0.0007	0.986	1.44×10^{-5}
Synthetic3d	$\mathcal{L}_{\text{REC}}$	0.5044	1.000	0.0353
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0013	1.000	2.58×10^{-5}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.2920	0.991	5.45×10^{-4}
	Full COIN	0.0012	1.000	1.50×10^{-5}
HandWritten	$\mathcal{L}_{\text{REC}}$	0.2220	0.883	0.0048
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0019	0.722	1.60×10^{-5}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.1823	0.590	5.18×10^{-4}
	Full COIN	0.0020	0.888	1.41×10^{-5}
TinyImageNet	$\mathcal{L}_{\text{REC}}$	0.3757	0.851	0.0060
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{INV}}$	0.0008	0.992	7.30×10^{-6}
	$\mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{IND}}$	0.3251	0.641	0.0016
	Full COIN	0.0007	0.990	5.06×10^{-6}