SELECTION-CONSISTENT COMPLETION FOR INCOMPLETE MULTI-VIEW UNSUPERVISED FEATURE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete multi-view unsupervised feature selection aims to identify a compact subset of original features when some views are missing. Existing joint methods combine missing-view completion with feature selection, but the completion process may still rely on features that are ultimately excluded from the selected subset, creating a mismatch between the selected-feature representation and the information used by completion. We introduce **selection–completion consistency**, which requires missing-view completion to rely only on currently available selected features, while allowing richer observed information to provide auxiliary supervision during training. Based on this principle, we propose **Selection-Consistent Multi-View Feature Selection (SC-MVFS)**, which integrates exact-budget feature selection, probabilistic multi-view posterior fusion and joint learning with a full-input teacher and a selected-input student. A conditional squared-risk decomposition further characterizes the predictive information difference between selected and all observed features. Experiments on eight real-world datasets and controlled synthetic data demonstrate competitive clustering performance. Additional controlled experiments examine the effects of using consistent feature inputs during training and final completion, as well as the conditions under which this consistency becomes more important.

1 INTRODUCTION

Multi-view data describe the same object from different perspectives, and the complementary information across views can provide richer representations than a single view alone (Baltrušaitis et al., 2018; Zhao et al., 2017; Ngiam et al., 2011). In real-world applications, however, some views are often unavailable for individual samples due to acquisition failures, device differences, or other practical factors (Ma et al., 2021; Reza et al., 2024). Incomplete multi-view unsupervised feature selection therefore aims to identify a small set of representative original features from high-dimensional multi-view data under partial observations (Xu et al., 2021; Huang et al., 2023). Because missing views weaken the complementary relations that can be exploited across views, recent methods increasingly couple missing-view completion with feature selection, using recovered cross-view information to support feature evaluation and selection (Huang et al., 2024; 2025).

However, many existing joint methods learn missing-view completion from full-dimensional observed views, while determining the final compact feature subset separately from learned feature weights or rankings. Consequently, even when only K features are retained, the completion module may still depend during training on observed features excluded from the final subset. The model may therefore achieve low reconstruction error by exploiting information outside the final selected-feature interface, without explicitly learning to complete missing views from the selected features alone. This creates a mismatch between the information used for completion and the outcome of feature selection.

We therefore introduce **selection–completion consistency**: given a selected subset, missing-view completion should depend only on the currently available selected features. Richer observed information may still provide auxiliary supervision during training, but it should not become a required input to the post-selection completion pathway.

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.

To instantiate this principle, we propose Selection-Consistent Multi-View Feature Selection (SC-MVFS). SC-MVFS uses an exact global feature budget. During training, the student performs inference and missing-view completion using only the selected observed features, while a teacher that sees all currently observed features provides additional supervision. The model combines information from the available views probabilistically and alternates between updating the teacher targets and learning the student parameters and feature subset. After training, a deterministic K-feature subset is fixed and used for both final inference and completion. We further characterize the information difference between selected and all observed features through a conditional squared-risk analysis. Our main contributions are as follows:

- We introduce **selection–completion consistency** and use a conditional squared-risk decomposition to characterize the population information cost of restricting completion to selected observed features.
- **SC-MVFS** is developed under an exact feature budget, where a teacher with access to all currently observed features provides auxiliary supervision, while the student is restricted to the selected features for posterior inference and missing-view completion. This design aligns the student completion inputs used during training with the final selected-feature completion interface.
- Extensive experiments on eight real-world datasets and controlled synthetic data show that SC-MVFS achieves competitive clustering performance. Additional controlled comparisons examine the effect of using different feature inputs during training and final completion, together with the roles of the main training components.

2 Related Work

Incomplete multi-view feature selection. Existing methods exploit incomplete multi-view information through different structure-recovery and feature-selection strategies. Xu et al. (2021) propose CVFS, which models instance- and feature-level cross-view relations to mitigate missing information. Huang et al. (2023) introduce C2IMUFS, which reconstructs complete sample-similarity structures while jointly learning complementary, consensus, and feature-selection representations. Huang et al. (2024) present UNIFIER, which combines graph-guided missing-view imputation with continuous feature weighting. Huang et al. (2025) propose TIME-FS, which uses tensor structure to coordinate missing-view recovery, low-dimensional representation, and row-sparse feature selection. Yu et al. (2025) introduce CLSG, which recovers missing information through latent representations and learns confident local graphs to guide feature selection. Yang et al. (2025) propose TERUIMUFS, which combines error-robust tensor representation, missing-sample recovery, and feature selection for unbalanced incomplete data. Yang et al. (2024) present UIMUFSLR, which uses an adaptively weighted unified graph and an explicit low-redundancy constraint for feature selection. Detailed method-level distinctions are provided in Appendix C.

Differentiable feature selection. Continuous relaxations enable discrete feature selection to be optimized with gradient-based methods. Maddison et al. (2016) introduce the Concrete distribution, which provides a differentiable relaxation of discrete random variables. Jang et al. (2016) propose Gumbel-Softmax, which enables differentiable sampling from categorical distributions. Balın et al. (2019) introduce the Concrete Autoencoder (CAE), which learns a subset of input features and reconstructs the original input from the selected features.

Latent-variable learning with missing views. Probabilistic latent-variable models provide a natural framework for learning from partially observed multi-view data. Wu & Goodman (2018) introduce the Multimodal Variational Autoencoder (MVAE), which uses a Product-of-Experts (PoE) formulation to fuse available modalities. Shi et al. (2019) propose the Mixture-of-Experts Multimodal Variational Autoencoder (MMVAE), which aggregates modality-specific information through a mixture-of-experts formulation. Sutter et al. (2021) introduce the generalized multimodal ELBO, which extends variational learning to different subsets of available modalities. Gao & Pu (2025) further model incomplete multi-view data through cyclic interactions among VAEs to enhance cross-view representation learning.

3 PROBLEM FORMULATION

Let $\mathcal{X} = \{\boldsymbol{X}^{(v)}\}_{v=1}^{V}$ denote a multi-view dataset with n samples, where $\boldsymbol{X}^{(v)} \in \mathbb{R}^{n \times d_v}$ is the v-th view and d_v denotes the number of features in that view. Let $o_{iv} \in \{0, 1\}$ indicate whether sample i is observed in view v, and assume each sample has at least one observed view. Write $\boldsymbol{O} = (o_{iv}) \in \{0, 1\}^{n \times V}$ for the observation mask and $\mathbf{x}_i^{(v)} \in \mathbb{R}^{d_v}$ for the feature vector of sample i in view v. We seek binary feature masks $\mathbf{m}^{(v)} \in \{0, 1\}^{d_v}$ satisfying

$$\sum_{v=1}^{V} \|\mathbf{m}^{(v)}\|_0 = K, \tag{1}$$

where K is the global feature budget. We write $\mathbf{m}$ for the concatenation of the per-view masks $\mathbf{m}^{(v)}$. In addition to this exact global cardinality constraint, we require that final posterior inference and missing-view completion rely only on the currently available selected features.

4 PROPOSED METHOD

Figure 1 provides an overview of SC-MVFS. To ensure that only the selected features are used during both training and final completion, the student performs posterior inference and reconstruction using only the currently selected features. Since richer observed information may still provide useful supervision during learning, a full-input teacher supplies auxiliary posterior and assignment targets, while remaining excluded from final completion. Product-of-Experts (PoE) fusion accommodates the varying sets of available views across samples, and an exact global feature budget determines the selected feature subset. After training, the resulting deterministic K-feature subset is fixed and used for both final inference and missing-view completion.

4.1 SELECTION–COMPLETION CONSISTENCY

Fix a selected feature subset S and an observation pattern $O = o$ with $\Pr(O = o) > 0$. Let U denote the observed features selected by S, W the observed features outside S, and $\mathbf{T}$ a square-integrable target vector from a missing view. Let g and h denote predictors of $\mathbf{T}$ based on U and (U, W), respectively. We write $\mathbb{E}_o[\cdot] := \mathbb{E}[\cdot \mid O = o]$ for expectation conditional on the observation pattern. Under squared loss, define

$$\begin{aligned} \mathcal{R}_{\text{sel}}(S, o) &= \inf_g \mathbb{E}_o \|\mathbf{T} - g(U)\|_2^2, \\ \mathcal{R}_{\text{full}}(S, o) &= \inf_h \mathbb{E}_o \|\mathbf{T} - h(U, W)\|_2^2. \end{aligned} \tag{2}$$

Proposition 1 (Population predictive information gap)**.** *If* $\mathbb{E}_o\|\mathbf{T}\|_2^2 < \infty$*, then*

$$\mathcal{R}_{\text{sel}}(S, o) - \mathcal{R}_{\text{full}}(S, o) = \mathbb{E}_o \left\|\mathbb{E}_o[\mathbf{T} \mid U, W] - \mathbb{E}_o[\mathbf{T} \mid U]\right\|_2^2 \geq 0. \tag{3}$$

The proof follows from the conditional-expectation projection identity (Appendix E.1). Proposition 1 quantifies, for fixed S and o, the population information loss induced by restricting prediction to the selected observed features. This proposition does not imply that using only the selected features yields higher predictive accuracy when richer observed information remains available. Rather, it quantifies the additional predictive risk incurred by restricting prediction to the selected observed features, relative to using all observed information.

4.2 PROBABILISTIC POSTERIOR MODELING

For sample i, let $\mathcal{V}_i$ denote the available views used for inference and let $\mathbf{z}_i$ be its shared latent variable. Each encoder f_v maps a view-specific input $\mathbf{u}_i^{(v)}$ to Gaussian parameters $(\boldsymbol{\mu}_{iv}, \boldsymbol{\sigma}_{iv}^2)$, where $\mathbf{u}_i^{(v)}$ denotes either the full view input or its selected-feature counterpart, depending on the inference path. This yields $q_v(\mathbf{z}_i \mid \mathbf{u}_i^{(v)}) = \mathcal{N}(\boldsymbol{\mu}_{iv}, \text{diag}(\boldsymbol{\sigma}_{iv}^2))$.

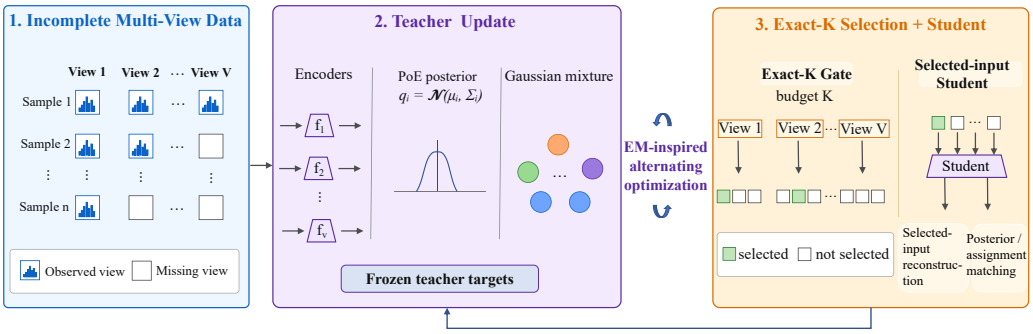


Figure 1: Overview of SC-MVFS. All currently observed features are used to construct probabilistic teacher targets, while the selected-input student performs posterior inference and reconstruction using only the currently selected features under randomly masked view patterns. After training, the final selected subset is fixed, and both posterior inference and missing-view completion use only the selected features.

Combining the available view-specific experts with a unit Gaussian prior through product-of-experts (PoE) fusion (Hinton, 2002; Wu & Goodman, 2018) gives

$$q_i(\mathbf{z}_i) \propto \mathcal{N}(\mathbf{z}_i; \mathbf{0}, \boldsymbol{I}) \prod_{v \in \mathcal{V}_i} q_v(\mathbf{z}_i \mid \mathbf{u}_i^{(v)}), \qquad q_i(\mathbf{z}_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{4}$$

With posterior precision $\boldsymbol{\Lambda}_i = \boldsymbol{\Sigma}_i^{-1}$,

$$\boldsymbol{\Lambda}_i = \boldsymbol{I} + \sum_{v \in \mathcal{V}_i} \operatorname{diag}(\boldsymbol{\sigma}_{iv}^{-2}), \qquad \boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i^{-1}, \qquad \boldsymbol{\mu}_i = \boldsymbol{\Sigma}_i \sum_{v \in \mathcal{V}_i} \operatorname{diag}(\boldsymbol{\sigma}_{iv}^{-2}) \boldsymbol{\mu}_{iv}. \tag{5}$$

An unavailable view contributes no expert, so the posterior uncertainty naturally reflects the available-view pattern.

We place a C-component diagonal Gaussian mixture over $\mathbf{z}$, where C is the known number of clusters. With mixture weights π_c, means $\boldsymbol{\nu}_c$ and diagonal covariances $\boldsymbol{\Sigma}_c$, the responsibility of component c for sample i is

$$r_{ic} \propto \pi_c \exp\left[-\frac{1}{2}\left(\log|\boldsymbol{\Sigma}_c| + \operatorname{tr}(\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\Sigma}_i) + (\boldsymbol{\mu}_i - \boldsymbol{\nu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\nu}_c)\right)\right], \tag{6}$$

with normalization over c.

4.3 Teacher Posterior and Mixture Update

At round t, with the neural parameters and selector fixed, we infer the teacher posterior $q_i^t = \mathcal{N}(\boldsymbol{\mu}_i^t, \boldsymbol{\Sigma}_i^t)$ from all currently observed features. Using the pre-refresh mixture parameters, we first compute preliminary responsibilities $\widetilde{r}_{ic}^t$ according to Formula 6 and then update the mixture from the posterior moments:

$$\begin{aligned} N_c^t &= \sum_i \widetilde{r}_{ic}^t, \quad \pi_c^t = \frac{N_c^t}{n}, \quad \boldsymbol{\nu}_c^t = \frac{1}{N_c^t}\sum_i \widetilde{r}_{ic}^t \boldsymbol{\mu}_i^t, \\ \operatorname{diag}(\boldsymbol{\Sigma}_c^t) &= \frac{1}{N_c^t}\sum_i \widetilde{r}_{ic}^t \left[\operatorname{diag}(\boldsymbol{\Sigma}_i^t) + (\boldsymbol{\mu}_i^t - \boldsymbol{\nu}_c^t)^{\odot 2}\right]. \end{aligned} \tag{7}$$

We then recompute the responsibilities r_{ic}^t using the updated mixture, let $\mathbf{r}_i^t = (r_{i1}^t, \ldots, r_{iC}^t)^\top$ and freeze the posterior and soft assignments as teacher targets:

$$\bar{q}_i^{\mathrm{T}} = \operatorname{sg}(q_i^t), \qquad \bar{\mathbf{r}}_i^{\mathrm{T}} = \operatorname{sg}(\mathbf{r}_i^t), \tag{8}$$

where $\operatorname{sg}(\cdot)$ denotes the stop-gradient operator. The updated mixture and teacher targets remain fixed during the subsequent neural update. Here diag extracts a matrix diagonal in Formula 7; applied to a vector in Formula 5, it constructs a diagonal matrix. The exponent $\odot 2$ denotes elementwise squaring.

4.4 BUDGET-CONSTRAINED FEATURE SELECTION

To enforce an exact global feature budget, we parameterize hierarchical view and within-view feature scores. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_V)^\top$ denote the view logits and $\mathbf{s}_v \in \mathbb{R}^{d_v}$ the feature logits of view v:

$$p_v = \mathrm{softmax}(\boldsymbol{\alpha}/T_{\mathrm{view}})_v, \qquad p_{vj} = p_v \,\mathrm{softmax}(\mathbf{s}_v)_j, \qquad a_{vj} = \log p_{vj}. \tag{9}$$

Here p_v controls the relative budget of view v, while a_{vj} gives the global score of feature j in that view. During training, we perturb the view and feature scores with independent standard Gumbel variables:

$$\widetilde{p}_v = \mathrm{softmax}\left(\frac{\boldsymbol{\alpha} + \gamma_s \mathbf{g}^{\mathrm{view}}}{T_{\mathrm{view}}}\right)_v, \qquad \widetilde{a}_{vj} = a_{vj} + \gamma_s g_{vj}, \tag{10}$$

where γ_s controls the perturbation magnitude at neural-update step s. These perturbations enable stochastic exploration of view allocations and within-view feature rankings.

We convert $K\widetilde{p}_v$ into integer per-view budgets K_v using largest-remainder allocation subject to view capacities, and retain the top-K_v features in each view, ensuring $0 \le K_v \le d_v, \sum_{v=1}^{V} K_v = K$. The exact allocation rule and surrogate-gradient details are given in Appendix D.3.

To optimize the discrete selector, we construct a global soft surrogate:

$$m_{vj}^{\mathrm{soft}} = \mathrm{sigmoid}\left(\frac{\widetilde{a}_{vj} - \tau}{T_{\mathrm{feat}}}\right), \qquad \sum_{v,j} m_{vj}^{\mathrm{soft}} = K, \tag{11}$$

where τ is determined by bisection and detached during backpropagation. Using the concatenated hard and soft masks, the straight-through estimator (Bengio et al., 2013)

$$\mathbf{m} = \mathrm{sg}(\mathbf{m}_{\mathrm{hard}} - \mathbf{m}_{\mathrm{soft}}) + \mathbf{m}_{\mathrm{soft}} \tag{12}$$

uses the exactly K-sparse hard mask in the forward pass and propagates gradients through the soft surrogate. The soft surrogate enforces only the global mass K; exact per-view allocation is imposed by the hard mask, with no gradient through K_v, largest-remainder allocation, or per-view top-K_v.

4.5 SELECTION-CONSISTENT COMPLETION

During training, the student performs posterior inference and reconstruction using only the features admitted by the hard mask. To expose the student to different incomplete-view patterns, we randomly hide observed views while retaining at least one. Let $\widetilde{o}_{iv} \le o_{iv}$ denote the resulting observation pattern and let g_v denote the decoder of view v.

Let $Q(\cdot)$ denote the PoE posterior map defined in Section 4.2. For a retained view, the selected student input is $\widetilde{\mathbf{x}}_i^{(v)} = \mathbf{m}^{(v)} \odot \mathbf{x}_i^{(v)}$. Here, $\widetilde{\mathbf{x}}_i^{(v)}$ denotes the masked selected-feature input used by the student during training. We distinguish the current full-input posterior from the selected-input student posterior as

$$q_i^{\mathrm{full}} = Q\Big(\{\mathbf{x}_i^{(v)} : o_{iv} = 1\}\Big), \qquad q_i^{\mathrm{sel}} = Q\Big(\{\widetilde{\mathbf{x}}_i^{(v)} : \widetilde{o}_{iv} = 1\}\Big). \tag{13}$$

The posterior means are $\boldsymbol{\mu}_i^{\mathrm{full}}$ and $\boldsymbol{\mu}_i^{\mathrm{sel}}$, respectively. Their corresponding responsibilities are $\mathbf{r}_i^{\mathrm{full}}$ and $\mathbf{r}_i^{\mathrm{sel}}$. The frozen teacher targets $\bar{q}_i^{\mathrm{T}}$ and $\bar{\mathbf{r}}_i^{\mathrm{T}}$ are obtained at the beginning of each alternating round as described in Section 4.3. In contrast, q_i^{full} and q_i^{sel} are computed using the current neural parameters during the subsequent updates.

The neural-update objective is

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask-rec}} + \lambda_{\mathrm{post}}\mathcal{L}_{\mathrm{post}} + \lambda_{\mathrm{assign}}\mathcal{L}_{\mathrm{assign}} \\ & + \lambda_{\mathrm{mix}}\mathcal{L}_{\mathrm{mix}} + \lambda_{2,1}\mathcal{L}_{2,1}.\end{aligned} \tag{14}$$

Let $\mathcal{O}_v = \{i : o_{iv} = 1\}$ denote the samples for which view v is naturally observed. The full-input and selected-input reconstruction losses are

$$\begin{aligned}\mathcal{L}_{\mathrm{rec}} &= \frac{1}{V}\sum_{v=1}^{V}\frac{1}{|\mathcal{O}_v| d_v}\sum_{i\in\mathcal{O}_v}\left\|\mathbf{x}_i^{(v)} - g_v(\boldsymbol{\mu}_i^{\mathrm{full}})\right\|_2^2, \\ \mathcal{L}_{\mathrm{mask-rec}} &= \frac{1}{V}\sum_{v=1}^{V}\frac{1}{|\mathcal{O}_v| d_v}\sum_{i\in\mathcal{O}_v}\left\|\mathbf{x}_i^{(v)} - g_v(\boldsymbol{\mu}_i^{\mathrm{sel}})\right\|_2^2.\end{aligned} \tag{15}$$

The first provides auxiliary reconstruction supervision from all currently observed features, while the second trains reconstruction from selected features under the reduced observation pattern. Because all naturally observed views are used as reconstruction targets, a retained source view may also serve as a target; thus the selected-input loss combines same-view reconstruction and cross-view reconstruction.

We further align the selected-input student with the frozen teacher targets at both the posterior and cluster-assignment levels:

$$\mathcal{L}_{\text{post}} = \frac{1}{n}\sum_i \mathrm{KL}\big(q_i^{\text{sel}} \| \bar{q}_i^{\text{T}}\big)\,, \qquad \mathcal{L}_{\text{assign}} = \frac{1}{n}\sum_i \mathrm{KL}\big(\bar{\mathbf{r}}_i^{\text{T}} \| \mathbf{r}_i^{\text{sel}}\big)\,. \tag{16}$$

The posterior loss aligns the selected-input posterior with the teacher distribution, while the assignment loss transfers soft cluster structure.

The current full-input posterior is additionally regularized by the refreshed latent mixture:

$$\mathcal{L}_{\text{mix}} = \frac{1}{n}\sum_i \sum_c r_{ic}^{\text{full}} \left[\mathrm{KL}\big(q_i^{\text{full}} \| \mathcal{N}(\boldsymbol{\nu}_c, \boldsymbol{\Sigma}_c)\big) + \log \frac{r_{ic}^{\text{full}}}{\pi_c}\right]. \tag{17}$$

This term encourages compatibility between the current full-input posterior and the refreshed Gaussian mixture.

We also use the auxiliary group regularizer $\mathcal{L}_{2,1} = \sum_{v,j} \left\| [\boldsymbol{W}_{\text{in}}^{(v)}]_{:,j} \right\|_2$, where $\boldsymbol{W}_{\text{in}}^{(v)}$ denotes the input weight matrix of the v-th view encoder. The $\ell_{2,1}$ regularizer encourages group-level sparsity (Nie et al., 2010), whereas the exact feature count is enforced by the hard forward gate in Section 4.4. Training alternates between the frozen teacher/mixture refresh in Section 4.3 and stochastic selected-input student updates. The complete training procedure is given in Appendix D.

4.6 Inference and Completion from Selected Features

After training, Gumbel perturbations are removed and the deterministic exact-K feature mask is fixed. For sample i, posterior inference uses only the selected features in its naturally observed views:

$$q_i^{\text{sel}} = Q\Big(\Big\{\mathbf{m}^{(v)} \odot \mathbf{x}_i^{(v)} : o_{iv} = 1\Big\}\Big) = \mathcal{N}(\boldsymbol{\mu}_i^{\text{sel}}, \boldsymbol{\Sigma}_i^{\text{sel}}). \tag{18}$$

Thus, features excluded by the final mask do not affect posterior inference.

Naturally observed views are copied directly, whereas missing views are reconstructed from the selected-input posterior:

$$\widehat{\mathbf{x}}_i^{(v)} = \begin{cases} \mathbf{x}_i^{(v)}, & o_{iv} = 1, \\ g_v(\boldsymbol{\mu}_i^{\text{sel}}), & o_{iv} = 0. \end{cases} \tag{19}$$

Here, $\widehat{\mathbf{x}}_i^{(v)}$ denotes the observed-or-completed view used to construct the final representation.

The final representation retains only the selected features of each observed or completed view:

$$\mathbf{y}_i = \mathrm{concat}_{v=1}^{V} [\widehat{\mathbf{x}}_i^{(v)}]_{\mathcal{S}_v} \in \mathbb{R}^K, \tag{20}$$

where $\mathcal{S}_v$ denotes the selected feature indices of view v and $\sum_{v=1}^{V} |\mathcal{S}_v| = K$. Stacking the sample representations gives $\boldsymbol{Y} = [\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top]^\top \in \mathbb{R}^{n \times K}$. Thus, the same selected subset governs posterior inference, missing-view completion, and the final representation. The corresponding architectural invariance property is established in Appendix E.2.

5 Experiments

We organize the experimental evaluation around four research questions. **RQ1: Clustering effectiveness.** How effectively does SC-MVFS perform relative to existing incomplete multi-view feature-selection methods under the central missingness and feature-budget setting? **RQ2: Robustness.** How does its clustering performance change as the missing-view rate and feature budget vary? **RQ3: Selection–completion consistency.** What happens when a completion model is trained with richer observed features but is restricted to selected features during final completion? Can training

Table 1: Main clustering results at $\mathrm{MR} = \mathrm{FR} = 0.3$ (%, reported as mean $\pm$ sample standard deviation over five random seeds). Average ranks are computed within each dataset and then averaged across datasets.

Metric	Dataset	SC-MVFS	TIME-FS	C2IMUFS	UNIFIER	CVFS	CLSG	TERUIMUFS	UIMUFSLR	Mean+CAE	Random-K
ACC	3Sources	38.07±3.80	**43.41±7.04**	34.67±1.84	41.19±3.94	36.74±2.99	37.04±1.57	33.93±3.20	34.81±0.52	36.44±2.69	36.59±3.17
	100Leaves	**44.44±0.88**	34.19±1.28	33.67±0.59	41.73±0.52	34.17±1.72	29.92±0.70	34.55±0.83	40.52±1.91	31.64±1.74	33.58±1.07
	BBCSport	**48.05±5.83**	35.59±0.94	41.10±2.94	38.30±3.62	40.69±3.08	39.68±5.30	33.66±4.79	43.68±3.09	38.11±6.26	36.37±1.43
	BDGP	**73.92±6.12**	57.00±1.77	35.68±2.21	33.65±2.79	35.87±3.18	48.93±5.68	26.71±3.30	56.53±3.73	41.40±5.01	46.00±2.89
	Caltech101-7	**51.11±3.36**	50.55±5.89	36.68±4.90	37.91±1.00	37.68±4.32	40.25±2.75	36.62±0.74	43.72±1.14	38.64±1.19	36.83±1.97
	CiteSeer	**36.25±4.67**	32.93±8.15	33.40±2.82	23.22±0.92	33.31±3.93	30.44±2.31	27.69±7.83	33.35±2.77	29.80±4.65	26.76±3.83
	Handwritten	**89.56±5.72**	41.86±2.10	48.75±4.72	82.91±1.57	49.01±2.73	53.86±4.43	21.26±2.50	55.23±3.16	41.04±1.47	40.09±7.60
	OutdoorScene	**63.29±1.82**	28.85±1.38	42.25±0.38	40.99±2.69	42.28±0.31	42.54±1.18	35.17±0.82	40.01±4.20	29.78±1.83	26.15±3.76
	Average rank	**1.25**	5.00	6.00	5.25	5.13	4.88	8.63	3.88	7.13	7.88
NMI	3Sources	11.60±3.60	**20.57±7.87**	7.84±1.40	17.36±6.70	10.77±1.60	10.43±1.68	7.67±2.41	8.06±1.33	12.26±3.68	12.51±2.49
	100Leaves	**69.63±0.98**	62.49±1.38	62.02±0.83	66.60±0.80	61.86±0.72	61.70±0.95	63.46±0.62	65.97±1.30	60.34±1.31	61.46±0.72
	BBCSport	**24.09±5.49**	4.46±2.69	11.88±4.59	8.87±2.79	13.01±7.22	10.80±4.15	4.84±3.41	16.71±4.67	8.77±5.81	4.36±2.65
	BDGP	**55.90±5.97**	31.37±1.99	15.61±3.36	10.35±3.41	15.38±3.67	29.65±3.96	3.73±1.97	39.07±4.32	31.35±2.54	34.06±2.82
	Caltech101-7	**57.66±0.56**	36.05±0.57	40.68±3.80	34.39±2.28	40.07±2.82	31.68±4.11	22.72±0.95	49.40±1.83	24.68±3.96	26.27±2.41
	CiteSeer	**15.08±3.89**	12.13±7.17	12.93±3.33	3.38±0.63	12.81±4.15	9.72±3.31	6.81±6.91	12.05±1.71	9.11±3.85	7.12±4.07
	Handwritten	**84.22±2.03**	44.24±1.94	45.68±2.35	71.94±0.82	45.38±0.63	53.53±4.41	13.98±4.17	52.46±2.40	40.79±1.35	40.16±4.58
	OutdoorScene	**52.12±0.91**	14.57±1.16	33.43±0.22	30.65±0.78	33.38±0.19	30.74±1.02	25.77±0.44	30.67±3.85	20.95±2.00	13.60±4.25
	Average rank	**1.50**	5.50	4.75	5.38	5.00	5.75	8.50	3.88	7.25	7.50

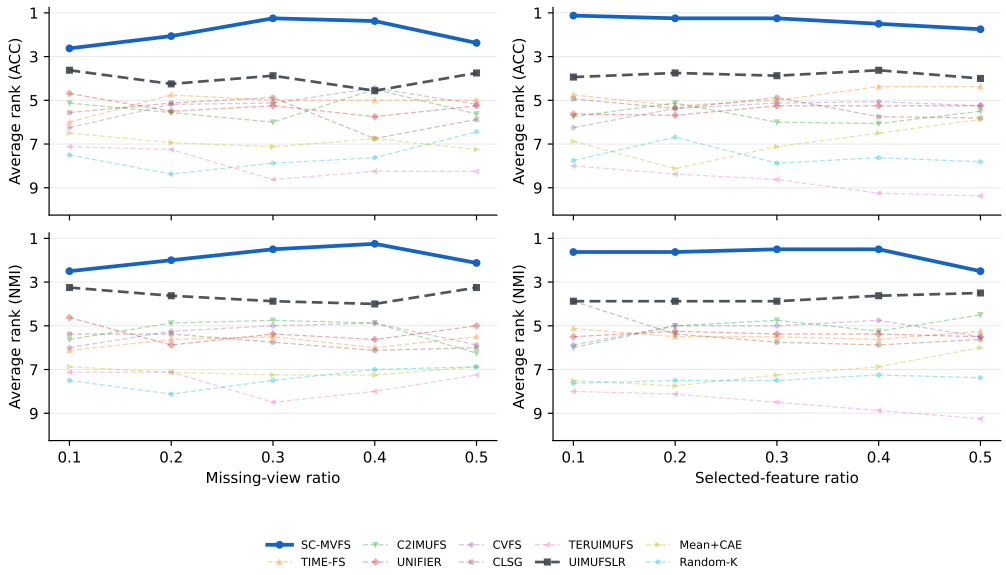


Figure 2: Average ACC and NMI ranks across the eight real-world datasets under varying missing-view rates and feature budgets (lower is better). The missing-view rate is varied with FR = 0.3, while the retained-feature ratio is varied with MR = 0.3. Methods are ranked within each dataset and then averaged across the eight datasets.

with the same selected-feature input reduce this discrepancy? **RQ4: Supporting mechanisms and controlled behavior.** How do the main training components contribute to SC-MVFS, how sensitive is its training behavior to key optimization choices, and under what controlled data and view-pattern conditions does selection–completion consistency become more important?

5.1 Experimental Setup

We evaluate SC-MVFS on eight public datasets; statistics are provided in Appendix B. Using missing-view ratio (MR) and retained-feature ratio (FR), we adopt a cross-design with $\mathrm{FR} = 0.3$ and $\mathrm{MR} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, together with $\mathrm{MR} = 0.3$ and $\mathrm{FR} \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, yielding nine conditions. Protocol details are provided in Appendix A. We compare SC-MVFS with TIME-FS, C2IMUFS, UNIFIER, CVFS, CLSG, TERUIMUFS, UIMUFSLR, mean imputation with Concrete Autoencoder and Random-K. Baseline details are provided in Appendix C. The code is available at `https://anonymous.4open.science/r/SC-MVFS-4768`.

5.2 Main Results

We address RQ1 by comparing SC-MVFS with the baselines under the central setting $(\mathrm{MR}, \mathrm{FR}) = (0.3, 0.3)$. As shown in Table 1, SC-MVFS achieves the best ACC and NMI on seven of the eight datasets and obtains the lowest average ranks of 1.25 and 1.50, respectively. Specifically, SC-MVFS achieves an NMI of 55.90% on BDGP, compared with 39.07% for the second-best method; 84.22% on Handwritten, compared with 71.94%; and 52.12% on OutdoorScene, compared with 33.43%. It also achieves the best ACC and NMI on 100Leaves, BBCSport, Caltech101-7 and CiteSeer. TIME-FS performs best on 3Sources, where SC-MVFS ranks third in ACC and fifth in NMI. Overall, SC-MVFS achieves the best results on most datasets and the lowest average rank under the central missingness and feature-budget setting.

5.3 Robustness to Missingness and Feature Budget

We address RQ2 by evaluating the robustness of SC-MVFS to different missing-view rates and feature budgets, while keeping the hyperparameters selected under the central setting fixed across all tested conditions. Figure 2 reports the average ACC and NMI ranks when varying the missing-view rate at $\mathrm{FR} = 0.3$ and the feature ratio at $\mathrm{MR} = 0.3$. SC-MVFS achieves the lowest average rank under every tested condition in both sweeps, showing that its relative clustering advantage is not limited to the central setting $(\mathrm{MR}, \mathrm{FR}) = (0.3, 0.3)$. Overall, SC-MVFS maintains favorable cross-dataset average rankings across different levels of missingness and feature budgets. Complete per-dataset results are reported in Appendix F, including five-seed means, standard-deviation bands and the complete average-rank results for all tested conditions.

5.4 Selection–Completion Consistency

We evaluate RQ3 at $\mathrm{MR} = \mathrm{FR} = 0.3$ using two complementary comparisons: a fixed-subset diagnostic that isolates training–completion input mismatch, and a joint control that evaluates the effect of reconstruction inputs when the selector is learned end to end. Table 2 summarizes both comparisons.

Table 2: Completion evidence at $\mathrm{MR} = \mathrm{FR} = 0.3$. The first three columns report pure-completion MSE with a fixed subset, and the last column reports the selected-input minus full-input-control RMSE under joint training. Lower is better, negative ΔRMSE favors selected-input training.

	Pure completion: MSE			Joint training
Dataset	Full→Full	Full→Selected	Selected→Selected	ΔRMSE
3Sources	1.8072 ± 0.2510	1.8144 ± 0.2574	1.8133 ± 0.2548	-0.0012
100Leaves	0.7254 ± 0.0592	1.2963 ± 0.0947	0.8352 ± 0.0618	-0.0374
BBCSport	2.4371 ± 0.1028	2.4469 ± 0.1080	2.4420 ± 0.1025	-0.0069
BDGP	1.0530 ± 0.0670	1.1470 ± 0.0540	1.0788 ± 0.0542	-0.0137
Caltech101-7	0.5324 ± 0.0503	0.6785 ± 0.0579	0.5455 ± 0.0404	-0.0008
CiteSeer	0.9624 ± 0.0265	0.9700 ± 0.0280	0.9646 ± 0.0292	-0.0015
Handwritten	0.3462 ± 0.0121	0.5800 ± 0.0292	0.3672 ± 0.0132	$+0.0002$
OutdoorScene	0.4811 ± 0.0150	0.6162 ± 0.0271	0.4911 ± 0.0214	-0.0058

Pure-completion diagnostic. The first three columns of Table 2 report a controlled completion diagnostic using the same fixed feature subset, with the learned selector and other joint-training components removed. We compare Full→Full, Full→Selected, and Selected→Selected under otherwise matched settings. As expected, Full→Full achieves the lowest MSE because richer inputs are available at both training and completion. The key constrained comparison is therefore Full→Selected versus Selected→Selected: training directly with selected inputs yields lower MSE on all eight datasets. Metric definitions and the complete protocol are provided in Appendix G.2.

Joint reconstruction control. The last column of Table 2 further compares two reconstruction settings within the full SC-MVFS training framework. The two settings use the same available views and retain all other training objectives, differing only in the features used by the reconstruction branch: one uses only the selected features, whereas the other uses all observed features from the same views.

After training, both settings perform completion using only the final selected features. On 7 of the 8 datasets, the selected-feature reconstruction setting achieves lower mean RMSE, suggesting that matching the reconstruction inputs used during training to those used for final completion can be beneficial. Complete experimental settings and per-dataset results are provided in Appendix G.3.

5.5 Mechanism Analysis and Controlled Studies

We address RQ4 through component ablations, sensitivity, optimization and efficiency analyses, controlled synthetic experiments and view-pattern controls.

Component ablations. The full model achieves the best aggregate ACC and NMI ranks across the ablation variants, although individual components are not uniformly beneficial on every dataset or metric. These results support their complementary roles at the aggregate level. Complete per-dataset results are reported in Appendix H.

Sensitivity, optimization and efficiency. The frozen hyperparameter defaults remain competitive across the tested settings. Among the tested numbers of alternating rounds, eight rounds achieve the best average ACC and NMI ranks. We additionally report computational cost, memory usage, and the number of alternating rounds to characterize the optimization and efficiency of SC-MVFS. Complete results are provided in Appendix I.

Controlled synthetic experiments. We further use controlled synthetic data to examine when selection–completion consistency becomes more important. The synthetic data include class-independent shortcut features that distribute cross-view predictive information across many candidate features. Over 20 paired synthetic datasets, the NMI advantage of SC-MVFS over a matched variant that uses all available features for completion during training but only selected features during post-selection completion increases from 3.56 to 11.51 percentage points. This result suggests that selection–completion mismatch becomes more harmful as useful cross-view information becomes increasingly distributed. Complete results are provided in Appendix J.

View-pattern controls. We use three variants to separate the effects of additional view hiding during training and additional naturally observed views used by the teacher. Randomly hiding more views lowers completion RMSE on all eight datasets and improves NMI on six. When the student's reduced view pattern is fixed, allowing the teacher to use all naturally observed views improves NMI on seven of the eight datasets, but lowers RMSE on only three. These results suggest that additional view hiding mainly benefits completion, whereas using more observed views in the teacher more consistently benefits clustering. Complete results are reported in Appendix K.

6 Conclusion

SC-MVFS addresses incomplete multi-view feature selection through selection–completion consistency, aligning selected-input training with the information admitted by the final selected-feature interface. A full-input teacher provides auxiliary supervision, while student inference and completion use only selected features. The method combines exact-budget feature selection with probabilistic multi-view modeling.

Conditional-risk analysis characterizes the difference in optimal prediction error between selected and full observed features, while controlled comparisons show that selected-input training can reduce completion error when final completion uses only selected features. Across eight real-world datasets, SC-MVFS achieves strong clustering performance under different missing-view rates and feature budgets. Ablation, view-pattern and synthetic experiments clarify component roles and show larger consistency benefits when useful cross-view information is distributed across many candidate features. Future work may consider richer posterior dependencies, flexible missingness mechanisms and adaptive feature budgets.

AI USE STATEMENT

We used generative AI tools to assist with checking language, notation, references, and formatting, identifying potentially relevant literature for the related work section, and polishing the text to improve readability. All AI-assisted checking, literature-search support, and polishing results were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including the adoption and modification of the aforementioned results.

REPRODUCIBILITY STATEMENT

We provide the complete model specification, optimization procedure, and evaluation protocol in the main paper and appendices. Dataset statistics, preprocessing, missing-view generation, hyperparameter settings, and additional per-dataset results are documented in the appendix. The source code is available through the anonymized repository link provided in the paper.

REFERENCES

Muhammed Fatih Balın, Abubakar Abid, and James Zou. Concrete autoencoders: Differentiable feature selection and reconstruction. In *International conference on machine learning*, pp. 444–453. PMLR, 2019.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, 2012.

Robert Duin. Multiple features. UCI Machine Learning Repository, 1998. URL `https://archive.ics.uci.edu/dataset/72/multiple+features`. Dataset.

Xin Gao and Jian Pu. Deep incomplete multi-view learning via cyclic permutation of VAEs. In *International Conference on Learning Representations*, 2025.

Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pp. 377–384, 2006.

Derek Greene and Pádraig Cunningham. A matrix factorization approach for integrating multiple data views. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 423–438. Springer, 2009.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Yanyong Huang, Zongxin Shen, Yuxin Cai, Xiuwen Yi, Dongjie Wang, Fengmao Lv, and Tianrui Li. C2IMUFS: Complementary and consensus learning-based incomplete multi-view unsupervised feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10681–10694, 2023. doi: 10.1109/TKDE.2023.3266595.

Yanyong Huang, Zongxin Shen, Tianrui Li, and Fengmao Lv. Unified view imputation and feature selection learning for incomplete multi-view data. *arXiv preprint arXiv:2401.10549*, 2024.

Yanyong Huang, Minghui Lu, Wei Huang, Xiuwen Yi, and Tianrui Li. Time-fs: joint learning of tensorial incomplete multi-view unsupervised feature selection and missing-view imputation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17503–17510, 2025.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

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.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Charles Mallah, James Cope, James Orwell, et al. Plant leaf classification using probabilistic integration of shape, texture and margin features. *Signal Processing, Pattern Recognition and Applications*, 5(1):45–54, 2013.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *Icml*, volume 11, pp. 689–696, 2011.

Feiping Nie, Heng Huang, Xiao Cai, and Chris H. Q. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *Advances in Neural Information Processing Systems*, volume 23, pp. 1813–1821, 2010.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 47(2):742–754, 2024.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multimodal deep generative models. *Advances in neural information processing systems*, 32, 2019.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal ELBO. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=5Y21V0RDBV`.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pp. 1073–1080, 2009.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

Yuanyuan Xu, Yu Yin, Jun Wang, Jinmao Wei, Jian Liu, Lina Yao, and Wenjie Zhang. Unsupervised cross-view feature selection on incomplete data. *Knowledge-Based Systems*, 234:107595, 2021.

Xuanhao Yang, Hangjun Che, Man-Fai Leung, and Shiping Wen. Unbalanced incomplete multiview unsupervised feature selection with low-redundancy constraint in low-dimensional space. *IEEE Transactions on Industrial Informatics*, 21(3):2679–2688, 2024.

Xuanhao Yang, Hangjun Che, and Man-Fai Leung. Tensor-based unsupervised feature selection for error-robust handling of unbalanced incomplete multi-view data. *Information Fusion*, 114:102693, 2025.

Hong-Wei Yu, Jun-Yun Wu, Jian-Sheng Wu, and Weidong Min. Confident local similarity graphs for unsupervised feature selection on incomplete multi-view data. *Knowledge-Based Systems*, 316: 113369, 2025.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

A Reproducibility and Evaluation Protocol

Data split and model selection. For each dataset, we use a 20/80 development/evaluation split. The development subset is used only for configuration selection under the central setting $\mathrm{MR} = \mathrm{FR} = 0.3$. Once the hyperparameters are selected, they are kept fixed for all subsequent control experiments. The main evaluation and joint-model control experiments adopt a transductive setting on the 80% evaluation subset. For SC-MVFS, the main objective weights are fixed globally, and only $\lambda_{2,1} \in \{10^{-5}, 10^{-3}\}$ is selected on the development subset.

SC-MVFS implementation details. Each view encoder and decoder uses one hidden layer with 128 units, and the shared latent dimension is 32. We optimize SC-MVFS using Adam (Kingma & Ba, 2014) with learning rate 10^{-3}, weight decay 10^{-5}, batch size 128 and gradient clipping at 5. Training uses five warm-up epochs, at most eight alternating rounds and three neural-update epochs per round. The stopping tolerance is 10^{-3} with patience two. We set the feature-surrogate and view-allocation temperatures to $T_{\mathrm{feat}} = 0.5$ and $T_{\mathrm{view}} = 1.0$, respectively, and anneal the Gumbel noise scale from 1.0 to 0.05. During selected-input training, each actually available view is additionally hidden with probability 0.35, while at least one observed view is retained for every sample. If independent hiding would remove every available view, one originally observed view is restored uniformly at random. This additional masking is applied only as a training augmentation after the benchmark observation pattern has been generated. The loss weights are fixed to $\lambda_{\mathrm{rec}} = \lambda_{\mathrm{mask}} = 1$, $\lambda_{\mathrm{post}} = \lambda_{\mathrm{assign}} = 0.5$, $\lambda_{\mathrm{mix}} = 0.1$, with $\lambda_{2,1}$ selected as described above. Mixture covariances are diagonal with a floor of 10^{-4}; component masses and mixture weights are floored at 10^{-3} and 10^{-8}, respectively, before renormalization for numerical stability. These default settings are kept fixed across the subsequent experiments, and their sensitivity is further examined in Appendix I.

Preprocessing and Evaluation. For all real-data experiments, each view is independently transformed by feature-wise min–max normalization. The minimum and range of each feature are estimated using only the samples for which that view is observed. MR denotes the fraction of sample–view entries removed from the complete observation matrix. FR denotes the retained-feature fraction over all views. For a fixed dataset, MR, and evaluation repetition, all methods use the same observation mask. The mask is also reused across different feature-budget settings. Using each fitted method's selected subset and completed views, we concatenate the views, retain the selected features, and apply K-means with the known number of clusters and 20 random initializations. ACC is computed after resolving the cluster–label permutation using the Hungarian algorithm. NMI uses arithmetic normalization following Vinh et al. (2009). All methods use the same development/evaluation split, missingness mask, feature budget, and clustering evaluator.

The feature selectors retain their method-specific treatment of incomplete views. For the common clustering evaluator, we use a method's recovered views when it exposes an explicit raw-feature completion: TIME-FS uses its learned $\widehat{\mathbf{X}}$, UNIFIER uses its jointly recovered views, CVFS uses its dynamic completion rule, CLSG uses its generator-based recovery, and TERUIMUFS uses its tensor-based completion. C2IMUFS reconstructs similarity structures rather than a complete raw-feature matrix, and UIMUFSLR likewise does not expose a complete matrix in the original feature space. We therefore apply the same observed-feature mean adapter to these two methods. Mean+CAE and Random-K also use observed-feature mean completion.

B Datasets

3Sources contains news stories collected from BBC, Reuters and The Guardian, with each source providing a distinct textual view (Greene & Cunningham, 2009). We use the subset of 169 stories reported by all three sources, covering six topical categories.

100Leaves contains 1,600 leaf images from 100 plant species, with 16 samples per species. Each image is described by three 64-dimensional feature sets characterizing shape, margin and texture (Mallah et al., 2013).

Table 3: Statistics of the eight real-world multi-view datasets.

Datasets	# Samples	# Views	# Classes	# Features per view
3Sources	169	3	6	3560 / 3631 / 3068
100Leaves	1600	3	100	64 / 64 / 64
BBCSport	544	2	5	3183 / 3203
BDGP	2500	2	5	1750 / 79
Caltech101-7	1474	6	7	48 / 40 / 254 / 1984 / 512 / 928
CiteSeer	3312	2	6	3312 / 3703
Handwritten	2000	6	10	216 / 76 / 64 / 6 / 240 / 47
OutdoorScene	2688	4	8	512 / 432 / 256 / 48

BBCSport contains 544 sports news documents from the BBC Sport website covering five categories, represented by two complementary textual views (Greene & Cunningham, 2006).

BDGP contains 2,500 Drosophila embryo image bags from five categories, where each sample is represented by a 1,750-dimensional visual feature vector and a 79-dimensional textual feature vector (Cai et al., 2012).

Caltech101-7 contains 1,474 images from seven object categories and provides six hand-crafted visual representations, including Gabor, wavelet moments, CENTRIST, HOG, GIST and LBP descriptors (Li et al., 2015).

CiteSeer consists of 3,312 scientific documents from six research categories and represents each document using content and citation-network information, providing two heterogeneous views (Sen et al., 2008).

Handwritten contains 2,000 samples of handwritten digits from 10 classes and describes each digit using six complementary feature families, including profile, Fourier, Karhunen–Loève, morphological, pixel, and Zernike descriptors (Duin, 1998).

OutdoorScene contains 2,688 images from eight outdoor scene categories (Oliva & Torralba, 2001). In the multi-view benchmark used here, each image is represented using GIST, HOG, LBP and Gabor features.

Detailed statistics for all eight datasets are summarized in Table 3. Overall, these datasets span text, citation networks, biological images, handwritten digits, objects and natural scenes, with sample sizes from 169 to 3,312, view counts from 2 to 6 and total dimensionalities from 192 to 10,259. They therefore cover both low-dimensional multi-view settings and high-dimensional, heterogeneous feature-selection scenarios.

C Baselines and Comparison Details

Specialized incomplete multi-view feature selection baselines. We compare SC-MVFS with seven representative methods designed for incomplete multi-view unsupervised feature selection. They cover complementary and consensus learning, graph reconstruction, joint imputation and selection, tensor modeling, and redundancy-aware representation learning.

CVFS models both instance-wise graph relations and feature-wise cross-view interactions. Dynamically weighted estimators mitigate the effect of missing instances, while the learned cross-view structure guides the selection of discriminative and nonredundant features (Xu et al., 2021).

C2IMUFS jointly exploits complementary and consensus information across views. It reconstructs complete view-specific sample-similarity graphs and learns $\ell_{2,p}$-regularized feature-selection matrices within an extended weighted matrix-factorization framework (Huang et al., 2023).

UNIFIER unifies missing-view imputation and feature selection. Sample- and feature-space similarity graphs guide the recovery of missing views, while continuous feature weights are jointly optimized to identify informative features (Huang et al., 2024).

TIME-FS jointly models missing-view recovery, discriminative feature selection, and low-dimensional representation. Tensor decomposition captures cross-view correlations in the learned representations and guides both imputation and row-sparse feature selection (Huang et al., 2025).

CLSG learns intact latent representations from incomplete observations and maps them back to the data space to recover missing information. Confident cross-view and view-specific local similarity graphs then guide unsupervised feature selection (Yu et al., 2025).

TERUIMUFS combines tensor low-rank self-representation, missing-sample recovery, graph learning, and error-robust feature selection to handle corruption and imbalance in unbalanced incomplete multi-view data (Yang et al., 2025).

UIMUFSLR handles incomplete samples through an adaptively weighted unified graph and performs feature selection in a low-dimensional space with an explicit low-redundancy constraint (Yang et al., 2024).

Closest-method distinction. Several baselines couple feature selection with missing-data or incomplete-structure recovery. UNIFIER, TIME-FS, and CLSG explicitly recover missing view data while learning feature-selection variables. C2IMUFS reconstructs complete sample-similarity graphs, whereas TERUIMUFS recovers missing-sample structure through tensor low-rank self-representation. In their published objectives, the final discrete subset is extracted from continuous feature-selection scores, without explicitly requiring the recovery pathway to depend only on that deterministic subset.

In contrast, SC-MVFS imposes an explicit selected-feature information constraint. During selected-input training, the student posterior and reconstruction pathway receive only features admitted by the current selector. Final inference and completion use only the resulting deterministic selected subset. The full-input teacher supplies privileged training targets but is unavailable to the final completion pathway.

Generic baselines and reference methods. We additionally include two generic budget-matched comparisons. **Mean+CAE** imputes each missing view using feature-wise means estimated from observed samples and subsequently applies the Concrete Autoencoder (Balın et al., 2019) for differentiable feature selection. **Random-K** applies the same preprocessing and mean-imputation protocol and then uniformly samples the same total number of features as the compared methods.

D Algorithm and Complexity

D.1 Training and Inference Procedure

Algorithm 1 summarizes the complete training and inference procedure of SC-MVFS. Training alternates between teacher/mixture refresh and neural and selector updates. The teacher targets and mixture parameters are frozen during each neural-update stage and refreshed at the beginning of the next alternating round. After training, the stochastic gate is replaced by a deterministic exact-K subset for final inference and missing-view completion.

D.2 Complexity Analysis

Let $D = \sum_{v=1}^{V} d_v$, let h denote the latent dimension, and let H denote the hidden-layer width. A forward pass through the view-specific encoders and decoders has worst-case cost $O(nH(D + Vh))$. Under diagonal covariance, posterior fusion costs $O(nVh)$ and responsibility evaluation costs $O(nCh)$. The exact-budget selector requires $O(D \log D)$ in the worst case for feature ranking. In practice, the encoder and decoder computations dominate the overall cost. Under the central experimental setting, the recorded fit time ranges from 1.39 to 27.61 seconds and peak allocated GPU memory ranges from 36.16 to 161.44 MiB, excluding data loading and K-means evaluation.

D.3 Exact-Budget Allocation and Surrogate Gradients

Exact per-view allocation. Given the perturbed view proportions $\{\widetilde{p}_v\}_{v=1}^{V}$, we form $b_v = K\widetilde{p}_v, \sum_{v=1}^{V} b_v = K$. The initial integer allocation is $K_v = \min\{d_v, \lfloor b_v \rfloor\}$. If $\sum_v K_v < K$,

Algorithm 1 Training and Inference Procedure of SC-MVFS

Require: Incomplete views $\mathcal{X} = \{\boldsymbol{X}^{(v)}\}_{v=1}^{V}$, observation mask $\boldsymbol{O}$, cluster number C, feature budget K, maximum rounds $R_{\max}$, neural-update epochs E, stopping tolerance ϵ
Ensure: Selected feature sets $\{\mathcal{S}_v\}_{v=1}^{V}$ and representation $\boldsymbol{Y} \in \mathbb{R}^{n\times K}$

```
Initialize encoders, decoders and selector parameters
Warm up the encoders and decoders using L_rec
Initialize the Gaussian mixture from full-input posterior means
EM-inspired alternating optimization
for t = 1 to R_max do
    Infer teacher posteriors {q_i^t}_{i=1}^n from all currently observed features
    Compute {r̃_ic^t}_{i,c} and update {π_c^t, ν_c^t, Σ_c^t}_{c=1}^C using Formula 7
    Recompute {r_i^t}_{i=1}^n and freeze {q̄_i^T, r̄_i^T}_{i=1}^n using Formula 8
    for e = 1 to E do
        Construct the straight-through selector mask m (exact-K in the forward pass) and sample reduced observation masks Õ; update the neural and selector parameters by minimizing Formula 14
    end for
    if mean absolute responsibility change < ε for two consecutive rounds then
        break
    end if
end for
Deterministic inference
Remove Gumbel perturbations and obtain {S_v}_{v=1}^V with Σ_{v=1}^V |S_v| = K
Infer selected-input posteriors and complete missing views using Formulas 18–19
Construct Y using Formula 20
return {S_v}_{v=1}^V, Y
```

the remaining slots are assigned one at a time to views with remaining capacity, in descending order of the fractional remainders $b_v - \lfloor b_v \rfloor$. Under $K \le \sum_v d_v$, this procedure guarantees $0 \le K_v \le d_v, \sum_{v=1}^{V} K_v = K$. The forward selector retains the K_v features with the largest perturbed scores $\widetilde{a}_{vj}$ in each view, yielding an exactly K-sparse hard mask.

Surrogate gradient path. The integer budgets K_v, remainder allocation and per-view top-K_v operations are discrete and are not differentiated. Instead, the backward pass uses the global soft surrogate

$$m_{vj}^{\mathrm{soft}} = \mathrm{sigmoid}\left(\frac{\widetilde{a}_{vj} - \tau}{T_{\mathrm{feat}}}\right), \qquad \sum_{v,j} m_{vj}^{\mathrm{soft}} = K, \tag{21}$$

where τ is obtained by bisection and detached during backpropagation. The straight-through mask

$$\mathbf{m} = \mathrm{sg}(\mathbf{m}_{\mathrm{hard}} - \mathbf{m}_{\mathrm{soft}}) + \mathbf{m}_{\mathrm{soft}} \tag{22}$$

uses the exactly K-sparse hard mask in the forward pass and the soft surrogate in the backward pass.

Recall that $a_{vj} = \log p_v + \log \mathrm{softmax}(\mathbf{s}_v)_j$ and $\widetilde{a}_{vj} = a_{vj} + \gamma_s g_{vj}$. Treating the sampled Gumbel perturbation and τ as constants during backpropagation, the gradient with respect to a view logit α_u is

$$\frac{\partial m_{vj}^{\mathrm{soft}}}{\partial \alpha_u} = \frac{m_{vj}^{\mathrm{soft}}(1 - m_{vj}^{\mathrm{soft}})}{T_{\mathrm{feat}} T_{\mathrm{view}}} \left(\mathbf{1}[u = v] - p_u\right). \tag{23}$$

Thus, the view logits are optimized through the global soft surrogate; no gradient passes through the discrete counts K_v, remainder allocation, hard top-K_v operation, or the implicit dependence of τ on the scores. The backward pass is therefore a straight-through surrogate, while exact cardinality is guaranteed by the hard forward mask.

E Analysis of Selection–Completion Consistency

We first analyze the population risk gap in Proposition 1 using classical conditional expectation. We then establish the architectural invariance property of the final selected-input mapping. The risk

analysis characterizes the predictive information loss induced by restricting prediction to selected observed features under a fixed selected set and observation pattern, while the architectural property shows that, for fixed model parameters and selected sets, the final mapping is invariant to discarded features.

E.1 Completion-Risk Gap Conditional on Observed Views

Fix a selected feature set S and an observation pattern o with $\mathbb{P}(O = o) > 0$. For a target view absent from o, let $\mathbf{T} \in \mathbb{R}^q$ denote the features to be completed. Let U collect the selected features in the observed views and W the corresponding discarded features. We work under the conditional law given $O = o$, with expectation denoted by $\mathbb{E}_o$. The risks in Formula 2 are population risks over all square-integrable predictors.

Proof of Proposition 1. Let $a(U) = \mathbb{E}_o[\mathbf{T} \mid U]$ and $b(U, W) = \mathbb{E}_o[\mathbf{T} \mid U, W]$. For any admissible predictor $g(U)$,

$$\mathbb{E}_o\|\mathbf{T} - g(U)\|_2^2 = \mathbb{E}_o\|\mathbf{T} - a(U)\|_2^2 + \mathbb{E}_o\|a(U) - g(U)\|_2^2, \tag{24}$$

because $\mathbb{E}_o[\mathbf{T} - a(U) \mid U] = 0$. Thus a attains $\mathcal{R}_{\text{sel}}$, and the same argument shows that b attains $\mathcal{R}_{\text{full}}$.

Next, write $\mathbf{T} - a = (\mathbf{T} - b) + (b - a)$. Since $b - a$ is measurable with respect to (U, W),

$$\mathbb{E}_o[(\mathbf{T} - b)^\top (b - a)] = \mathbb{E}_o\big[\mathbb{E}_o[\mathbf{T} - b \mid U, W]^\top (b - a)\big] = 0.$$

Therefore,

$$\mathcal{R}_{\text{sel}}(S, o) = \mathcal{R}_{\text{full}}(S, o) + \mathbb{E}_o\|b - a\|_2^2.$$

The last term is nonnegative and vanishes exactly when $b = a$ almost surely, which proves the proposition. □

Interpretation. This result follows from the L^2 projection property of conditional expectation (Williams, 1991). It characterizes the population predictive information gap between the selected information U and the richer observed information (U, W) under a fixed selected set and observation pattern. It does not compare finite fitted networks or imply an ordering of clustering performance. Proposition 1 does not imply that selected-input prediction is preferable when the richer information remains available; rather, it quantifies the information cost of restricting prediction to the selected observed features.

Selected-input predictor excess risk. Let $\Delta(S, o) = \mathcal{R}_{\text{sel}}(S, o) - \mathcal{R}_{\text{full}}(S, o)$. For any square-integrable selected-input predictor $f(U)$, combining Formula 24 with Proposition 1 gives

$$\mathbb{E}_o\|\mathbf{T} - f(U)\|_2^2 - \mathcal{R}_{\text{full}}(S, o) = \Delta(S, o) + \mathbb{E}_o\|a(U) - f(U)\|_2^2. \tag{25}$$

The first term is the unavoidable information loss caused by restricting prediction to U, whereas the second measures the excess error of the selected-input predictor relative to the optimal predictor using U.

For a full-input predictor $h(U, W)$, the discarded features W are excluded from the final selected-feature interface. If they are replaced by a rule $c(U)$, the resulting predictor is

$$f_h(U) = h(U, c(U)). \tag{26}$$

A predictor that performs well when both U and W are available need not remain optimal after W is removed. In particular, $f_h(U)$ need not be close to the optimal predictor based only on U, $a(U) = \mathbb{E}_o[\mathbf{T} \mid U]$. Therefore, low full-input error alone does not guarantee low error when prediction is restricted to the selected features.

E.2 Architectural Property

Property 1 (Architectural inference invariance)**.** *Fix the trained parameters, preprocessing statistics, observation mask and deterministic selected sets $\{\mathcal{S}_v\}_{v=1}^V$. If two versions of sample i agree on every available selected feature, then Formulas 19–20 produce the same $\mathbf{y}_i$, regardless of their discarded features. Consequently, wherever the network is differentiable, $\frac{\partial \mathbf{y}_i}{\partial x_{ij}^{(v)}} = \mathbf{0}, \qquad j \notin \mathcal{S}_v$.*

Proof. At inference, each observed view enters the posterior only through its selected features. Changing a discarded feature therefore leaves the fused posterior unchanged. For an observed view, Formula 19 returns $\mathbf{x}_i^{(v)}$, but Formula 20 retains only the features indexed by $\mathcal{S}_v$. For a missing view, the selected-input posterior and hence the decoder output are unchanged, and again only the features indexed by $\mathcal{S}_v$ are retained. Thus the final representation $\mathbf{y}_i$ is unchanged, which also gives the zero-derivative statement wherever the network is differentiable. □

F Complete Main Results and Robustness

Table 4 reports the average ACC and NMI ranks across all nine missing-view and feature-budget conditions. Figures 3–4 and Figures 5–6 provide the corresponding complete per-dataset robustness curves.

Table 4: Average ACC/NMI ranks across eight datasets under all cross-design conditions (lower is better). Each entry reports ACC rank/NMI rank.

MR	FR	SC-MVFS	TIME-FS	C2IMUFS	UNIFIER	CVFS	CLSG	TERUIMUFS	UIMUFSLR	Mean+CAE	Random-K
0.1	0.3	**2.63/2.50**	6.00/6.13	5.13/5.63	4.69/4.63	6.25/6.00	5.56/5.38	7.13/7.13	3.63/3.25	6.50/6.88	7.50/7.50
0.2	0.3	**2.06/2.00**	4.75/5.63	5.56/4.88	5.50/5.88	5.19/5.25	5.13/5.38	7.25/7.13	4.25/3.63	6.94/7.13	8.38/8.13
0.3	0.1	**1.13/1.63**	4.75/5.13	5.75/6.00	5.63/5.50	6.25/5.88	4.94/3.88	8.00/8.00	3.94/3.88	6.88/7.50	7.75/7.63
0.3	0.2	**1.25/1.63**	5.25/5.50	5.13/5.00	5.69/5.25	5.38/5.00	5.38/5.38	8.38/8.13	3.75/3.88	8.13/7.75	6.69/7.50
0.3	0.3	**1.25/1.50**	5.00/5.50	6.00/4.75	5.25/5.38	5.13/5.00	4.88/5.75	8.63/8.50	3.88/3.88	7.13/7.25	7.88/7.50
0.3	0.4	**1.50/1.50**	4.38/5.63	6.06/5.25	5.25/5.38	5.06/4.75	5.75/5.88	9.25/8.88	3.63/3.63	6.50/6.88	7.63/7.25
0.3	0.5	**1.75/2.50**	4.38/5.25	5.50/4.50	5.25/5.50	5.25/5.50	5.81/5.63	9.38/9.25	4.00/3.50	5.88/6.00	7.81/7.38
0.4	0.3	**1.38/1.25**	5.00/6.00	4.50/4.88	5.75/5.63	4.44/4.88	6.75/6.13	8.25/8.00	4.56/4.00	6.75/7.25	7.63/7.00
0.5	0.3	**2.38/2.13**	5.00/5.50	5.63/6.25	5.25/5.00	5.19/5.88	5.88/6.00	8.25/7.25	3.75/3.25	7.25/6.88	6.44/6.88
Average		**1.70/1.85**	4.94/5.58	5.47/5.24	5.36/5.35	5.35/5.35	5.56/5.49	8.28/8.03	3.93/3.65	6.88/7.06	7.52/7.42

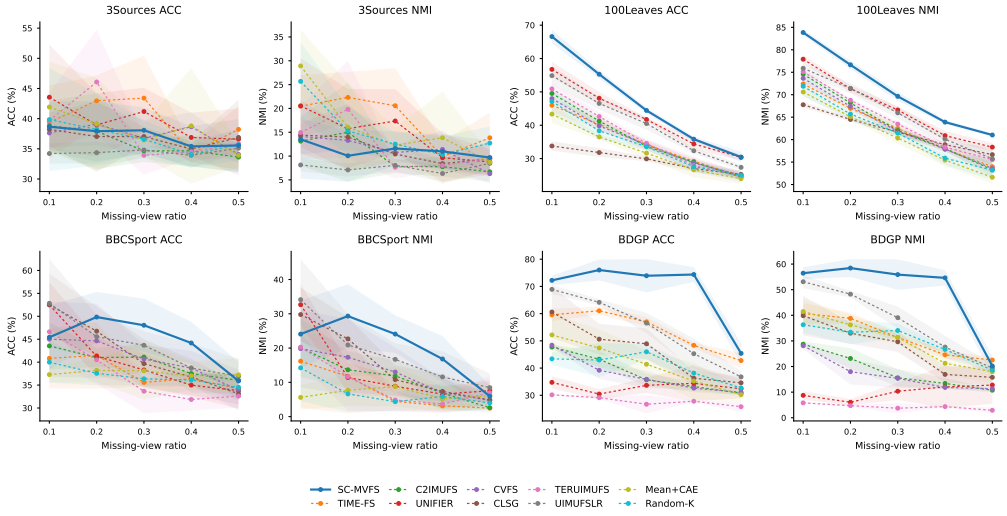


Figure 3: Per-dataset robustness to the missing-view ratio at fixed FR $= 0.3$ (part I: 3Sources, 100Leaves, BBCSport and BDGP). Each panel reports the five-seed mean, with shaded regions indicating one standard deviation.

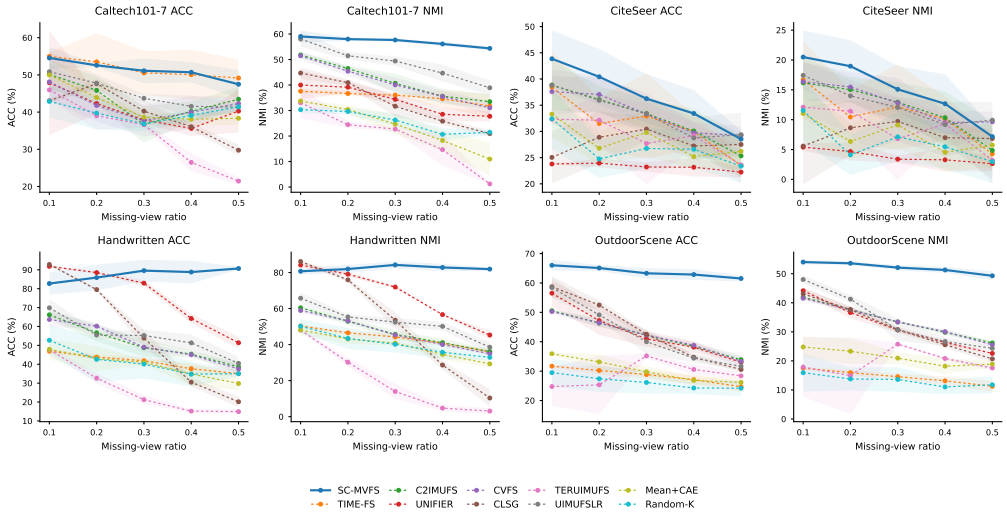


Figure 4: Per-dataset robustness to the missing-view ratio at fixed $\mathrm{FR} = 0.3$ (part II: Caltech101-7, CiteSeer, Handwritten and OutdoorScene). Plotting conventions are identical to Figure 3.

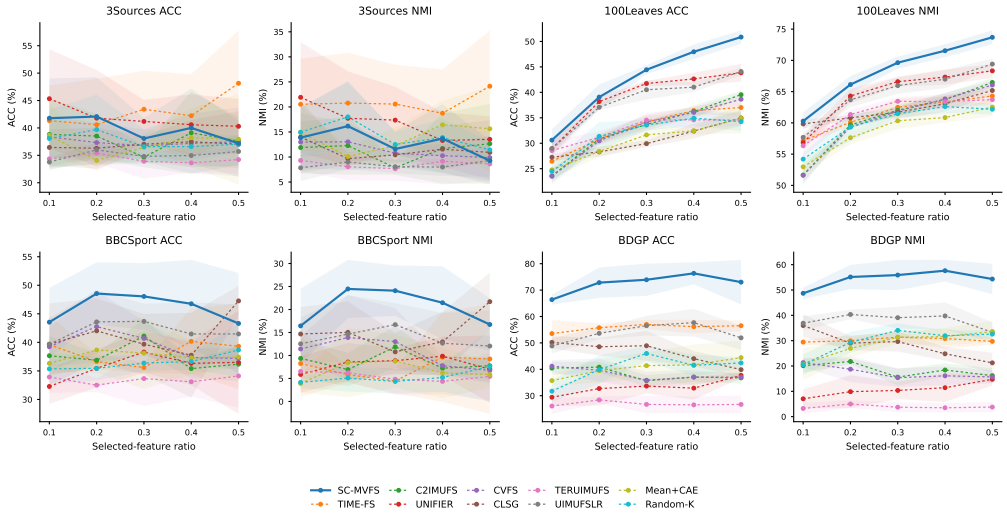


Figure 5: Per-dataset robustness to the selected-feature ratio at fixed $\mathrm{MR} = 0.3$ (part I: 3Sources, 100Leaves, BBCSport and BDGP). Each panel reports the five-seed mean, with shaded regions indicating one standard deviation.

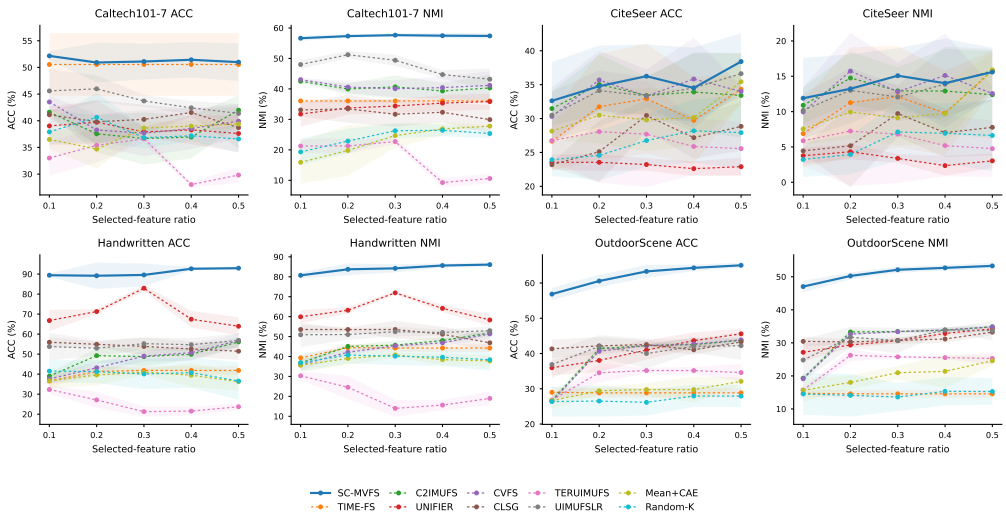


Figure 6: Per-dataset robustness to the selected-feature ratio at fixed $\mathrm{MR} = 0.3$ (part II: Caltech101-7, CiteSeer, Handwritten and OutdoorScene). Plotting conventions and budget protocols are identical to Figure 5.

G Restricted-Input and Completion Diagnostics

G.1 Completion Metrics

Let $\mathcal{H}_v$ denote the samples whose v-th view is hidden for evaluation, and let $\mathcal{V}_{\mathrm{eval}}$ denote the set of evaluated views. For view v, let

$$\mathbf{c}_v = (c_{v1}, \ldots, c_{vd_v})^\top \tag{27}$$

denote the feature-wise evaluation scales. For min–max preprocessed data, c_{vj} is the observed-data range of feature j in view v, with values below 10^{-12} replaced by 1; when no such normalization is used, we set $\mathbf{c}_v = \mathbf{1}$.

The normalized per-view MSE is

$$\mathrm{MSE}_v = \frac{1}{|\mathcal{H}_v| d_v} \sum_{i \in \mathcal{H}_v} \left\| \left(\mathbf{x}_i^{(v)} - \widehat{\mathbf{x}}_i^{(v)} \right) \oslash \mathbf{c}_v \right\|_2^2, \tag{28}$$

where $\oslash$ denotes element-wise division. Its macro-average is

$$\mathrm{MSE}_{\mathrm{macro}} = \frac{1}{|\mathcal{V}_{\mathrm{eval}}|} \sum_{v \in \mathcal{V}_{\mathrm{eval}}} \mathrm{MSE}_v. \tag{29}$$

For RMSE, we first take the square root within each view,

$$\mathrm{RMSE}_v = \sqrt{\mathrm{MSE}_v}, \tag{30}$$

and then macro-average over views:

$$\mathrm{RMSE}_{\mathrm{macro}} = \frac{1}{|\mathcal{V}_{\mathrm{eval}}|} \sum_{v \in \mathcal{V}_{\mathrm{eval}}} \mathrm{RMSE}_v = \frac{1}{|\mathcal{V}_{\mathrm{eval}}|} \sum_{v \in \mathcal{V}_{\mathrm{eval}}} \sqrt{\mathrm{MSE}_v}. \tag{31}$$

G.2 Fixed-Subset Pure-Completion Diagnostic

We first learn a fixed feature subset S from the original 20% development split and reuse it for all completion models. The remaining samples are further divided into completion training and test sets, yielding approximately 20/64/16% development/training/test proportions.

Table 5: Downstream clustering results for the pure-completion diagnostic. All three paths use the same fixed K-feature subset S whenever selected inputs are required, and share the same test samples, masks, and class count.

	Full→Full		Full→Selected		Selected→Selected	
Dataset	ACC	NMI	ACC	NMI	ACC	NMI
3Sources	41.48 ± 4.83	27.63 ± 3.63	41.48 ± 4.83	27.63 ± 3.63	41.48 ± 4.83	27.63 ± 3.63
100Leaves	56.56 ± 0.97	84.83 ± 0.61	49.84 ± 2.74	81.32 ± 0.93	54.30 ± 1.77	83.81 ± 0.55
BBCSport	41.84 ± 1.31	11.88 ± 0.85	42.76 ± 1.89	12.89 ± 2.62	42.53 ± 1.63	12.68 ± 2.76
BDGP	70.45 ± 10.55	56.53 ± 7.28	72.45 ± 2.19	51.92 ± 2.77	74.15 ± 7.08	56.28 ± 4.74
Caltech101-7	57.71 ± 10.99	60.54 ± 4.83	56.61 ± 3.89	59.46 ± 3.54	51.86 ± 5.79	56.21 ± 2.96
CiteSeer	32.98 ± 2.56	12.17 ± 3.01	28.99 ± 4.43	10.22 ± 1.97	30.61 ± 2.34	10.31 ± 2.66
Handwritten	78.56 ± 3.05	77.20 ± 0.61	78.25 ± 4.21	76.34 ± 2.21	78.39 ± 3.16	76.80 ± 2.01
OutdoorScene	59.67 ± 1.01	52.70 ± 1.04	56.56 ± 3.96	50.74 ± 1.93	58.70 ± 2.19	51.36 ± 2.03

In this experiment, only the completion encoders and decoders are trained. Originally observed views are artificially hidden and used as reconstruction targets, whereas naturally missing views are never used as targets. The teacher, mixture model, matching losses, group regularization and learned selector are all removed. The Full and Selected predictors use the same architecture, initialization, training masks and optimization settings. Full→Full uses all observed features during both training and testing. Full→Selected is trained with all features, but at test time only the selected features are retained, with the discarded features replaced by their training means. Selected→Selected uses only the fixed selected features during both training and testing.

Table 5 further reports downstream clustering results obtained from the completed selected-feature representations. Compared with Full→Selected, Selected→Selected improves ACC on five datasets, ties on one, and decreases on two; NMI exhibits the same 5/1/2 pattern. Under this diagnostic, Selected→Selected achieves lower mean completion error on all eight datasets and improves downstream clustering on the majority of datasets.

G.3 Joint Reconstruction-Source Control

We compare two reconstruction-input settings within the full SC-MVFS training objective. Variant C follows the original selected-feature reconstruction setting, in which the reconstruction branch uses only the selected features from the student's currently available views. Variant F modifies only this branch to use all currently observed features from the same views. Unlike the fixed-subset diagnostic in Appendix G.2, each joint-training variant learns its own selected subset. The student view set $\mathcal{V}_i^{\mathrm{red}} = \{v : \widetilde{o}_{iv} = 1\}$ is identical for the two variants. All other training components are kept unchanged, including the teacher, the posterior- and assignment-matching objectives, mixture modeling, and regularization terms. Each variant learns its own exact-K feature subset, and both use only the selected features for final inference and completion. For this paired control, both variants use a fixed patience of 9 for matched training rounds and separate random-number streams for view hiding and feature gating.

Table 6 reports the complete per-dataset RMSE, ACC, and NMI results. In addition to the aggregate results reported in Section 5.4, C achieves higher NMI than F on five datasets. Its ACC is higher on five datasets and tied on one. These results suggest that selected-feature reconstruction tends to reduce completion error and can improve downstream clustering on the majority of datasets. Completion error and cluster separability are not equivalent objectives, so lower RMSE need not coincide with higher ACC or NMI.

Table 6: Joint reconstruction-source control at $\mathrm{MR} = \mathrm{FR} = 0.3$. C uses selected-source reconstruction, whereas F uses all currently observed source features under the same student view pattern. Lower RMSE and higher ACC/NMI are better.

	RMSE		ACC (%)		NMI (%)	
Dataset	C	F	C	F	C	F
3Sources	1.4158 ± 0.0266	1.4170 ± 0.0264	35.41 ± 0.97	35.41 ± 2.48	9.21 ± 1.18	9.97 ± 1.15
100Leaves	0.9801 ± 0.0228	1.0175 ± 0.0307	45.72 ± 1.13	44.34 ± 0.93	70.13 ± 0.53	69.77 ± 0.29
BBCSport	1.6346 ± 0.0495	1.6415 ± 0.0495	48.37 ± 3.62	48.02 ± 2.73	23.20 ± 2.21	22.58 ± 2.60
BDGP	1.0138 ± 0.0142	1.0275 ± 0.0143	74.85 ± 6.23	74.26 ± 6.45	56.74 ± 6.22	56.03 ± 6.35
Caltech101-7	0.7410 ± 0.0101	0.7418 ± 0.0102	52.43 ± 0.34	53.81 ± 2.42	57.57 ± 0.44	57.89 ± 1.17
CiteSeer	0.9440 ± 0.0131	0.9455 ± 0.0130	36.47 ± 5.97	38.93 ± 7.35	14.64 ± 4.65	16.51 ± 5.97
Handwritten	0.6666 ± 0.0033	0.6664 ± 0.0033	89.85 ± 6.11	89.79 ± 5.79	84.75 ± 1.99	84.44 ± 1.62
OutdoorScene	0.7980 ± 0.0674	0.8038 ± 0.0667	63.44 ± 1.89	63.40 ± 0.93	52.10 ± 0.93	51.97 ± 0.46

H Component Ablations

To examine the roles of the main training mechanisms in SC-MVFS, we ablate three components corresponding to the core designs introduced in the method section. First, posterior matching transfers information from the full-input teacher to the selected-input student. Second, masked reconstruction explicitly trains the completion pathway to reconstruct from selected features only. Third, reduced-view augmentation additionally hides part of the observed views during training, exposing the student to more incomplete-view patterns. We therefore remove each of these three mechanisms individually while keeping all remaining settings fixed to evaluate their respective contributions.

No posterior matching sets $\lambda_{\mathrm{post}} = 0$, thereby removing $\mathcal{L}_{\mathrm{post}}$. *No masked reconstruction* sets $\lambda_{\mathrm{mask}} = 0$, thereby removing $\mathcal{L}_{\text{mask-rec}}$. *No reduced-view augmentation* sets the additional view-hiding probability from 0.35 to 0, so that the student uses all naturally observed views during training. All variants retain the same exact-K feature selection rule and use only the selected features for final inference and completion.

Table 7 reports the complete per-dataset ACC and NMI results together with the corresponding average ranks across the four variants. The full model achieves the best aggregate average rank for both ACC and NMI. Removing any of the three mechanisms worsens the corresponding aggregate ranking, although some ablated variants may perform better on individual datasets or metrics. These results support the roles of teacher–student posterior transfer, selected-feature reconstruction, and reduced-view augmentation in the overall training framework.

Table 7: Component ablations at $\mathrm{MR} = \mathrm{FR} = 0.3$ (%, mean $\pm$ sample standard deviation over five seeds). The bottom panel reports average ranks across the eight datasets among the four listed variants; lower rank is better.

Dataset	Variant	ACC	NMI	Dataset	Variant	ACC	NMI
3Sources	SC-MVFS	38.07 ± 3.80	11.60 ± 3.60	Caltech101-7	SC-MVFS	51.11 ± 3.36	57.66 ± 0.56
	No posterior matching	34.81 ± 1.05	8.98 ± 2.61		No posterior matching	51.21 ± 2.90	57.42 ± 1.08
	No masked reconstruction	38.52 ± 5.42	11.33 ± 4.25		No masked reconstruction	49.69 ± 4.05	56.80 ± 0.38
	No reduced-view augmentation	37.63 ± 3.90	11.76 ± 5.23		No reduced-view augmentation	53.47 ± 0.50	56.75 ± 0.29
100Leaves	SC-MVFS	44.44 ± 0.88	69.63 ± 0.98	CiteSeer	SC-MVFS	36.25 ± 4.67	15.08 ± 3.89
	No posterior matching	44.84 ± 2.14	69.67 ± 1.12		No posterior matching	31.87 ± 6.95	11.81 ± 4.95
	No masked reconstruction	44.39 ± 2.13	69.49 ± 1.09		No masked reconstruction	33.42 ± 6.75	13.01 ± 6.02
	No reduced-view augmentation	41.66 ± 0.89	68.45 ± 0.28		No reduced-view augmentation	32.65 ± 5.61	11.44 ± 4.98
BBCSport	SC-MVFS	48.05 ± 5.83	24.09 ± 5.49	Handwritten	SC-MVFS	89.56 ± 5.72	84.22 ± 2.03
	No posterior matching	44.23 ± 1.46	18.89 ± 2.48		No posterior matching	91.81 ± 0.41	84.47 ± 0.69
	No masked reconstruction	43.54 ± 8.89	18.25 ± 7.22		No masked reconstruction	89.71 ± 5.69	84.45 ± 2.22
	No reduced-view augmentation	44.28 ± 6.46	20.37 ± 9.10		No reduced-view augmentation	77.65 ± 3.12	77.43 ± 1.06
BDGP	SC-MVFS	73.92 ± 6.12	55.90 ± 5.97	OutdoorScene	SC-MVFS	63.29 ± 1.82	52.12 ± 0.91
	No posterior matching	72.50 ± 1.45	53.57 ± 1.88		No posterior matching	62.85 ± 0.72	50.69 ± 0.80
	No masked reconstruction	74.57 ± 4.78	56.05 ± 4.47		No masked reconstruction	63.54 ± 1.82	52.02 ± 1.08
	No reduced-view augmentation	67.00 ± 9.18	49.60 ± 5.62		No reduced-view augmentation	65.59 ± 1.41	52.49 ± 0.95

Variant	ACC rank	NMI rank
SC-MVFS	**2.13**	**1.75**
No posterior matching	2.75	2.63
No masked reconstruction	2.38	2.63
No reduced-view augmentation	2.75	3.00

I SENSITIVITY, OPTIMIZATION AND EFFICIENCY

Hyperparameter sensitivity. We examine three key hyperparameters on 3Sources, BDGP and Caltech101-7 under the setting $\mathrm{MR} = \mathrm{FR} = 0.3$. The tested view-dropout rates are $\{0.1, 0.2, 0.35, 0.5, 0.7\}$, with 0.35 as the default. The latent dimension h is varied over $\{16, 32, 64, 128\}$, with $h = 32$ as the default. The posterior-matching weight λ_{post} is varied over $\{0.1, 0.25, 0.5, 1, 2\}$, with $\lambda_{\mathrm{post}} = 0.5$ as the default. Figure 7 reports the corresponding NMI results. No single tested value consistently performs best across all three datasets, while the default settings remain competitive throughout the tested ranges.

Alternating-round sensitivity. Table 8 compares maximum round counts of 4, 8 and 12 across all eight datasets. Eight rounds achieve the best average ACC and NMI ranks, both 1.25. These results support the adopted eight-round setting among the tested choices.

Optimization and efficiency. Table 9 reports the per-dataset computational and optimization statistics at $\mathrm{MR} = \mathrm{FR} = 0.3$, including fit time, peak memory, and the number of alternating rounds over five seeds. These measurements characterize the computational footprint of SC-MVFS across datasets.

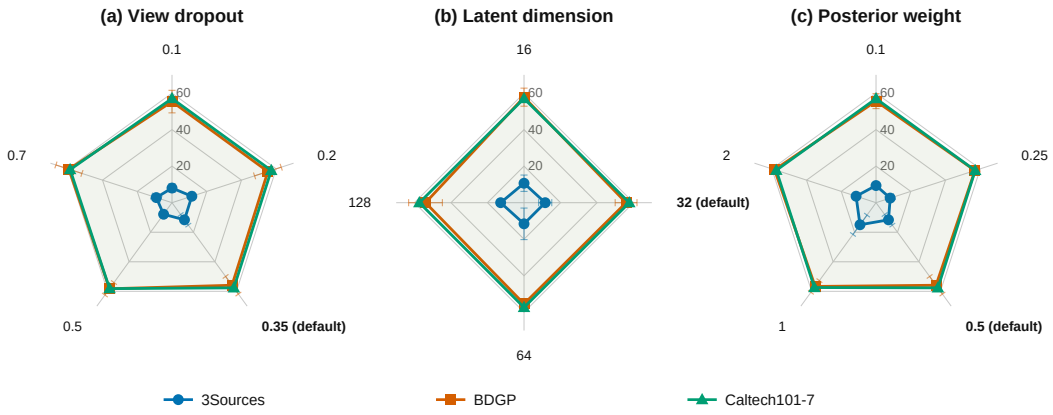


Figure 7: Key-parameter sensitivity at $\mathrm{MR} = \mathrm{FR} = 0.3$. Each radar axis corresponds to one tested value and the radial axis reports NMI (%). Solid curves show the five-seed means on 3Sources, BDGP and Caltech101-7, with radial whiskers indicating one standard deviation. Bold axis labels mark the frozen defaults.

Table 8: Alternating-round sensitivity at $\mathrm{MR} = \mathrm{FR} = 0.3$. The last row gives the average rank across datasets; lower is better.

	4 rounds		8 rounds		12 rounds	
Dataset	ACC	NMI	ACC	NMI	ACC	NMI
3Sources	36.15±2.74	11.59±3.27	38.07±3.80	11.60±3.60	35.68±1.99	8.24±1.85
100Leaves	43.41±2.47	69.10±1.16	44.44±0.88	69.63±0.98	45.69±1.85	70.45±1.22
BBCSport	41.24±5.09	16.26±6.51	48.05±5.83	24.09±5.49	44.32±4.94	19.77±2.78
BDGP	72.84±3.14	55.51±3.07	73.92±6.12	55.90±5.97	73.31±5.80	55.72±5.61
Caltech101-7	50.83±4.51	57.58±1.41	51.11±3.36	57.66±0.56	50.52±0.76	57.45±0.72
CiteSeer	36.88±4.82	15.31±4.68	36.25±4.67	15.08±3.89	35.71±5.84	14.84±3.85
Handwritten	87.64±5.66	82.35±1.95	89.56±5.72	84.22±2.03	88.94±0.61	83.86±0.89
OutdoorScene	63.02±0.98	51.96±0.37	63.29±1.82	52.12±0.91	62.74±1.75	51.84±1.11
Average rank ↓	2.38	2.38	**1.25**	**1.25**	2.38	2.38

Table 9: Per-dataset optimization and efficiency at $\mathrm{MR} = \mathrm{FR} = 0.3$. Values are mean $\pm$ sample standard deviation over five seeds.

Dataset	Fit time (s)	Peak memory (MiB)	Rounds
3Sources	1.39±0.20	117.37±0.51	7.40±1.34
100Leaves	12.64±1.55	85.52±0.00	7.00±2.24
BBCSport	2.96±0.11	91.06±0.54	8.00±0.00
BDGP	11.34±0.25	52.14±0.09	8.00±0.00
Caltech101-7	11.25±0.52	74.67±0.08	8.00±0.00
CiteSeer	27.61±10.85	161.44±1.09	8.00±0.00
Handwritten	16.52±0.87	36.16±0.00	7.80±0.45
OutdoorScene	14.76±0.24	46.66±0.00	8.00±0.00

J Controlled Synthetic Experiments

J.1 Experimental Setup

We construct a controlled synthetic dataset with 600 samples, three views, and five balanced classes. Each sample has a shared six-dimensional class-dependent latent representation

$$\mathbf{z}_i \boldsymbol{\kappa}_{c_i} + 0.55 s_i \boldsymbol{\xi}_i, \qquad \boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_6), \tag{32}$$

where the class centers are sampled from a Gaussian distribution, centered across classes, and rescaled to Euclidean norm 2.5. Sample-specific variation is introduced by

$$s_i = \exp(0.3 t_i - 0.045), \qquad t_i \sim \mathcal{N}(0, 1). \tag{33}$$

Each view contains eight primary features. For feature j in view v, we sample $\mathbf{g}_j^{(v)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_6)$ and define

$$\widetilde{\mathbf{a}}_j^{(v)} 0.20, \mathbf{g}_j^{(v)} + \mathbf{e}_{(j+2v) \bmod 6}, \qquad \mathbf{a}_j^{(v)} \frac{\widetilde{\mathbf{a}}_j^{(v)}}{\max(|\widetilde{\mathbf{a}} j^{(v)}|2, 10^{-12})}, \tag{34}$$

where j and v are zero-indexed in the generator. The clean primary signal is $\widetilde{p} i j^{(v)} = \mathbf{a}_j^{(v)\top} \mathbf{z}_i$. Let

$$s j, \mathrm{clean}^{(v)} \operatorname{Std} i\big(\widetilde{p} i j^{(v)}\big), \tag{35}$$

where sample standard deviation uses $\mathrm{ddof} = 1$. With the fixed base $\mathrm{SNR} = 2.0$, no view-noise asymmetry, and observation-noise multiplier equal to one, the primary features are

$$p_{ij}^{(v)} \widetilde{p}_{ij}^{(v)} + s_i \frac{s_{j,\mathrm{clean}}^{(v)}}{\sqrt{2}} \epsilon_{ij}^{(v)}, \qquad \epsilon_{ij}^{(v)} \sim \mathcal{N}(0, 1). \tag{36}$$

Thus, $\mathrm{SNR} = 2.0$ determines the base feature-wise noise scale before the sample-dependent factor s_i is applied; it does not imply that the realized finite-sample SNR is exactly two. Each primary feature has one noisy redundant copy with perturbation coefficient 0.15. For view v, define the common feature scale

$$\beta_v \frac{1}{8} \sum_{j=1}^{8} \max\left(s_{j,\mathrm{clean}}^{(v)}, 10^{-6}\right). \tag{37}$$

Each view additionally contains 32 independent Gaussian nuisance features with standard deviation β_v. To introduce class-independent cross-view information, we sample

$$\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{48}). \tag{38}$$

For each of the 96 shortcut features in view v, we independently sample $\widetilde{\mathbf{w}} j^{(v)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}48)$ and normalize

$$\mathbf{w}_j^{(v)} \frac{\widetilde{\mathbf{w}}_j^{(v)}}{\max(|\widetilde{\mathbf{w}}_j^{(v)}|2, 10^{-12})}. \tag{39}$$

Writing

$$g i j^{(v)} \frac{\mathbf{w}_j^{(v)\top} \mathbf{u}_i}{\max!\left(\operatorname{Std} i(\mathbf{w}_j^{(v)\top} \mathbf{u}_i), 10^{-12}\right)}, \tag{40}$$

with $\mathrm{ddof} = 1$, the shortcut feature is

$$qij^{(v)}\beta_v \left[\eta g_{ij}^{(v)} + \sqrt{1-\eta^2}, e_{ij}^{(v)}\right], \qquad e_{ij}^{(v)} \sim \mathcal{N}(0,1), \tag{41}$$

where $\eta \in 0, 0.8, 0.9$. Increasing η strengthens the shared cross-view component while keeping the shortcut scale approximately stable. Each view therefore contains $8 + 8 + 96 + 32 = 144$ features, giving 432 features in total. Feature columns are independently permuted within each view. The global selection budget is $K = 24$. We use a 20/80 development/evaluation split and approximately 30% view-level missingness while ensuring that each sample retains at least one observed view. No additional feature normalization is applied.

Table 10: Configuration of the controlled synthetic experiments.

Setting	Value	Setting	Value
Samples	600	Views	3
Classes	5	Class latent dimension	6
Class-center radius	2.5	Sample-scale coefficient	0.55
Primary features/view	8	Redundant features/view	8
Shortcut features/view	96	Nuisance features/view	32
Shortcut latent dimension	48	Base SNR	2.0
Redundant perturbation	0.15	Loading perturbation	0.20
Features/view	144	Total features	432
Selected features K	24	Shortcut strength η	$\{0,\ 0.8,\ 0.9\}$
View-level missing ratio	0.3	Dev./eval. split	20/80
Synthetic repetitions	20	K-means restarts	20

J.2 Distributed Shortcuts and Selection–Completion Consistency

We compare SC-MVFS with a matched full-input completion variant under $\eta \in \{0, 0.8, 0.9\}$. Both variants use the same global feature budget $K = 24$, model architecture, optimization settings, and generated datasets. Their only difference is the information provided to the completion pathway during training: SC-MVFS uses only the currently selected features, whereas the full-input variant uses all features from the same available views. At inference, both variants are restricted to their final selected features. The full-input variant therefore isolates the effect of training completion with information that is unavailable after feature selection.

Results are reported over 20 independently generated synthetic datasets. For every repetition, the two variants are trained and evaluated on the same generated dataset, producing 20 paired comparisons at each shortcut strength. Table 11 reports NMI as mean $\pm$ sample standard deviation across these 20 repetitions.

For the cross-view prediction diagnostic, we use ridge regression ($\alpha = 1$) over all six ordered source–target view pairs with a fixed 60/40 train/test split. The prediction target consists of the 96 shortcut features in the target view. The all-feature setting uses all 96 shortcut features from the source view, whereas the budget-subset setting uses eight randomly sampled shortcut features. The latter result is averaged over eight independent subset draws.

As shown in Table 11, SC-MVFS consistently outperforms the matched full-input variant. At $\eta = 0$, their NMI values are 67.73 ± 4.58 and 64.17 ± 3.60, respectively. When the shortcut strength increases to 0.8, the corresponding results become 56.54 ± 9.13 and 48.41 ± 8.24. At $\eta = 0.9$, SC-MVFS obtains 43.51 ± 11.62, compared with 32.00 ± 12.66 for full-input completion. The resulting paired mean advantages of SC-MVFS are 3.56, 8.14, and 11.51 NMI percentage points, respectively. Thus, although the task becomes more difficult for both variants as the shortcut strength increases, the disadvantage of training completion with subsequently discarded information grows more rapidly.

The prediction diagnostics confirm that the shortcut information becomes increasingly useful only when it is aggregated across many features. The all-feature cross-view prediction R^2 increases from -0.371 at $\eta = 0$ to 0.244 at $\eta = 0.8$ and 0.556 at $\eta = 0.9$. In contrast, the budget-subset R^2 remains much lower, changing only from -0.030 to 0.038 and 0.080. Hence, no small budget-sized shortcut subset captures the predictive information available to the full-input completion pathway.

The selected subsets exhibit the corresponding behavior. At $\eta = 0.8$, the full-input variant allocates 47.5% of its selected budget to shortcut features, compared with 28.8% for SC-MVFS. At $\eta = 0.9$, these proportions increase to 66.9% and 53.8%, respectively. Meanwhile, SC-MVFS retains a larger proportion of primary features at every shortcut strength. Together, these results support the proposed mechanism: when cross-view predictive information is distributed across many features that cannot all be retained under the final budget, training completion with the full input creates a larger mismatch with the selected-feature inference interface. Restricting completion to selected evidence reduces this mismatch.

Table 11: Controlled synthetic results under increasing shortcut strength. NMI is reported as mean $\pm$ sample standard deviation over 20 independently generated synthetic datasets. Individual correlation measures the average absolute correlation of a single shortcut feature with the cross-view prediction target. All-feature and budget-subset R^2 measure cross-view prediction using all shortcut features and a budget-sized subset, respectively. The remaining columns report averages over the 20 repetitions. The last four columns give the percentages of shortcut and primary features in the final selected subset.

Shortcut	NMI (%)		Individual	All features	Budget-subset	Shortcut selected (%)		Primary selected (%)	
strength η	SC-MVFS	Full-input completion	correlation	R^2	R^2	SC-MVFS	Full-input completion	SC-MVFS	Full-input completion
0.0	$\mathbf{67.73 \pm 4.58}$	64.17 ± 3.60	0.033	-0.371	-0.030	0.4	5.2	50.0	47.3
0.8	$\mathbf{56.54 \pm 9.13}$	48.41 ± 8.24	0.081	0.244	0.038	28.8	47.5	35.8	27.1
0.9	$\mathbf{43.51 \pm 11.62}$	32.00 ± 12.66	0.099	0.556	0.080	53.8	66.9	22.5	17.1

K VIEW-PATTERN CONTROLS

We compare three variants to separate two effects: the effect of randomly hiding additional observed views during student training and the effect of allowing the teacher to use more observed views than the student. Let $\mathcal{V}_i^{\mathrm{obs}} = \{v : o_{iv} = 1\}$ and $\mathcal{V}_i^{\mathrm{red}} = \{v : \widetilde{o}_{iv} = 1\}$. Variant A uses the naturally observed view set $\mathcal{V}_i^{\mathrm{obs}}$ for both the teacher and the student, with no additional view dropout. Variant B retains the additional view dropout, but both the teacher and the student use the same reduced view set $\mathcal{V}_i^{\mathrm{red}}$. Variant C follows the original setting: the student uses $\mathcal{V}_i^{\mathrm{red}}$, while the teacher uses all naturally observed views in $\mathcal{V}_i^{\mathrm{obs}}$. All other training settings are kept unchanged, and each variant learns its own exact-K feature subset. For this control, we use the same paired experimental protocol as in Appendix G.3, including a fixed patience of 9 and separate random-number streams for view hiding and feature gating.

Comparing B with A isolates the effect of reduced-view augmentation when the teacher and student use the same views. As shown in Table 12, B achieves lower completion RMSE than A on all eight datasets, while NMI also improves on six datasets. This suggests that exposing the model to more incomplete view patterns during training improves completion under this setting.

Comparing C with B isolates the effect of allowing the teacher to use the additional naturally observed views while keeping the student input unchanged. C achieves higher NMI on seven of the eight datasets, while completion RMSE is lower on only three datasets. This suggests that the extra views available to the teacher provide more consistent benefits for clustering than for completion.

Table 12: View-pattern controls at $\mathrm{MR} = \mathrm{FR} = 0.3$. Variant A removes additional view dropout, Variant B uses the same reduced view pattern for teacher and student and Variant C is the original setting in which the teacher uses all naturally observed views while the student uses the reduced pattern. Lower RMSE and higher NMI are better.

Dataset	A	B	C
RMSE			
3Sources	1.4200 ± 0.0263	1.4169 ± 0.0268	1.4158 ± 0.0266
100Leaves	1.0210 ± 0.0138	0.9760 ± 0.0146	0.9801 ± 0.0228
BBCSport	1.6388 ± 0.0495	1.6369 ± 0.0498	1.6346 ± 0.0495
BDGP	1.0850 ± 0.0247	1.0006 ± 0.0109	1.0138 ± 0.0142
Caltech101-7	0.7414 ± 0.0093	0.7225 ± 0.0105	0.7410 ± 0.0101
CiteSeer	0.9468 ± 0.0131	0.9445 ± 0.0133	0.9440 ± 0.0131
Handwritten	0.6922 ± 0.0047	0.6643 ± 0.0067	0.6666 ± 0.0033
OutdoorScene	0.8271 ± 0.0634	0.7917 ± 0.0657	0.7980 ± 0.0674
NMI (%)			
3Sources	11.27 ± 4.79	8.61 ± 1.52	9.21 ± 1.18
100Leaves	68.63 ± 0.58	69.45 ± 0.95	70.13 ± 0.53
BBCSport	20.87 ± 8.76	22.89 ± 4.81	23.20 ± 2.21
BDGP	52.07 ± 3.54	56.37 ± 5.48	56.74 ± 6.22
Caltech101-7	56.04 ± 0.48	56.81 ± 0.76	57.57 ± 0.44
CiteSeer	13.12 ± 5.77	12.22 ± 4.52	14.64 ± 4.65
Handwritten	78.94 ± 1.90	80.85 ± 1.84	84.75 ± 1.99
OutdoorScene	53.49 ± 1.42	53.62 ± 0.99	52.10 ± 0.93