RETHINKING MULTI-VIEW CLUSTERING THROUGH SEMANTIC RECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering (MVC) seeks to discover meaningful groups by integrating complementary descriptions of the same samples. However, geometric consistency alone does not ensure semantic correctness: nuisance variation can fragment one class across clusters, while shared contextual cues can mix distinct classes. Correcting these errors calls for semantic evidence that reveals both shared identities across clusters and distinct identities within them. We propose Semantic Recomposition (SemReC), which starts from a geometric partition, extracts semantic descriptions from sparse, coverage-oriented representatives using a multimodal large language model, and organizes them into frozen semantic seeds. These seeds guide split proposals, while a joint geometric–semantic objective evaluates coupled merge–split moves that preserve the cluster count. Across four benchmarks spanning visual, image–text, and audio–text settings, SemReC achieves an average ACC of 84.97% and the highest ACC on all four datasets among the compared baselines.

1 INTRODUCTION

Multi-view clustering (MVC) groups unlabeled samples by integrating complementary information from heterogeneous views. Geometry-driven approaches include traditional methods that recover shared structure through matrix factorization, latent embedding, and affinity learning (Wen et al., 2018; Chen et al., 2021; Guo et al., 2023), as well as deep methods that learn nonlinear representations and adaptive fusion mechanisms (Wu et al., 2024; Du et al., 2025). These approaches primarily optimize reconstruction, structural coherence, and cross-view agreement to obtain consistent partitions. However, geometric coherence does not guarantee semantic correctness.

Two opposing errors frequently arise. **Semantic fragmentation** occurs when one semantic class is scattered across multiple clusters because of nuisance variation such as viewpoint, pose, wording, or recording condition. **Semantic mixing** occurs when distinct semantic classes are grouped together because they share contextual or structural cues. Figure 1 illustrates both on COCO-2017 ($K = 44$): Bear is fragmented across three nearly pure clusters due to appearance variation, while person and dog are mixed in one cluster because of shared natural-scene context. These errors corrupt the semantic meaning of clusters themselves. Mixing makes a cluster-level prototype, pseudo-label, or retrieved neighborhood internally contradictory, whereas fragmentation represents one concept as several artificial identities. Consequently, downstream procedures that rely on cluster structure inherit incorrect grouping decisions rather than merely noisy individual assignments.

Semantic guidance could in principle correct such errors, but existing semantic-aware MVC methods still obtain their supervisory signals from the same learned feature space used to form the partition, for example through cluster-assignment contrast, attention-weighted fusion, or pseudo-label guidance (Chen et al., 2023a; Xu et al., 2022; Sun et al., 2025). Such signals can improve representation consistency, yet they are not independent evidence of semantic identity and may inherit the same geometric biases that produced the errors. Raw-content semantic models, particularly multimodal large language models (MLLMs), offer a different source of evidence because they can directly interpret the observed content rather than infer semantics only from embedding geometry. Existing LLM-assisted MVC methods, however, mainly use language models upstream to adjust fusion or representation learning. For example, LLM-DAMVC (Xu & Wang, 2025) uses an LLM to deter-

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.

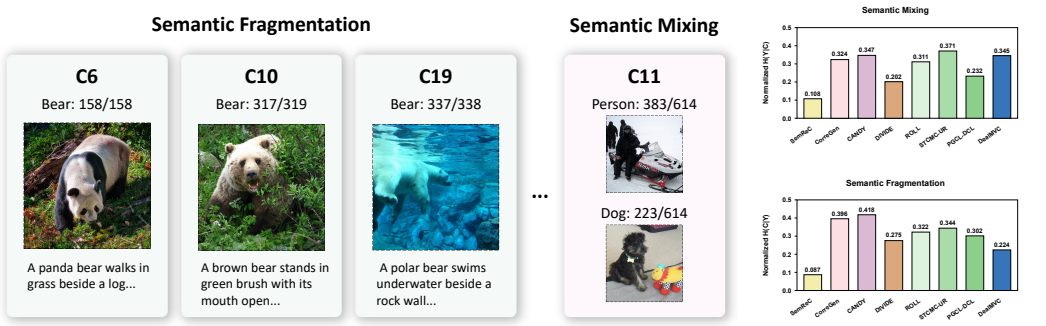

Figure 1: Semantic fragmentation and mixing: examples on COCO (left) and mean label-based diagnostics across four datasets using seed 17 (right).

mine view-fusion policies, but does not use semantic evidence to diagnose and structurally correct an already formed partition.

The two errors correspond naturally to the complementary conditional entropies underlying V-measure (Rosenberg & Hirschberg, 2007). $H(Y|C)$ measures **within-cluster mixing**, asking whether each cluster corresponds to a single semantic class, while $H(C|Y)$ measures **cross-cluster fragmentation**, asking whether each semantic class is concentrated in one cluster. Importantly, these objectives are not independent under a fixed cluster budget K. Merging two clusters can reduce fragmentation but may increase mixing; splitting a mixed cluster can reduce mixing but may create additional fragmentation. Semantic correction is therefore inherently **bidirectional and coupled**. Our post-hoc analysis of seven MVC baselines across four image, text, and audio benchmarks confirms this tension: all exhibit substantial mixing and fragmentation, and no method jointly minimizes both errors (Figure 1).

Addressing this coupled objective raises two distinct challenges. **First, the true semantic variable Y is unavailable during unsupervised clustering.** We therefore construct a semantic surrogate S: an MLLM describes the raw multimodal content of representative samples, an anonymous audit normalizes these descriptions into comparable semantic phrases, and coverage-weighted grouping converts them into frozen semantic seeds. This yields computable counterparts $H(S|C)$ and $H(C|S)$ that preserve the same two structural directions as the unavailable $H(Y|C)$ and $H(C|Y)$. **Second, semantic evidence should remain sparse rather than replace clustering with dense semantic labeling.** We therefore query only a coverage-oriented set of representatives and weight their semantic evidence by the samples they represent. Semantic information is thus used to identify suspicious partition boundaries, while the underlying geometric structure remains responsible for constraining how the partition may change.

Based on this principle, we present Semantic Recomposition (SemReC), a semantic-geometric framework that assigns semantic evidence a constrained structural role: deciding when geometrically formed clusters should be recomposed rather than replacing geometry itself. SEMREC has three stages. **Stage I, Geometric Bootstrap**, forms the initial multi-view partition and its representative-based local units. **Stage II, Semantic Evidence Formation**, generates modality-grounded cards, audits them into normalized phrases, embeds the phrases, and groups them into K frozen semantic seeds. **Stage III, Geometry-Constrained Recomposition**, evaluates atomic merge–split candidates under a joint objective combining geometric crossing cost with $H(S|C)$ and $H(C|S)$. Each candidate merges a pair of current clusters and splits a semantic subgroup from a third; acceptance preserves exactly K nonempty clusters and strictly decreases the full objective.

Our contributions are threefold. **(i)** We identify semantic fragmentation and semantic mixing as coupled structural failures of geometric MVC, and formulate a bidirectional semantic objective that captures both directions simultaneously. **(ii)** We introduce sparse semantic surrogates and atomic merge-split optimization, allowing raw-content semantic evidence to correct partition structure without replacing geometric clustering. **(iii)** Across four heterogeneous benchmarks, our method achieves the best ACC among eight baselines; permutation placebo and post-hoc structural diagnostics further

show that the improvement depends on semantic identity and coincides with simultaneous reductions in both mixing and fragmentation.

2 Background and Related Work

Geometry-driven multi-view clustering. Geometry-driven MVC includes traditional optimization methods and deep learning approaches. Traditional methods recover shared structure through matrix factorization and latent embedding (Wen et al., 2018; Chen et al., 2021), or model cross-view dependencies through low-rank subspace and tensor learning (Guo et al., 2022; Chen et al., 2023b; Guo et al., 2023; Qin et al., 2024). Anchor-based formulations improve scalability while preserving local and higher-order relationships (Xia et al., 2023; Guo et al., 2025; Liang et al., 2026). Other developments address unpaired or incomplete observations through graph matching, graph filtering, and structured embedding learning (Li et al., 2025; Du et al., 2024; Gu et al., 2026), or improve clustering optimization through optimal transport (Lin et al., 2023). Deep approaches introduce trainable representations and fusion mechanisms. For example, self-weighted contrastive fusion uses feature-distribution discrepancies to adjust view contributions (Wu et al., 2024), while LargeMvC-Net unfolds anchor-based optimization into a reconstruction-driven network (Du et al., 2025). These methods primarily optimize reconstruction, affinity structure, and representation agreement, without explicitly evaluating the semantic identities of the resulting clusters.

Semantic-aware MVC and language-model guidance. Semantic-aware MVC introduces inferred category information or mined relations that serve as proxies for semantic membership. MFLVC combines reconstruction, common features, and learned semantic labels (Xu et al., 2022), while joint contrastive triple-learning incorporates cluster-level consistency (Hu et al., 2023). Assignment-based methods improve semantic guidance through different mechanisms: DealMVC calibrates feature similarities against high-confidence pseudo-label relations (Yang et al., 2023); STCMC-UR selectively uses pseudo-labels according to uncertainty (Hu et al., 2025a); PGCL-DCL combines pseudo-label-guided contrastive learning with dual correlation learning (Hu et al., 2025b); and ROLL learns robust pseudo-labels under noisy correspondence (Sun et al., 2025). Relation-based approaches include contextual semantic mining in CANDY (Guo et al., 2024), high-order random-walk correction of contrastive targets in DIVIDE (Lu et al., 2024), and cluster-informed correspondence recovery in CorreGen (Zhou et al., 2026). DREAM further combines semantic and structural representations through dynamic fusion and multi-granularity alignment (Kong et al., 2026). These approaches derive internal semantic proxies from learned representations and feed them back into training. Language models introduce additional guidance: LLM-DAMVC uses clustering-quality statistics to obtain a view-fusion policy (Xu & Wang, 2025), whereas Coordinator generates semantic representations with an LVLM for dynamic fusion and encoder refinement (Lan et al., 2026). Such guidance primarily shapes representations or fusion. Our method instead freezes semantic evidence extracted from sparse representative bundles and uses it to guide split proposals and evaluate coupled merge–split moves alongside geometric constraints. This directly addresses within-cluster semantic mixing and cross-cluster fragmentation on an existing partition while preserving the number of clusters.

3 Method

Given N samples observed from V views and a prescribed number of clusters K, we aim to correct semantic errors in an initially geometric multi-view partition without access to class labels. Figure 2 summarizes the complete pipeline: geometric bootstrap, sparse semantic evidence formation, and geometry-constrained merge–split recomposition at a fixed cluster count. Figure 2 previews Stage III: semantic seeds propose split boundaries, while the quotient graph scores their geometric cost. We first characterize two complementary semantic failures (Sec. 3.1). We then construct sparse semantic evidence from geometry-selected representatives (Sec. 3.2), before combining this evidence with the original geometry to perform fixed-K semantic recomposition (Sec. 3.3).

3.1 Problem Formulation

Let $\mathcal{X} = \{(x_i^{(1)}, \ldots, x_i^{(V)})\}_{i=1}^N$ denote N samples observed through V views. Given the known cluster count K, multi-view clustering seeks a partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ into exactly K nonempty clusters. Let the random variable C denote cluster identity under the current partition.

Class labels are unavailable during feature processing, clustering, representative selection, semantic prompting, and optimization; ACC, NMI, and ARI are computed only after predictions have been saved.

To characterize semantic correctness independently of a particular clustering algorithm, we introduce an oracle semantic variable Y, used only for problem definition and post-hoc evaluation. A partition can exhibit two complementary errors. Semantic mixing occurs when one cluster contains multiple semantic identities, measured by

$$D_{\mathrm{mix}}(\mathcal{C}; Y) = \frac{H(Y \mid C)}{\log K}, \tag{1}$$

whereas semantic fragmentation occurs when one semantic identity is distributed across multiple clusters, measured by

$$D_{\mathrm{frag}}(\mathcal{C}; Y) = \frac{H(C \mid Y)}{\log K}. \tag{2}$$

The two errors imply opposite structural pressures. Reducing mixing requires separating heterogeneous content within a cluster, while reducing fragmentation requires aggregating semantically redundant clusters. Under a fixed cluster budget K, these corrections are inherently coupled.

However, Y is unavailable during unsupervised clustering. The central question is therefore how to obtain semantic evidence for these two directions without replacing geometric clustering by direct semantic labeling.

3.2 Sparse Semantic Evidence Construction

We obtain semantic evidence only from a sparse set of representatives. Geometry first determines where these representatives are placed and how much data each of them covers; an MLLM then provides semantic observations at those locations.

Geometric bootstrap and representative cover. For each view v, a frozen encoder produces a feature $z_i^{(v)}$, which is independently ℓ_2-normalized. Frozen feature blocks are normalized and fused with equal weights to obtain $\mathbf{f}_i$.

$$\mathbf{f}_i = \mathrm{Norm}\left([\sqrt{\omega_1}\bar{z}_i^{(1)}; \cdots; \sqrt{\omega_V}\bar{z}_i^{(V)}]\right), \qquad \omega_v = \frac{1}{V}, \tag{3}$$

where $\bar{z}_i^{(v)}$ is the normalized feature and ω_v is its fusion weight. Cosine K-means (MacQueen, 1967) yields the initial geometric partition $\mathcal{C}^{(0)}$.

Our goal is to correct this partition while preserving local geometric coherence.

We then select $B = mK$ observed representatives using a facility-location cover. One medoid is first retained from every initial cluster. For a candidate representative r in cluster $C_k^{(0)}$, its marginal coverage gain is

$$\Delta(r) = \sum_{i \in C_k^{(0)}} \left[d(i, \mathcal{R}_k) - \min\{d(i, \mathcal{R}_k), d(i, r)\}\right], \tag{4}$$

where $d(i, r) = 1 - \mathbf{f}_i^\top \mathbf{f}_r$, $\mathcal{R}_k$ is the set of representatives already selected in $C_k^{(0)}$, and $d(i, \mathcal{R}_k) = \min_{r' \in \mathcal{R}_k} d(i, r')$. The remaining representative budget is allocated by lazy-greedy selection of the largest marginal gain. Thus, complex clusters can receive more representatives than simple ones while every initial cluster remains covered.

Each sample is then assigned to its nearest representative within its initial cluster. Let $g(i) \in \mathcal{R}_k$ denote the representative assigned to sample $i \in C_k^{(0)}$. Representative $r \in \mathcal{R}_k$ therefore defines a local geometric group

$$U_r = \{i : g(i) = r\}, \qquad r \in \mathcal{R}_k, \qquad a_r = |U_r|, \tag{5}$$

where a_r is its coverage. The map g is frozen thereafter, so later semantic correction operates on these local groups rather than arbitrarily reassigning individual samples.

Semantic Pass 1: Card Generation. Stage II contains two MLLM passes, followed by embedding and semantic K-means. In Pass 1, the MLLM independently observes each representative's available modalities and produces a card with three fields:

$$\{\texttt{semantic_label}, \texttt{summary}, \texttt{keywords}\}.$$

The label gives a concise identity, the summary records factual evidence, and the keywords capture salient cues. The prompt suppresses incidental background, pose, viewpoint, and recording variation.

Semantic Pass 2: Anonymous Audit. Pass 2 receives only frozen cards, without reopening raw inputs, and maps representative r to a normalized semantic phrase

$$\phi_r = \mathcal{A}(r), \tag{6}$$

where $\mathcal{A}$ denotes the audit process. It normalizes synonymous descriptions while preserving materially different semantic identities.

A frozen semantic encoder maps each phrase to

$$\mathbf{e}_r = E(\phi_r),$$

and coverage-weighted K-means over these embeddings assigns every supported representative a semantic seed

$$s_r \in \{1, \dots, K\}. \tag{7}$$

Representatives judged uncertain contribute no semantic mass. Seeds are frozen before structural optimization and define an organization of sparse semantic evidence, not estimates of the unknown labels Y.

Let $\ell_r \in \{1, \dots, K\}$ denote the current cluster assignment of local group U_r, and let $\tilde{a}_r = a_r$ for semantically supported representatives and 0 otherwise. We define the coverage-weighted cluster–semantic contingency matrix

$$M_{cs} = \sum_r \tilde{a}_r \, \mathbf{1}[\ell_r = c] \, \mathbf{1}[s_r = s]. \tag{8}$$

After normalizing M into a joint distribution over cluster identity C and semantic seed S, we obtain two observable discrepancies:

$$H_w(S \mid C), \qquad H_w(C \mid S). \tag{9}$$

The former measures semantic heterogeneity within clusters, while the latter measures how broadly one semantic seed is dispersed across clusters.

Proposition 1 (Bidirectional semantic discrepancy)**.** *With equal weighting,*

$$\boxed{H_w(S \mid C) + H_w(C \mid S) = \mathrm{VI}_w(C, S),} \tag{10}$$

where VI_w *is the coverage-weighted variation of information between the current clustering and the semantic-evidence organization. Hence, the two directions form a symmetric partition discrepancy rather than two unrelated semantic penalties. The proof is given in Appendix C.3.*

3.3 Geometry-Constrained SemReC

With the local geometric units and semantic seeds fixed, Stage III revises the initial partition through coupled merge–split updates. Semantic seeds identify candidate subgroups for splitting, while a joint objective evaluates each update by balancing geometric connectivity against semantic mixing and fragmentation. Each update moves complete units and preserves exactly K nonempty clusters.

A cosine nearest-neighbor graph is first constructed on the fused sample features. Let $A_{ij} \geq 0$ denote the affinity between samples i and j. Using the frozen map g, we aggregate this graph into a quotient affinity between local groups:

$$Q_{rr'} = \sum_{i:g(i)=r} \sum_{j:g(j)=r'} A_{ij}. \tag{11}$$

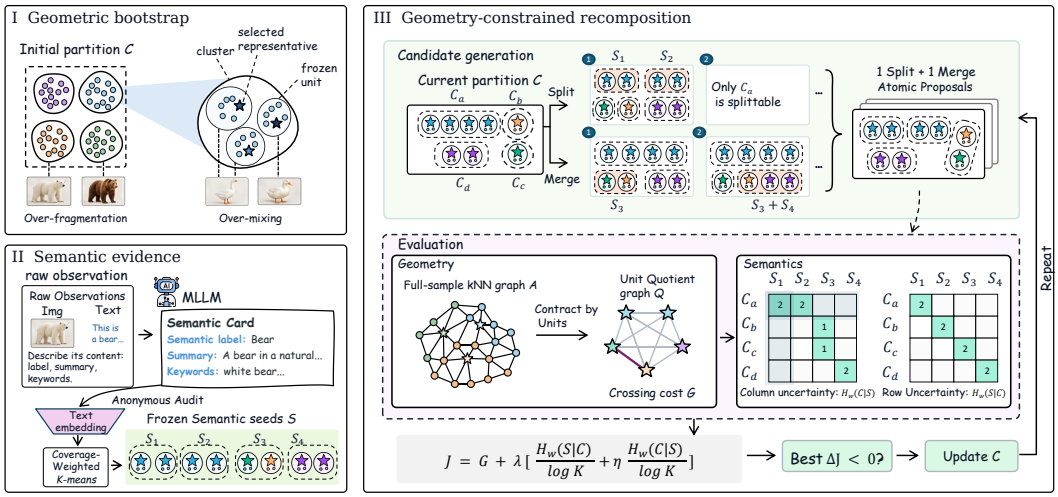

Figure 2: Framework overview: geometric bootstrap, sparse semantic evidence formation, and geometry-constrained merge–split recomposition with a fixed cluster count.

For the group-label vector

$$\boldsymbol{\ell} = (\ell_1, \ldots, \ell_B),$$

we define the normalized geometric crossing cost

$$G(\boldsymbol{\ell}) = \frac{\sum_{r<r'} Q_{rr'} \mathbf{1}[\ell_r \neq \ell_{r'}]}{\sum_{r<r'} Q_{rr'}}. \tag{12}$$

The final objective combines geometric connectivity with the two semantic directions:

$$\boxed{J(\boldsymbol{\ell}) = G(\boldsymbol{\ell}) + \lambda \left[\frac{H_w(S \mid C)}{\log K} + \eta \frac{H_w(C \mid S)}{\log K} \right],} \tag{13}$$

where λ controls semantic weight and η balances fragmentation against mixing. The hard restrictions preserve each unit's membership, exactly K nonempty clusters, and the specified merge–split rules. In contrast, G is a soft penalty: a move with $\Delta G > 0$ can be accepted if its semantic improvement yields $\Delta J < 0$.

Lemma 1 (Quotient-graph preservation)**.** *For any group-level partition lifted to samples by*

$$\hat{c}_i = \ell_{g(i)},$$

the crossing-edge mass computed on the quotient graph Q equals that of the corresponding sample-level partition on A. Contraction therefore preserves this objective within the allowed partition family (Appendix C.4).

The two semantic errors naturally suggest opposite structural edits: fragmentation creates pressure to merge clusters, whereas mixing creates pressure to split them. Suppose a structural correction contains n_m binary merges and n_s binary splits. Its cluster count becomes

$$K' = K - n_m + n_s. \tag{14}$$

Because the output must retain exactly K clusters,

$$K' = K \quad \Longrightarrow \quad n_m = n_s. \tag{15}$$

Proposition 2 (Minimal cardinality-preserving recomposition)**.** *Among nontrivial structural corrections composed of binary merges and splits, the minimum operation preserving exactly K clusters consists of one merge and one split.*

An atomic candidate merges a pair of current clusters and assigns the released label to a proper nonempty semantic-seed subgroup of a third cluster. All legal merge pairs are evaluated without a

semantic-similarity prefilter (Appendix C.5). The joint objective favors beneficial recompositions; merge candidates are not assumed semantically consistent.

For every legal candidate transformation T, we compute the exact objective change

$$\Delta J(T) = J(T(\boldsymbol{\ell})) - J(\boldsymbol{\ell}). \tag{16}$$

At each iteration, the candidate with the largest objective decrease is accepted if $\Delta J(T) < 0$; otherwise optimization terminates. Final sample assignments inherit the optimized label of their local geometric group:

$$\hat{c}_i = \ell_{g(i)}. \tag{17}$$

Thus, semantic evidence proposes splits, hard restrictions define legal moves, and the joint semantic–geometric objective determines acceptance.

4 Experiments

4.1 Experimental setup

We evaluate on COCO-2017 ($K = 44$), Caltech101-20 ($K = 20$), AwA2 ($K = 50$), and Fluent Speech Commands ($K = 31$) (Lin et al., 2014; Fei-Fei et al., 2004; Lampert et al., 2009; Lugosch et al., 2019). The frozen views are DINOv2 and ConvNeXt V2 image features with SBERT, TF–IDF/SVD, and GloVe text features for COCO; dense-SIFT and LBP for Caltech; DINOv2 and MAE for AwA2; and BEATs audio with BGE-M3 text for Fluent (Oquab et al., 2024; Woo et al., 2023; Reimers & Gurevych, 2019; He et al., 2021; Chen et al., 2022; 2024). Every method uses the same exported samples and feature blocks. We compare with CorreGen, CANDY, DIVIDE, ROLL, STCMC-UR, PGCL-DCL, DealMVC, and LLM-DAMVC. The representative budget is $B = 10K$; the graph uses $q = 30$; and $\lambda = \eta = 1$. We report means and sample standard deviations over seeds 17, 18, and 19. ACC, NMI, and ARI are evaluated only after predictions are saved. Appendix A specifies the data, parameters, and prompts.

4.2 Main Results

Table 1 summarizes the mean performance on all four benchmarks. SemReC achieves the highest ACC on COCO (90.51%), Caltech (87.16%), AwA2 (87.36%), and Fluent (74.85%), improving on the strongest baseline by 13.25, 41.35, 1.79, and 6.19 percentage points, respectively. The gains cover both correspondence-driven baselines and methods guided by internal assignments or pseudo-labels, which is consistent with the motivation that sparse raw-content semantic evidence can complement the geometric signals used to form the initial partition.

4.3 Component ablation

Table 2 examines the contributions of the semantic terms and geometric crossing penalty on Caltech. Removing the semantic terms from the acceptance objective reduces ACC from 87.16% to 52.61%, accompanied by substantial decreases in NMI and ARI. These terms evaluate candidate updates for within-cluster mixing and cross-cluster fragmentation, providing the semantic direction needed for partition correction. Removing the geometric penalty reduces ACC to 83.00%, with NMI and ARI also declining. The crossing penalty discourages updates that separate strongly connected units, complementing sparse semantic evidence with connectivity information from the full sample graph. Together, these results support the complementary roles of semantic guidance and geometric regularization.

4.4 Post-hoc diagnostics and accepted operations

To examine whether recomposition addresses the two targeted semantic errors, Figure 3 compares the initial and final Caltech partitions using normalized mixing, $H(Y \mid C)/\log K$, and fragmentation, $H(C \mid Y)/\log K$. Lower mixing indicates that clusters contain fewer distinct semantic identities, while lower fragmentation indicates that samples sharing one identity are concentrated in fewer clusters. Both quantities decrease in the final partition, alongside a lower optimization

Table 1: Clustering results (%), mean ± sample SD over seeds 17, 18, and 19. SemReC is bold.

COCO-2017				Caltech101-20			
Method	ACC	NMI	ARI	Method	ACC	NMI	ARI
CorreGen	77.26 ± 1.01	90.30 ± 0.43	76.65 ± 1.03	CorreGen	42.71 ± 2.63	61.81 ± 0.78	31.33 ± 0.11
CANDY	68.09 ± 0.36	87.02 ± 0.36	64.99 ± 0.35	CANDY	37.45 ± 1.08	56.70 ± 1.69	26.59 ± 0.74
DIVIDE	53.96 ± 0.28	78.83 ± 0.92	43.93 ± 2.53	DIVIDE	36.87 ± 1.04	58.84 ± 1.19	26.84 ± 1.37
ROLL	44.53 ± 0.80	66.64 ± 1.43	13.81 ± 1.09	ROLL	35.78 ± 0.43	55.46 ± 0.30	22.97 ± 0.48
STCMC-UR	62.10 ± 0.69	85.41 ± 0.58	64.23 ± 0.90	STCMC-UR	34.97 ± 1.26	47.18 ± 0.90	25.43 ± 1.77
PGCL-DCL	61.30 ± 2.25	85.51 ± 0.42	60.38 ± 0.97	PGCL-DCL	30.54 ± 1.38	49.18 ± 0.66	20.68 ± 1.01
DealMVC	65.45 ± 7.68	80.38 ± 6.42	65.04 ± 13.65	DealMVC	35.64 ± 1.18	50.17 ± 0.21	29.34 ± 0.65
LLM-DAMVC	72.25 ± 4.04	88.36 ± 0.57	73.23 ± 3.04	LLM-DAMVC	45.81 ± 2.88	54.95 ± 1.93	40.23 ± 1.59
SemReC	**90.51 ± 2.23**	**93.78 ± 0.85**	**91.72 ± 2.45**	**SemReC**	**87.16 ± 1.20**	**82.04 ± 0.91**	**93.02 ± 0.38**
AwA2				**Fluent Speech Commands**			
Method	ACC	NMI	ARI	Method	ACC	NMI	ARI
CorreGen	84.71 ± 1.46	92.48 ± 0.45	83.17 ± 1.23	CorreGen	44.57 ± 0.64	71.21 ± 0.42	37.08 ± 0.49
CANDY	83.81 ± 0.97	91.48 ± 0.47	81.94 ± 0.82	CANDY	45.43 ± 0.35	69.78 ± 1.23	35.82 ± 0.86
DIVIDE	78.18 ± 2.29	89.29 ± 0.87	74.33 ± 2.38	DIVIDE	44.48 ± 3.15	69.75 ± 2.54	35.53 ± 3.04
ROLL	69.15 ± 3.64	83.97 ± 1.46	56.02 ± 7.34	ROLL	36.88 ± 1.54	60.30 ± 5.00	24.26 ± 5.10
STCMC-UR	36.00 ± 2.94	60.33 ± 0.37	13.37 ± 2.00	STCMC-UR	29.91 ± 3.08	53.60 ± 3.65	16.50 ± 5.89
PGCL-DCL	65.57 ± 1.09	81.53 ± 0.45	59.86 ± 0.44	PGCL-DCL	43.69 ± 2.52	68.68 ± 0.98	35.45 ± 2.23
DealMVC	62.29 ± 8.58	80.82 ± 5.58	58.08 ± 9.73	DealMVC	26.39 ± 3.57	47.08 ± 6.02	18.01 ± 3.53
LLM-DAMVC	85.57 ± 1.49	90.43 ± 1.20	82.64 ± 1.68	LLM-DAMVC	68.66 ± 3.37	87.13 ± 1.83	66.36 ± 3.82
SemReC	**87.36 ± 0.56**	**94.12 ± 0.17**	**83.42 ± 0.45**	**SemReC**	**74.85 ± 4.00**	**88.42 ± 1.40**	**72.69 ± 5.98**

Table 2: Caltech ablation (%). Three-seed mean ± sample SD.

Variant	ACC	NMI	ARI
KMeans	46.98±2.66	66.29±1.43	38.00±3.34
w/o semantic terms	52.61±1.20	46.38±3.60	20.96±3.27
w/o geometric penalty	83.00±1.20	79.89±1.35	91.67±0.35
w/o $H(S \mid C)$	47.99±4.60	38.82±7.51	14.81±4.05
w/o $H(C \mid S)$	85.26±0.79	81.39±0.27	92.53±0.35
SemReC	**87.16±1.20**	**82.04±0.91**	**93.02±0.38**

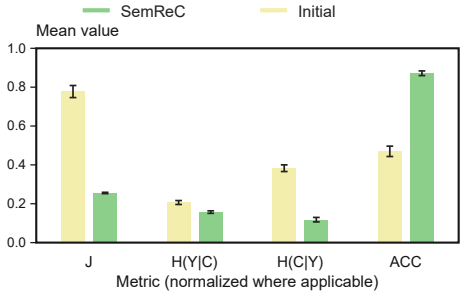

Figure 3: Initial and final Caltech partitions. Error bars show three-seed sample SD.

objective and higher ACC. These changes show that recomposition improves the partition in both semantic directions.

Figure 4 illustrates the underlying correction: two zebra clusters are merged while a motorcycle subgroup is separated from a mixed vehicle cluster, preserving the cluster count.

4.5 Additional controls, sensitivity, and cost

Figure 5 varies λ and η on Caltech with frozen semantic evidence and $q = 30$; $\lambda = 0$ is omitted because η is then inactive. Appendix B provides the Caltech example, accepted-operation trajectory, direct semantic-assignment control, and analyses of graph degree, permutation, fixed-unit oracle performance, and representative budget. Appendix A reports accepted MLLM seconds per semantic pass.

4.6 Frozen-baseline Stage 2/3 plugin

Table 3 evaluates Stages II and III on frozen partitions from CorreGen, CANDY, and DIVIDE. The plugin improves ACC, NMI, and ARI for all three baselines on both datasets. These results show that semantic recomposition can refine partitions produced by different clustering backbones.



Figure 4: Accepted COCO merge–split example (seed 17). Red borders mark separated units.

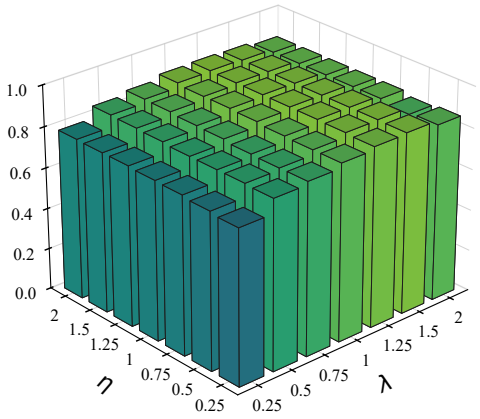

Figure 5: Joint sensitivity to λ and η on Caltech at $q = 30$. Bars show three-seed mean ACC.

Table 3: Frozen-baseline Stage 2/3 plugin results (%). Mean ± sample SD over seeds 17, 18, and 19.

		Frozen baseline			+ Stages II–III		
Baseline	Dataset	ACC	NMI	ARI	ACC	NMI	ARI
CorreGen	Caltech	42.71±2.63	61.81±0.78	31.33±0.11	85.18±1.66	80.40±0.89	91.83±0.81
CorreGen	COCO	77.26±1.01	90.30±0.43	76.65±1.03	89.16±1.19	93.33±0.17	90.58±1.07
CANDY	Caltech	37.45±1.08	56.70±1.69	26.59±0.74	83.81±0.46	79.40±0.94	90.61±1.43
CANDY	COCO	68.09±0.36	87.02±0.36	64.99±0.35	89.07±2.25	93.02±0.80	89.89±2.34
DIVIDE	Caltech	36.87±1.04	58.84±1.19	26.84±1.37	85.88±1.13	81.14±1.12	91.82±1.00
DIVIDE	COCO	53.96±0.28	78.83±0.92	43.93±2.53	89.44±1.83	92.57±0.47	90.28±1.62

5 Limitations and Reproducibility

Limitations. **Semantic scope.** The method assumes a known class count K, frozen views, and a fixed semantic granularity. A stronger semantic model might infer an appropriate granularity or adapt the evidence protocol when the class count and ontology are not prescribed. **Unit-level capacity.** Fixed local units constrain correction, and the oracle gaps retain unused capacity; Appendix B analyzes this limitation through the fixed-unit oracle and representative-budget controls.

Reproducibility. Appendix A.2 specifies the locked feature, clustering, graph, and semantic-pass settings, and Appendix A.3 provides the prompt templates and dataset-specific semantic contracts. The implementation, configurations, frozen evidence records, accepted operations, and evaluation scripts will be released with the paper.

6 Conclusion

We presented Semantic Recomposition (SemReC), which uses sparse semantic evidence to correct geometric multi-view partitions. Its coupled merge–split updates address semantic mixing and fragmentation while preserving local units and the cluster count. SemReC achieves higher clustering accuracy than the compared baselines across four benchmarks. The results highlight the value of raw-content semantic evidence for improving the semantic coherence of existing partitions. Future work will explore adaptive semantic granularity when the class count is unknown. Another direction is to jointly adapt representative selection and local units, enabling finer corrections under limited semantic-query budgets.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and polishing parts of the manuscript and with writing and debugging the visualization code. All AI-assisted text and code were reviewed, verified, and corrected by the authors. We have not used generative AI tools for research ideation, experiment design, data analysis, or the interpretation of results. Note that the studied method itself queries a multimodal LLM by design; those queries are part of the experimental system, not writing assistance. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16706–16715, 2023a. doi: 10.1109/ICCV51070.2023.01536. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Chen_Deep_Multiview_Clustering_by_Contrasting_Cluster_Assignments_ICCV_2023_paper.pdf.

Man-Sheng Chen, Ling Huang, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. Relaxed multi-view clustering in latent embedding space. *Information Fusion*, 2021. URL https://www.sciencedirect.com/science/article/abs/pii/S1566253520303821.

Man-Sheng Chen, Chang-Dong Wang, and Jian-Huang Lai. Low-rank tensor based proximity learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2023b. URL https://www.computer.org/csdl/journal/tk/2023/05/09714775/1B2CRSssXUA.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers, 2022. URL https://arxiv.org/abs/2212.09058.

Liang Du, Yukai Shi, Yan Chen, Peng Zhou, and Yuhua Qian. Fast and scalable incomplete multi-view clustering with duality optimal graph filtering. In *ACM Multimedia*, 2024. doi: 10.1145/3664647.3681346. URL https://dl.acm.org/doi/10.1145/3664647.3681346.

Shide Du, Chunming Wu, Zihan Fang, Wendi Zhao, Yilin Wu, Changwei Wang, and Shiping Wang. LargeMvC-Net: Anchor-based deep unfolding network for large-scale multi-view clustering. *arXiv preprint arXiv:2507.20980*, 2025. URL https://arxiv.org/abs/2507.20980.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178, 2004. doi: 10.1109/CVPR.2004.383.

Zhibin Gu, Zhenhao Zhong, Xi Zhang, and Bing Li. Constant degree matrix-driven incomplete multi-view clustering via connectivity-structure and embedding tensor learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/88438dc62fc5c8777e2b5f1b4f6d37a2-Abstract-Conference.html.

Jipeng Guo, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin. Rank consistency induced multiview subspace clustering via low-rank matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. URL https://pubmed.ncbi.nlm.nih.gov/33882005/.

Jipeng Guo, Yanfeng Sun, Junbin Gao, Yongli Hu, and Baocai Yin. Logarithmic schatten-p norm minimization for tensorial multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. URL https://jipengguo95.github.io/publications/.

Jipeng Guo, Yanfeng Sun, Xin Ma, Junbin Gao, Yongli Hu, Youqing Wang, and Baocai Yin. Globality meets locality: An anchor graph collaborative learning framework for fast multiview subspace clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. URL https://jipengguo95.github.io/publications/.

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/dbe81b08f7dc4dd8b43bc62dedfd9662-Abstract-Conference.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

Shizhe Hu, Guoliang Zou, Chaoyang Zhang, Zhengzheng Lou, Ruilin Geng, and Yangdong Ye. Joint contrastive triple-learning for deep multi-view clustering. *Information Processing and Management*, 2023. URL https://www.sciencedirect.com/science/article/pii/S0306457323000213.

Shizhe Hu, Binyan Tian, Weibo Liu, and Yangdong Ye. Self-supervised trusted contrastive multi-view clustering with uncertainty refined. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025a. doi: 10.1609/aaai.v39i16.33902. URL https://ojs.aaai.org/index.php/AAAI/article/view/33902.

Shizhe Hu, Chengkun Zhang, Guoliang Zou, Zhengzheng Lou, and Yangdong Ye. Deep multiview clustering by pseudo-label guided contrastive learning and dual correlation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b. doi: 10.1109/TNNLS.2024.3354731.

Wenyuan Kong, Zhibin Gu, and Bing Li. Dual-branch representations with dynamic gated fusion and triple-granularity alignment for deep multi-view clustering. In *International Conference on Learning Representations*, pp. 27677–27692, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2eb74636e69ca26ab6cba8b724b0776e-Abstract-Conference.html.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 951–958, 2009. doi: 10.1109/CVPR.2009.5206594.

Wei Lan, Yinghao Guo, Qingfeng Chen, Shichao Zhang, Shirui Pan, Huiyu Zhou, and Yi Pan. Coordinator: Semantic-coordinated dynamic fusion for multi-view clustering. *Information Fusion*, 132:104239, 2026. doi: 10.1016/j.inffus.2026.104239.

Xingfeng Li, Yuangang Pan, Yuan Sun, Yinghui Sun, Quansen Sun, Zhenwen Ren, and Ivor W. Tsang. Scalable unpaired multi-view clustering with bipartite graph matching. *Information Fusion*, 2025. URL https://www.sciencedirect.com/science/article/pii/S1566253524005645.

Cheng Liang, Wenchao Zang, Daoyuan Wang, and Fei Guo. High-order correlation and consistency-aware multi-view clustering via anchor graph learning. *Neural Networks*, 2026. URL https://www.sciencedirect.com/science/article/abs/pii/S0893608026002571.

Renjie Lin, Shide Du, Shiping Wang, and Wenzhong Guo. Multi-view clustering via optimal transport algorithm. *Knowledge-Based Systems*, 2023. URL https://www.sciencedirect.com/science/article/abs/pii/S0950705123007049.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024. doi: 10.1609/aaai.v38i13.29330. URL https://ojs.aaai.org/index.php/AAAI/article/view/29330.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. In *Proceedings of Interspeech*, 2019. URL https://arxiv.org/abs/1904.03670.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press, 1967. URL https://projecteuclid.org/euclid.bsmsp/1200512992.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Yalan Qin, Zhenjun Tang, Hanzhou Wu, and Guorui Feng. Flexible tensor learning for multi-view clustering with markov chain. *IEEE Transactions on Knowledge and Data Engineering*, 36: 1552–1565, 2024. doi: 10.1109/TKDE.2023.3305624. URL https://doi.org/10.1109/TKDE.2023.3305624.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks, 2019. URL https://arxiv.org/abs/1908.10084.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007. URL https://aclanthology.org/D07-1043/.

Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. ROLL: Robust noisy pseudo-label learning for multi-view clustering with noisy correspondence. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.02862.

Jie Wen, Zheng Zhang, Yong Xu, and Zuofeng Zhong. Incomplete multi-view clustering via graph regularized matrix factorization. In *ECCV Workshops*, 2018. URL https://openaccess.thecvf.com/content_eccv_2018_workshops/w23/html/Wen_Incomplete_Multi-view_Clustering_via_Graph_Regularized_Matrix_Factorization_ECCVW_2018_paper.html.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders, 2023. URL https://arxiv.org/abs/2301.00808.

Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, 26:9150–9162, 2024. doi: 10.1109/TMM.2024.3387298.

Wei Xia, Quanxue Gao, Qianqian Wang, Xinbo Gao, Chris Ding, and Dacheng Tao. Tensorized bipartite graph learning for multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. URL https://pubmed.ncbi.nlm.nih.gov/35786549/.

Haiming Xu and Qianqian Wang. LLM-DAMVC: A large language model assisted dynamic agent for multi-view clustering. In *Advances in Neural Information Processing Systems*, volume 38, pp. 119515–119534, 2025. doi: 10.52202/085713-3990. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ad48f017e6c3d474caf511208e600459-Abstract-Conference.html.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16030–16039, 2022. doi: 10.1109/CVPR52688.2022.01558. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Xu_Multi-Level_Feature_Learning_for_Contrastive_Multi-View_Clustering_CVPR_2022_paper.pdf.

Xihong Yang, Jiaqi Jin, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. DealMVC: Dual contrastive calibration for multi-view clustering, 2023. URL https://arxiv.org/abs/2308.09000.

Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. Uncover underlying correspondence for robust multi-view clustering. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8c1d92835eb4e601f396c97ec60439fe-Abstract-Conference.html.