BEYOND FEATURE GEOMETRY: SEMANTIC EVIDENCE FOR CONTRASTIVE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering (MVC) aims to discover semantically meaningful groups by exploiting consistency and complementarity across views. Many contrastive MVC methods construct supervision from learned representations or model predictions. However, feature geometry may capture attributes that do not align with the task-specific semantic criterion. Consequently, even high-confidence supervision can reinforce semantically incorrect relationships. We refer to this mismatch as *Semantic-Blind Correspondence* (SBC). To address SBC, we propose *Semantic Evidence for Contrastive Multi-View Clustering* (SECMVC), which introduces semantic evidence through a small budget of multimodal large language model (MLLM) queries. SECMVC first constructs verified semantic anchors under a fixed task criterion and partitions the remaining samples into reliable and unreliable sets. Reliable samples learn directly from semantic prototypes without additional queries. For unreliable samples, SECMVC queries a representative subset and uses confirmed assignments to guide reference-based contrastive learning, while uncertain responses provide no semantic supervision. In this way, SECMVC propagates semantic evidence when feature-based assignments are reliable and directs MLLM queries toward uncertain cases. Experiments on four multimodal clustering benchmarks demonstrate the effectiveness of SECMVC.

1 INTRODUCTION

Multimodal data describe the same instances through heterogeneous modalities such as images, text, and audio, providing complementary views of their underlying semantics. Multi-view clustering (MVC) seeks to jointly exploit these views to discover a common partition without category annotations. Deep MVC methods learn representations that capture shared cluster structure across views (Trosten et al., 2021; Lu et al., 2024); contrastive learning shapes the space by attracting positive pairs and separating negative pairs. Accordingly, the quality of the supervisory relationships that define these pairs plays a central role in contrastive MVC (Yang et al., 2023; Guo et al., 2024).

However, a supervisory relationship can be reliable in feature space while remaining inconsistent with the semantic criterion of the clustering task. As illustrated in Fig. 1, consider an office chair, a dining chair, an office desk, and a dining table. These items can be grouped by furniture type into chairs and tables, or by intended use into office and dining furniture. If the representation mainly captures furniture type while the task requires intended use, the chairs form a positive pair while the office chair and desk are separated. Both relationships are wrong under the task criterion, though consistent with the feature geometry. We refer to this mismatch between geometry-derived supervision and task semantics as *Semantic-Blind Correspondence* (SBC). Training on such relationships can reinforce a partition that is coherent in the embedding space but inconsistent with the task.

Recent studies have improved contrastive MVC by refining pair relationships or filtering unreliable pseudo-labels. DIVIDE uses high-order random walks to construct positive and negative relationships beyond local neighborhoods (Lu et al., 2024), while DealMVC guides contrastive learning with high-confidence pseudo-labels (Yang et al., 2023). STCMC-UR further reduces unreliable supervision through uncertainty estimation (Hu et al., 2025a). Although these methods use different strategies, their supervisory signals are still derived from learned representations or model predictions. They can therefore identify relationships that are uncertain in feature space, but cannot determine whether a geometrically reliable relationship follows the semantic criterion specified by the task.

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.

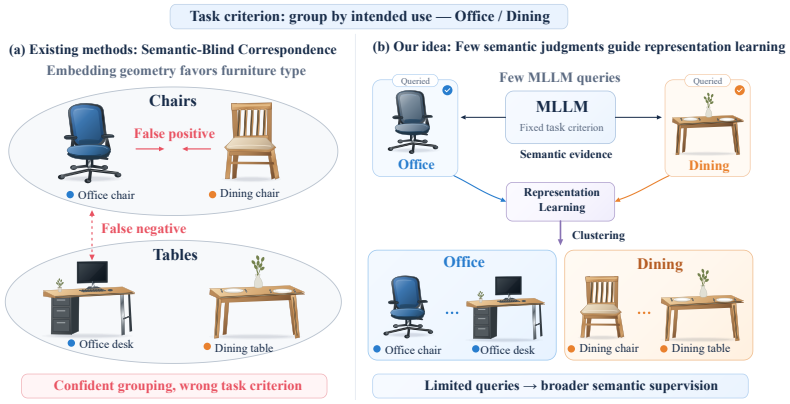


Figure 1: Semantic-blind correspondence. Feature geometry groups the chairs yet separates the office chair from the desk; under an intended-use criterion both relations are wrong.

To address SBC, we use multimodal large language models (MLLMs) to obtain semantic evidence by examining sample content in light of the task-specific semantic criterion. However, querying every sample would be costly and would leave the potential to share this evidence across related samples unexplored. The challenge is therefore to extend the supervision provided by a limited number of semantic judgments to samples that have not been queried.

We propose *Semantic Evidence for Contrastive Multi-View Clustering* (SECMVC) to address SBC under a limited MLLM query budget. SECMVC first queries representative samples to construct verified semantic anchors under a fixed task criterion. It then partitions the remaining samples into reliable and unreliable sets using cross-view consensus and prototype-margin separation. Reliable samples learn directly from semantic prototypes, allowing verified evidence to be propagated without additional queries. For unreliable samples, SECMVC selects a representative subset for direct MLLM assessment and uses confirmed assignments to guide reference-based contrastive learning. Uncertain responses provide no semantic supervision. In this way, SECMVC propagates semantic evidence only when the current representation provides consistent support and reserves additional queries for assignments that remain ambiguous. Our main contributions are summarized as follows:

- We identify *Semantic-Blind Correspondence* (SBC) in contrastive MVC, where supervisory relationships can remain reliable in feature space while conflicting with the task-specific semantic criterion. Unlike conventional noisy correspondence, SBC can therefore persist even when feature-space relationships are confident and internally consistent.
- We propose *Semantic Evidence for Contrastive Multi-View Clustering* (SECMVC), which introduces criterion-conditioned semantic evidence under a limited MLLM query budget. SECMVC constructs verified semantic anchors and uses reliability-aware two-path learning: reliable samples learn from semantic prototypes, while uncertain assignments are selectively verified and guided by concrete semantic references.
- Experiments on four benchmarks against eighteen baselines verify the effectiveness of SECMVC: it improves ACC over the strongest baseline by 8.2 to 32.2 points and achieves the best ACC on all four datasets.

2 Related Work

Contrastive Multi-View Clustering. Deep MVC learns representations that capture consistency and complementarity across views. Contrastive learning is a widely used means to this end, since it shapes representations by encouraging agreement between positive pairs and separating negative pairs (Tian et al., 2020; Trosten et al., 2021). However, inaccurate pair relationships can introduce misleading supervision and impair clustering. Existing methods address this issue by refining sample relationships or calibrating contrastive supervision. DIVIDE uses high-order random walks to identify positive and negative relationships beyond local neighborhoods (Lu et al., 2024). CANDY combines cross-view similarity context with spectral denoising to mitigate false negative and false positive

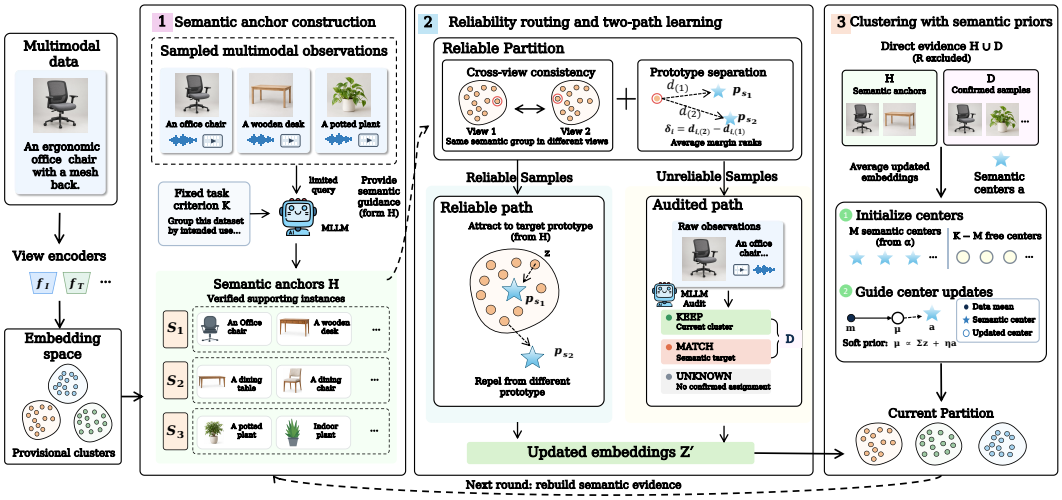


Figure 2: Overview of SECMVC. Representative samples form verified semantic anchors. The remaining samples are split into a reliable set, which learns from semantic prototypes, and an unreliable set, from which a small audited subset learns from frozen concrete references.

correspondences (Guo et al., 2024). DealMVC aligns feature similarities with high-confidence pseudo-label relationships through local and global contrastive calibration (Yang et al., 2023). SEM instead down-weights unreliable samples through self-weighted contrastive learning to resist representation degeneration (Xu et al., 2023a), and GCFAgg aggregates global and cross-view features to strengthen the shared consensus (Yan et al., 2023). Beyond pair calibration, sample-level graph clustering refines relationship construction by learning a sample-level graph that captures higher-order affinities among samples (Tan et al., 2023), and anchor-based strategies enhance anchor graphs from cross-view correlation for scalable clustering (Liu et al., 2024). In the same scalable vein, EASEMVC selects informative instances and clusters through a dual selection mechanism to reduce redundancy (Xiao et al., 2025), while granular-ball and universum constructions improve robustness by coarsening noisy pairs and introducing pseudo-label-guided universum samples (Su et al., 2025; Wang et al., 2025). These methods improve supervision using learned representations and model predictions. Nevertheless, reliable relationships within the embedding space may still conflict with the task-specific semantic criterion. SECMVC addresses this mismatch by introducing external semantic evidence to refine contrastive supervision.

Clustering with External Semantic Guidance. External semantic knowledge can help align clustering with task-relevant meanings. TAC transfers WordNet semantics through cross-modal neighborhood distillation (Li et al., 2024), while IVS-GMV constructs vision-language semantic graphs for multi-view clustering (Ke et al., 2024). LiSA uses LLM-generated topic words and LLM-assisted assignment correction to align document and topic spaces through consistency and neighborhood matching objectives (Liu et al., 2025b). Related work also incorporates user-specified criteria. CRL transforms image representations using a criterion-specific semantic basis (Liu et al., 2025a), while the guideline-driven image clustering agent generates criterion-aware descriptions and uses selective LLM queries to guide cluster merging (Zhong et al., 2026). These methods establish several ways to incorporate external semantics into clustering. SECMVC addresses the complementary challenge of extending a few semantic judgments to supervision for many unqueried samples: semantic prototypes supervise samples with reliable cross-view assignments, additional queries target a representative subset of uncertain samples, and confirmed assignments provide instance-level contrastive supervision.

3 METHOD

SECMVC uses a limited number of MLLM queries to align contrastive supervision with the task-specific semantic criterion. As illustrated in Fig. 2, the framework consists of three stages repeated across rounds. Before the loop, view-specific encoders map multimodal observations into an em-

bedding space, and warm-up training plus spherical K-means produce an initial partition. Stage 1 constructs semantic anchors: SECMVC selects samples nearest to each cluster center and queries the MLLM on their raw multimodal observations under the fixed task criterion. Stage 2 routes and trains: cross-view consensus and prototype-margin separation split the remaining samples into reliable and unreliable sets; reliable samples learn from semantic prototypes, and a representative audited subset of unreliable samples learns from confirmed reference assignments. Stage 3 applies fixed-K spherical clustering with semantic priors: semantic centers derived from directly supported samples initialize and softly guide the cluster centers. The three stages repeat across rounds to refine the partition.

3.1 Problem Formulation

Let $\mathcal{X} = \{X_i\}_{i=1}^N$ be an unlabeled dataset. Each instance has raw observations $o_i = (o_i^{(1)}, \ldots, o_i^{(M)})$ from M modalities and V aligned feature views $\{x_i^{(v)}\}_{v=1}^V$, where $x_i^{(v)} \in \mathbb{R}^{d_v}$. Multiple feature views may originate from the same modality. SECMVC uses feature views for representation learning and clustering, and raw observations for semantic assessment.

Given a task criterion κ and a cluster count K, our goal is to learn representations supporting a common partition $\mathbf{c} \in \{1, \ldots, K\}^N$ consistent with κ. The criterion specifies which attributes determine the partition and which should be ignored. For each unordered pair $q = \{i, j\}$, define $s_\kappa(q) = 1$ if the two instances should share a cluster under κ, and $s_\kappa(q) = 0$ otherwise. These relations are not observed during training. Geometry-derived contrastive supervision may attract pairs with $s_\kappa(q) = 0$ or repel pairs with $s_\kappa(q) = 1$. We call this mismatch *Semantic-Blind Correspondence* (SBC). The following result formalizes why criterion-independent information alone cannot resolve all such mismatches.

Theorem 1 (Criterion-agnostic supervision limit). Fix $\mathcal{X}$ and criterion-independent information $\mathcal{G}(\mathcal{X})$. Let $\kappa \sim \mu$ and $\mathcal{Q} = \{\{i, j\} : 1 \le i < j \le N\}$. Consider any possibly randomized supervision rule A satisfying $\widehat{\mathbf{s}}_A \perp \kappa \mid \mathcal{G}(\mathcal{X})$. Define $p_q = \Pr_{\kappa\sim\mu}[s_\kappa(q) = 1]$ and its expected pairwise error as

$$R_\mu(A) = \mathbb{E}_{\kappa,A}\left[\frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathbf{1}[\widehat{s}_A(q) \neq s_\kappa(q)]\right]. \tag{1}$$

Then

$$R_\mu(A) \ge \mathcal{A}_\mu(\mathcal{X}) := \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \min\{p_q, 1 - p_q\}. \tag{2}$$

If $\mathcal{A}_\mu(\mathcal{X}) > 0$, no such rule can recover the desired relations for every criterion in the support of μ. The proof is provided in Appendix A.1. The distribution μ describes alternative criteria for the same data; SECMVC uses one fixed criterion per run. Theorem 1 motivates introducing criterion-conditioned evidence, rather than relying exclusively on feature geometry.

For each view, an encoder $f_{\theta_v} : \mathbb{R}^{d_v} \to \mathbb{R}^d$ produces

$$h_i^{(v)} = f_{\theta_v}(x_i^{(v)}), \qquad z_i^{(v)} = \nu(h_i^{(v)}), \qquad \bar{z}_i = \nu([z_i^{(1)}; \ldots; z_i^{(V)}]), \tag{3}$$

where ν denotes ℓ_2 normalization and $[\cdot;\cdot]$ denotes concatenation. The embeddings $z_i^{(v)}$ support learning and reliability assessment, while $\bar{z}_i$ is used for clustering. A frozen pretrained MLLM assesses selected observations o_i under κ to construct semantic supervision. Both the MLLM and κ remain fixed throughout training.

3.2 Constructing Semantic Anchors

SECMVC constructs semantic anchors by querying an MLLM on samples selected from the current clustering. The clustering identifies candidate groups, while the MLLM determines their shared meaning and supporting samples under the task criterion κ. Before the first round, we warm up the representation model using reconstruction and cross-view learning, then apply spherical K-means (MacQueen, 1967; Dhillon & Modha, 2001) to the fused embeddings. In subsequent rounds, we use the partition obtained by reclustering after the preceding representation update. Let c_i denote the current assignment of instance i and $C_k = \{i : c_i = k\}$ the corresponding cluster. We denote the per-view embeddings at the beginning of the round by $\widetilde{z}_i^{(v)}$.

Criterion-conditioned evidence. For each cluster C_k, we compute its normalized fused center and select up to n_a members with the highest cosine similarity to that center. Let $\mathcal{A}_k$ denote this construction subset. The MLLM summarizes the available raw modalities of these instances and combines the summaries into factual evidence. Conditioned on κ, it then identifies a specific shared meaning and its supporting instance IDs $S_k \subseteq \mathcal{A}_k$, or abstains. A candidate is accepted only if its description is nonempty and specific, its support count satisfies $|S_k| \geq n_{\min}$, and its support ratio $\pi_k = |S_k|/|\mathcal{A}_k|$ satisfies $\pi_k \geq \pi_{\min}$. Here, π_k measures support within the selected subset, rather than the purity of the entire cluster.

Semantic identity. Different clusters may express the same meaning under κ. To avoid duplicate audit targets, we process candidates in descending order of $|C_k|\pi_k$ and compare each with previously accepted meanings. A new meaning receives a semantic ID; a uniquely matched duplicate adds its nonconflicting support samples to the existing identity. Duplicates are identified by identical normalized canonical descriptions or MLLM confirmation under κ; unresolved candidates are discarded. Only explicit support samples receive semantic assignments, not entire clusters.

Semantic anchors. Let $\mathcal{I} = \{1, \ldots, K_s\}$ denote the accepted semantic identities, where $K_s \leq K$, and let H_u contain the support samples assigned to identity u. Write $H = \bigcup_{u \in \mathcal{I}} H_u$ and $y_j = u$ for $j \in H_u$. Separate MLLM judgments under κ establish a symmetric relation $\mathcal{E}_{\neq}$ of explicitly different meanings, with $\Delta(u) = \{w : (u, w) \in \mathcal{E}_{\neq}\}$. Distinct IDs alone do not imply a negative relation; unresolved relations remain unspecified. Each semantic description and its support set H_u form an anchor, from which we construct per-view prototypes and reference embeddings:

$$p_u^{(v)} = \nu\left(\frac{1}{|H_u|}\sum_{j \in H_u} \widetilde{z}_j^{(v)}\right), \qquad b_j^{(v)} = \widetilde{z}_j^{(v)}, \quad j \in H_u. \tag{4}$$

3.3 Reliability Partition

Semantic anchors provide criterion-conditioned reference points, but nearest-prototype assignments remain provisional, since the current representation may still encode variations unrelated to κ. Treating all such assignments as supervision can propagate pseudo-label errors through representation learning. SECMVC uses an assignment for direct prototype supervision only when cross-view agreement and a clear separation from competing prototypes support it; the remaining samples are routed to the audit path.

Cross-view consensus. The partition is computed on the samples that are not construction supports, $\mathcal{O} = \{1, \ldots, N\} \setminus H$. For each view, a sample is assigned to its nearest semantic prototype $q_i^{(v)} = \arg\min_{u \in \mathcal{I}} d_{iu}^{(v)}$, where $d_{iu}^{(v)} = 1 - \langle z_i^{(v)}, p_u^{(v)} \rangle$ is the cosine distance. An assignment is consensus only when every view selects the same semantic:

$$g_i = \frac{1}{V}\max_{u \in \mathcal{I}} \sum_{v=1}^{V} \mathbf{1}\left[q_i^{(v)} = u\right], \tag{5}$$

where $g_i = 1$ means full cross-view agreement and $g_i < 1$ means at least one view dissents. The consensus candidate set is $\mathcal{C} = \{i \in \mathcal{O} : g_i = 1\}$; samples outside it are excluded from prototype supervision.

Separation from competing prototypes. Cross-view agreement alone may still retain samples near a prototype decision boundary. For each $i \in \mathcal{C}$ and view v, let $d_{i,(1)}^{(v)} \leq d_{i,(2)}^{(v)}$ be the smallest and second-smallest cosine distances from $z_i^{(v)}$ to the semantic prototypes, and define the prototype margin $\delta_i^{(v)} = d_{i,(2)}^{(v)} - d_{i,(1)}^{(v)}$. We rank these margins within each view and average the ranks into a reliability score $\rho_i \in [0, 1]$. Instead of hand-picking a threshold, we split the score distribution with a one-dimensional Otsu criterion, which chooses the threshold θ_{Otsu} that maximizes the between-group variance of $\{\rho_i\}_{i \in \mathcal{C}}$ (details in Appendix A.4). The reliable and unreliable sets are then $R = \{i \in \mathcal{C} : \rho_i \geq \theta_{\mathrm{Otsu}}\}$ and $U = \mathcal{O} \setminus R$.

3.4 Two-Path Representation Learning

The reliability partition determines how semantic evidence guides representation learning. Reliable samples learn from semantic prototypes, while a representative subset of unreliable samples is audited and learns from reference instances when its assignments are confirmed.

3.4.1 Prototype Learning for Reliable Samples

For $i \in R$, let q_i denote its assigned semantic identity. We encourage each reliable sample to approach its assigned prototype and separate from prototypes with explicitly different meanings:

$$\mathcal{L}_R = -\frac{1}{|R|V} \sum_{i \in R} \sum_{v=1}^{V} \log \frac{\exp(\langle z_i^{(v)}, p_{q_i}^{(v)} \rangle / \tau)}{\sum_{u \in \{q_i\} \cup \Delta(q_i)} \exp(\langle z_i^{(v)}, p_u^{(v)} \rangle / \tau)}, \tag{6}$$

where τ is the temperature and $\Delta(q_i)$ contains semantics confirmed to differ from q_i. This objective extends the semantic evidence in the anchor supports to reliable samples. SECMVC restricts this supervision to the reliable set selected in Section 3.3, since nearest-prototype assignments are not reliable for every sample.

3.4.2 Audit-Guided Learning for Unreliable Samples

For unreliable samples, feature geometry does not provide sufficiently consistent evidence for direct prototype supervision, so the MLLM assesses a representative subset of them:

$$\mathcal{T}_k = \underset{i \in C_k \cap U_{\text{new}}}{\text{Top-}a_k} \left[\frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k(i)} \langle \bar{z}_i, \bar{z}_j \rangle \right], \tag{7}$$

where U_{new} contains previously unaudited samples in U, $\mathcal{N}_k(i)$ denotes the nearest neighbors of i within C_k, excluding i itself, and Top-a_k returns the indices of the a_k highest-scoring candidates. The per-cluster allocations a_k are capped and satisfy $\sum_k a_k \le B_{\text{audit}}$.

For each selected sample, the MLLM first interprets its raw multimodal observations under κ, without access to the accepted semantic inventory. When this interpretation is judged uncertain, a second query matches it against the accepted meanings; when it is already decisive, the sample proceeds with a single query. A sample whose interpretation is confirmed to match an accepted meaning receives that semantic target. Let $D \subseteq U$ denote the samples with verified matches, with y_i denoting the confirmed semantic assignment of $i \in D$. Samples without a verified match receive no semantic target. Each confirmed sample then learns from the instances that support its assigned meaning. These reference instances provide concrete examples of what the sample should approach in the representation space. References with explicitly different meanings provide negative targets. Let $J_u(i) \subseteq H_u \setminus \{i\}$ denote the selected references for semantic u. For a confirmed sample i, the positive and negative reference sets are $\mathcal{P}_i^H = J_{y_i}(i)$ and $\mathcal{N}_i^H = \bigcup_{u \in \Delta(y_i)} J_u(i)$, respectively. Their reference embeddings $b_j^{(v)}$, constructed at the beginning of the round, remain fixed while the encoders update the sample embedding $z_i^{(v)}$. The reference-based contrastive loss over the confirmed samples is

$$\mathcal{L}_D = -\frac{1}{|D|V} \sum_{i \in D} \sum_{v=1}^{V} \frac{1}{|\mathcal{P}_i^H|} \sum_{j \in \mathcal{P}_i^H} \log \frac{\exp(\langle z_i^{(v)}, b_j^{(v)} \rangle / \tau)}{\sum_{k \in \mathcal{P}_i^H \cup \mathcal{N}_i^H} \exp(\langle z_i^{(v)}, b_k^{(v)} \rangle / \tau)}. \tag{8}$$

This objective attracts each confirmed sample toward the references supporting its meaning and separates it from references with explicitly different meanings. By updating the shared encoders, semantic targets for a small set of representative samples also reshape the representations of other samples, allowing limited semantic evidence to influence the broader representation space.

3.4.3 Joint Optimization

The two learning paths introduce task-relevant semantic supervision into each view. To coordinate their effects across views, we encourage representations of the same instance to remain aligned:

$$\mathcal{L}_{\text{cross}} = \frac{1}{NV(V-1)} \sum_{i=1}^{N} \sum_{v \neq w} \left(1 - \left\langle z_i^{(v)}, z_i^{(w)} \right\rangle\right), \tag{9}$$

where $z_i^{(v)}$ denotes the unit-normalized representation of instance i in view v. This objective couples the view-specific representations, encouraging semantic refinement within individual views to preserve cross-view consistency. It applies to all instances, including those without an accepted semantic assignment.

We further retain information from each input view through reconstruction, so that representation learning remains informed by the observed data as semantic supervision reshapes the embedding space. For the normalized input $\widetilde{x}_i^{(v)} = \nu(x_i^{(v)})$, a view-specific decoder reconstructs $\widehat{x}_i^{(v)} = g_{\phi_v}(h_i^{(v)})$ from the encoder output $h_i^{(v)}$. The reconstruction objective is

$$\mathcal{L}_{\text{rec}} = \frac{1}{NV} \sum_{i=1}^{N} \sum_{v=1}^{V} \frac{\|\widehat{x}_i^{(v)} - \widetilde{x}_i^{(v)}\|_2^2}{d_v}. \tag{10}$$

Combining these objectives with the two semantic learning paths gives

$$\mathcal{L} = \lambda_{\text{cross}} \mathcal{L}_{\text{cross}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_R \mathcal{L}_R + \lambda_D \mathcal{L}_D, \tag{11}$$

where the coefficients control the relative contributions of the four objectives. Cross-view alignment and reconstruction preserve consistency and input information across the dataset, while $\mathcal{L}_R$ and $\mathcal{L}_D$ guide reliable and confirmed audited samples toward their semantic targets. Jointly optimizing the encoders and decoders therefore integrates task-specific semantic guidance with multi-view representation learning.

3.4.4 Clustering with Semantic Priors

Each round ends with full-N fixed-K spherical K-means on the fused embeddings $\bar{z}_i$. For each accepted semantic u, the fused center of its direct evidence $H_u \cup D_u$ provides one initial cluster center; the remaining $K - K_s$ centers come from k-means++. During the updates, each anchored cluster is additionally pulled toward its semantic center with a soft prior of weight $w = 20$. Verified evidence therefore steers the partition while the data geometry of all N samples keeps the clustering grounded.

4 Experiments

We evaluate SECMVC by addressing five questions. (Q1) Does SECMVC outperform state-of-the-art MVC methods when feature geometry partially conflicts with the task-specific semantic criterion (Section 4.2)? (Q2) Do the two learning paths contribute, and can auditing a small representative subset refine a large unreliable set (Section 4.3)? (Q3) Can SECMVC adapt to different semantic criteria on the same data, while geometry-based methods cannot (Appendix A.7)? (Q4) How does the audit budget affect performance (Appendix A.8)? (Q5) How sensitive is SECMVC to hyperparameters (Appendix A.9)?

4.1 Setup

Datasets. We evaluate on four public benchmarks: Microsoft Common Objects in Context (COCO) (Lin et al., 2014), Caltech101-20, a 20-class subset of Caltech-101 (Fei-Fei et al., 2004; Tao et al., 2017), Fluent Speech Commands (Lugosch et al., 2019), and UCF50, the UCF YouTube human-action dataset (Reddy & Shah, 2013). Dataset statistics, construction, and feature extraction details are given in Appendix A.2.

Baselines. We compare against eighteen methods in three families. Geometric methods construct pairs from learned representations: SDMVC (Xu et al., 2023b), DEMVC (Xu et al., 2021), SiMVC and CoMVC (Trosten et al., 2021), MFLVC (Xu et al., 2022), and CVCL (Chen et al., 2023). Correspondence-robust methods reweight pairs or views in feature space: CANDY (Guo et al., 2024), DIVIDE (Lu et al., 2024), ROLL (Sun et al., 2025), DealMVC (Yang et al., 2023), PAUSE (Wang et al., 2025), PGCL-DCL (Hu et al., 2025b), AICN-MLM (Shu et al., 2025), DUA-Nets (Geng et al., 2021), STCMC-UR (Hu et al., 2025a), DCMVC (Cui et al., 2024), and CSOT (Zhang et al., 2024). ALPC (Chen et al., 2025) is anchor-based. Baselines do not use the MLLM or semantic paths.

Table 1: Main results (%). Best results are in bold and second-best results are underlined. All methods use the same views; only SECMVC uses MLLM queries.

Method	COCO			Caltech101-20			Fluent			UCF50		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
SDMVC	69.78	87.34	72.68	40.65	50.17	31.89	38.57	66.30	30.13	27.26	42.04	14.37
DEMVC	75.85	88.26	78.26	36.34	42.29	32.42	<u>38.78</u>	<u>67.25</u>	<u>32.90</u>	28.36	47.38	17.63
SiMVC	61.43	84.30	62.32	34.03	55.11	24.67	29.44	54.58	20.69	24.25	35.91	12.12
CoMVC	61.44	84.51	62.27	35.00	53.37	25.76	32.89	58.72	23.08	24.91	39.40	13.19
MFLVC	<u>79.01</u>	88.59	81.99	44.26	48.71	<u>40.78</u>	36.50	65.67	29.77	29.23	46.40	18.13
CVCL	61.68	85.10	57.34	33.07	27.98	20.71	35.07	60.99	24.91	37.14	53.06	24.53
CANDY	72.98	88.65	72.20	40.44	57.43	26.51	34.72	59.49	23.95	30.97	43.69	18.44
DIVIDE	60.68	81.59	57.90	40.40	58.56	28.39	33.69	55.55	22.36	31.07	43.50	17.57
ROLL	57.46	81.91	54.52	36.92	<u>63.20</u>	25.76	34.56	55.60	24.31	<u>54.29</u>	<u>67.09</u>	<u>40.94</u>
DealMVC	66.45	83.17	68.89	40.07	49.33	28.63	30.24	56.30	21.74	27.38	43.16	16.16
PAUSE	65.80	86.10	64.10	43.00	62.74	33.46	34.37	59.32	27.05	26.07	37.53	14.77
PGCL-DCL	63.56	86.07	63.16	33.40	55.91	23.64	36.19	63.44	27.37	33.27	48.61	19.56
AICN-MLM	76.11	89.58	<u>82.21</u>	35.58	59.12	29.58	35.53	63.64	28.72	25.85	39.74	13.92
DUA-Nets	71.82	89.83	72.65	<u>45.39</u>	61.07	37.13	36.62	62.06	28.06	34.99	48.07	21.08
STCMC-UR	69.26	85.34	70.65	43.71	56.54	34.88	35.82	64.21	28.45	24.34	44.93	10.86
DCMVC	76.27	<u>90.56</u>	75.52	36.84	58.65	28.13	37.87	67.18	31.01	36.55	47.90	22.81
CSOT	62.46	85.38	61.93	24.81	42.96	17.61	34.25	61.80	26.60	24.11	40.09	12.52
ALPC	68.83	87.38	68.23	44.76	55.50	28.89	36.25	64.69	28.05	46.46	57.83	30.36
SECMVC (ours)	**87.22**	**92.75**	**88.51**	**77.54**	**75.03**	**85.41**	**50.19**	**71.52**	**39.14**	**62.88**	**68.57**	**48.80**

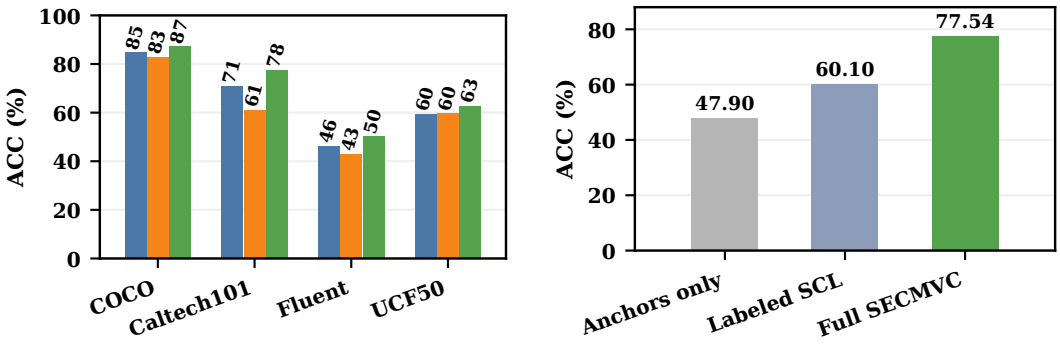
Implementation details. We use Qwen3-Omni (Xu et al., 2025) to extract semantic evidence from raw samples and audit selected instances. The audit budget is $B_{\text{audit}} = 120$ per round; we sample 16 instances per cluster for semantic extraction and train for 10 semantic rounds of 20 epochs each. Further training, inference, and export details are provided in Appendices A.5 and A.3.

4.2 Main Results (Q1)

Table 1 reports clustering performance on four benchmarks. From these results, we obtain the following observations. (1) SECMVC achieves the best ACC, NMI, and ARI on all four datasets, improving ACC over the strongest baseline by 8.21 to 32.15 points. Since all methods share the same views and training budget and only SECMVC receives semantic supervision, these gains come from semantic targets that redirect geometry-derived supervision conflicting with the task criterion, not from a stronger feature extractor. (2) Correspondence-robust baselines do not consistently dominate the geometric ones, and their ranking varies sharply across datasets. On Fluent, the two strongest baselines are geometric (SDMVC 38.57 and DEMVC 38.78), while ROLL, the strongest baseline on UCF50 (54.29), reaches only 57.46 ACC on COCO, the lowest among all baselines. Because these methods refine pairs with the same geometry that produces them, they cannot flag a relationship that is geometrically reliable yet semantically wrong; the remaining error is criterion-level. (3) The gain tracks the distance between the available views and the criterion axis. On Caltech101-20, whose GIST and HOG views describe appearance structure rather than object identity, the strongest baseline reaches 45.39 ACC and SECMVC improves it by 32.15 points; on UCF50, whose motion descriptors already align well with fine-grained actions, the gain is 8.59 points over the strongest baseline (54.29). Semantic evidence therefore matters most where feature geometry is furthest from the criterion. (4) SECMVC is the only method that leads on all three metrics simultaneously. On Caltech101-20, three different baselines hold the best ACC, NMI, and ARI (DUA-Nets, ROLL, and MFLVC), while SECMVC holds all three on every dataset: criterion-conditioned evidence aligns the metrics instead of trading one for another.

4.3 Ablation Study (Q2)

We remove one learning path at a time under the frozen-evidence protocol: each variant restores the recorded warm-up checkpoint and replays the live semantic snapshot without new MLLM calls. The



(a) Reliable-only (blue), audit-only (orange), and full two-path (green) ACC.

(b) Caltech101-20 with the same frozen evidence used in three ways.

Figure 3: Q2 path ablation under the frozen-evidence protocol. (a) Each single path falls short of the full two-path objective. (b) On Caltech101-20, the same evidence through the two paths beats anchor-only initialization or one supervised contrastive loss.

reliable variant keeps $\mathcal{L}_R$ with $\lambda_D = 0$; the audit variant keeps $\mathcal{L}_D$ with $\lambda_R = 0$; everything else—warm-up, reconstruction, cross-view learning, path weights, and training budget—stays unchanged.

Fig. 3(a) leads to two observations. First, either removal lowers ACC on all four datasets: removing the reliable path costs 4.22, 16.35, 7.12, and 3.14 points on COCO, Caltech101-20, Fluent, and UCF50, respectively, and removing the audit path costs 2.31, 6.63, 3.86, and 3.37 points. Neither path is therefore dispensable. Second, the drops differ across datasets because the two paths address different failure modes. The reliable path propagates one verified prototype to many cross-view-consistent samples at no additional query cost—at export it covers 29.0% to 45.1% of the samples (Appendix A.10)—while the audit path supplies frozen instance-level references for the ambiguous assignments that prototype supervision deliberately excludes. The combination provides both coverage and targeted instance-level guidance. On UCF50 the two single paths nearly tie (59.51 vs. 59.74), while on Caltech101-20 removing the reliable path alone costs 16.35 points, so the weight of each failure mode is dataset-specific.

Fig. 3(b) replays the same Caltech101-20 evidence three ways: using the anchors only to initialize K-means reaches 47.90 ACC, placing the same labels in one supervised contrastive loss reaches 60.10, and routing them through the two paths reaches 77.54. Initialization alone provides no learning signal after the first assignment, and one labeled contrastive loss applies one target type to a heterogeneous set. The two-path design keeps the reliable and audited targets disjoint and applies each to the samples it was built for; the same evidence therefore works substantially better through the two paths than through either control. These comparisons isolate how the evidence is used; they do not measure MLLM judgment accuracy.

5 Conclusion

We identified semantic-blind correspondence in contrastive MVC: positives and negatives are defined by geometry that cannot see meaning, and the errors persist even when the feature-space correspondence is confident. SECMVC guides those pairs with a frozen criterion and a small MLLM budget: fail-closed anchors with explicit DIFFERENT relations split samples into a reliable set that learns against prototypes without queries and an unreliable set whose double-confirmed matches learn against frozen references. The two paths together achieve the best ACC on all four benchmarks against eighteen baselines, adapt to a changed criterion on the same data by swapping only the criterion card, and spend an MLLM budget that scales with the cluster count rather than the dataset size. Limitations are the assumption of complete views and a hand-supplied criterion card; extending fail-closed anchors to missing views and interactive criterion refinement are promising next steps.

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

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12449–12460, 2020.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yawei Chen, Huibing Wang, Jinjia Peng, and Yang Wang. Anchor learning with potential cluster constraints for multi-view clustering. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 15939–15947, 2025.

Jinrong Cui, Yuting Li, Han Huang, and Jie Wen. Dual contrast-driven deep multi-view clustering. *IEEE Transactions on Image Processing*, 33:4753–4764, 2024.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 886–893, 2005. doi: 10.1109/CVPR.2005.177.

John G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985. doi: 10.1364/JOSAA.2.001160.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, 2001. doi: 10.1023/A:1007612920971.

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 Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 178–178, 2004. doi: 10.1109/CVPR.2004.383.

Yu Geng, Zongbo Han, Changqing Zhang, and Qinghua Hu. Uncertainty-aware multi-view representation learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 7545–7553, 2021.

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 (NeurIPS)*, 2024.

Shizhe Hu, Binyan Tian, Weibo Liu, and Yangdong Ye. Self-supervised trusted contrastive multi-view clustering with uncertainty refined. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025a.

Shizhe Hu, Chengkun Zhang, Guoliang Zou, Zhengzheng Lou, and Yangdong Ye. Deep multi-view clustering by pseudo-label guided contrastive learning and dual correlation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3646–3658, 2025b.

JunLong Ke, Zichen Wen, Yechenhao Yang, Chenhang Cui, Yazhou Ren, Xiaorong Pu, and Lifang He. Integrating vision-language semantic graphs in multi-view clustering. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4273–4281, 2024.

Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27890–27902, 2024.

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 (ECCV)*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Honglin Liu, Chao Sun, Peng Hu, Yunfan Li, and Xi Peng. Conditional representation learning for customized tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025a.

Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, and Yining Li. LLM-guided semantic-aware clustering for topic modeling. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 18420–18435, 2025b.

Suyuan Liu, Ke Liang, Zhibin Dong, Siwei Wang, Xihong Yang, Sihang Zhou, En Zhu, and Xinwang Liu. Learn from view correlation: An anchor enhancement strategy for multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26151–26161, 2024. doi: 10.1109/CVPR52733.2024.02471.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. In *Interspeech*, pp. 814–818, 2019. doi: 10.21437/Interspeech.2019-1705.

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, 1967.

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. doi: 10.1023/A:1011139631724.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. doi: 10.3115/v1/D14-1162.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision (ECCV)*, pp. 143–156, 2010. doi: 10.1007/978-3-642-15561-1_11.

Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013. doi: 10.1007/s00138-012-0450-4.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0.

sentence-transformers. all-mpnet-base-v2 (model card). Hugging Face model card: https://huggingface.co/sentence-transformers/all-mpnet-base-v2, 2021.

Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer. Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6783–6787, 2021. doi: 10.1109/ICASSP39728.2021.9414560.

Zhenqiu Shu, Teng Sun, Yunwei Luo, and Zhengtao Yu. Ambiguous instance-aware contrastive network with multi-level matching for multi-view document clustering. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 20479–20487, 2025.

Peng Su, Shudong Huang, Weihong Ma, Deng Xiong, and Jiancheng Lv. Multi-view granular-ball contrastive clustering. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yuze Tan, Yixi Liu, Shudong Huang, Wentao Feng, and Jiancheng Lv. Sample-level multi-view graph clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23966–23975, 2023. doi: 10.1109/CVPR52729.2023.02295.

Zhiqiang Tao, Hongfu Liu, Sheng Li, Zhengming Ding, and Yun Fu. From ensemble clustering to multi-view clustering. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2843–2849, 2017.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision (ECCV)*, pp. 776–794, 2020.

Daniel J. Trosten, Sigurd Løkse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1255–1265, 2021.

Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3551–3558, 2013. doi: 10.1109/ICCV.2013.441.

Zhenxi Wang, Zongyao Yin, Yujie Hou, and Xianchuan Yu. Robust multi-view clustering via pseudo label guided universum learning. In *ACM International Conference on Multimedia (MM)*, 2025.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, 2023.

Baili Xiao, Zhibin Dong, Ke Liang, Suyuan Liu, Siwei Wang, Tianrui Liu, Xingchen Hu, En Zhu, and Xinwang Liu. EASEMVC: Efficient dual selection mechanism for deep multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20716–20726, 2025.

Jie Xu, Yazhou Ren, Guofeng Li, Lili Pan, Ce Zhu, and Zenglin Xu. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573:279–290, 2021.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S. Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2023b. doi: 10.1109/TKDE.2022.3193569.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. GCFAgg: Global and cross-view feature aggregation for multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19863–19872, 2023.

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. In *ACM International Conference on Multimedia (MM)*, pp. 337–346, 2023. doi: 10.1145/3581783.3611951.

Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, Yonghuai Liu, and Wei Zhang. Learning common semantics via optimal transport for contrastive multi-view clustering. *IEEE Transactions on Image Processing*, 33:4501–4515, 2024.

Wenliang Zhong, Rob Barton, Lucas Goncalves, Kushal Kumar, Feng Jiang, Hehuan Ma, Yuzhi Guo, Vidit Bansal, Karim Bouyarmane, and Junzhou Huang. Universal guideline-driven image clustering via a hybrid LLM agent. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

A Proofs, Details, and Additional Results

A.1 Proof of Theorem 1

Fix a pair $q \in \mathcal{Q}$ and let $r_q = \Pr_A[\widehat{s}_A(q) = 1 \mid \mathcal{G}(\mathcal{X})]$. Since A is criterion-agnostic, $\widehat{\mathbf{s}}_A \perp \kappa \mid \mathcal{G}(\mathcal{X})$, the quantity r_q does not depend on the realized criterion κ. The expected error on q therefore decomposes as

$$\begin{aligned} L_q &= \mathbb{E}_{\kappa,A}[\mathbf{1}[\widehat{s}_A(q) \neq s_\kappa(q)] \mid \mathcal{G}(\mathcal{X})] \\ &= p_q\,(1 - r_q) + (1 - p_q)\,r_q. \end{aligned} \tag{12}$$

The right-hand side is a convex combination of p_q and $1 - p_q$ with weights $1 - r_q$ and r_q. Since $r_q \in [0, 1]$, any such combination lies between p_q and $1 - p_q$, and hence

$$L_q \geq \min\{p_q, 1 - p_q\}. \tag{13}$$

Averaging over all pairs $q \in \mathcal{Q}$ and taking the expectation over $\kappa \sim \mu$ gives

$$R_\mu(A) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} L_q \geq \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \min\{p_q, 1 - p_q\} = \mathcal{A}_\mu(\mathcal{X}). \tag{14}$$

Finally, suppose $\mathcal{A}_\mu(\mathcal{X}) > 0$ and consider a rule A that recovered the desired relations for every criterion in the support of μ. Such a rule would satisfy $\widehat{s}_A(q) = s_\kappa(q)$ for every pair and every realized κ, hence $R_\mu(A) = 0$, contradicting the bound. Therefore no criterion-agnostic supervision rule can recover the desired relations for all criteria simultaneously. □

The quantity $\mathcal{A}_\mu(\mathcal{X})$ measures the average semantic ambiguity of the pairs under the criterion distribution μ; it is positive whenever at least one pair is drawn together under some criteria and apart under others. The bound shows that any supervision built only from criterion-independent information pays this ambiguity as an unavoidable error, which is what motivates SECMVC's criterion-conditioned evidence.

A.2 Datasets and Feature Extraction

The four datasets are Microsoft Common Objects in Context (COCO), Caltech101-20, Fluent Speech Commands, and UCF50. COCO is the 44-category subset of the COCO2017 training images that carry captions: we retain the categories with at least 100 annotated instances, which leaves 44 of the 80 categories and 24,278 samples (103 to 1,795 per category). Caltech101-20 is the standard 20-class subset of Caltech-101 with 2,386 images. Fluent Speech Commands contains 30,043 utterances across 31 intents. UCF50 contains 6,681 clips across 50 human-action categories. COCO uses five views: DINOv2 ViT-B/14 (768-D) and ConvNeXtV2-Base (1024-D) image features (Oquab et al., 2023; Woo et al., 2023); Sentence-BERT all-mpnet-base-v2 text embeddings (768-D) (Reimers & Gurevych, 2019; sentence-transformers, 2021); TF-IDF/SVD text features (256-D); and TF-IDF-weighted GloVe-300 text features (Salton & Buckley, 1988; Deerwester et al., 1990; Pennington et al., 2014). Caltech101-20 uses a 512-D GIST/Gabor-grid view and a 1764-D HOG view (our OpenCV extraction) (Oliva & Torralba, 2001; Daugman, 1985; Dalal & Triggs, 2005). Fluent uses 768-D wav2vec2-base and 1024-D Emformer-RNNT audio features (Baevski et al., 2020; Shi et al., 2021), 256-D TF-IDF/SVD text features, and 50-topic LDA transcript features (Salton & Buckley, 1988; Deerwester et al., 1990; Blei et al., 2003). UCF50 uses two 500-D PCA-reduced Fisher-vector views built from Improved Dense Trajectory HOG and HOF descriptors (Wang & Schmid, 2013; Perronnin et al., 2010). The task criteria are primary-entity identity for COCO and Caltech101-20, complete command intent for Fluent, and the finest observable action for UCF50. All views are row-L2-normalized before the encoder. The MLLM receives raw images, captions, audio, transcripts, or videos, not the extracted clustering features.

A.3 Representation Learning Parameters

Each dataset supplies only its aligned views, sample count N, cluster count K, and a frozen criterion card; the path weights (λ_R, λ_D) are the only scalars that differ across datasets. Per-view MLP encoders and decoders use hidden size 512 and a unit-norm latent of 128, with input row-L2 geometry preserved. The fused embedding is an equal-block concatenation of the unit per-view embeddings followed by L2 normalization. Warm-up trains for 20 epochs on $0.1\,\mathcal{L}_{\text{rec}} + 1.0\,\mathcal{L}_{\text{cross}}$ with no cross-item pairs, and each of the 10 semantic rounds then trains for 20 epochs. All phases use AdamW with learning rate 10^{-4}, weight decay 10^{-5}, and gradient-norm cap 5; the optimizer state is reset before every semantic round. Construction samples up to $n_a = 16$ members per cluster; a candidate is accepted only with at least 3 supports and a support ratio of at least 0.5. The audit budget is $B_{\text{audit}} = 120$ unreliable samples per round with a per-cluster cap of 16. Recorded path weights are $(\lambda_R, \lambda_D) = (0.25, 1.00)$ on COCO and UCF50 and $(0.50, 0.50)$ on Caltech101-20 and Fluent. Round-end clustering is full-N fixed-K spherical K-means initialized with the fused H/D centers under a soft prior of weight 20; the round-10 assignment is exported and scored with ACC, NMI, and ARI.

A.4 Reliability Partition Details

For each $i \in \mathcal{C}$, the prototype margin in view v is $\delta_i^{(v)} = d_{i,(2)}^{(v)} - d_{i,(1)}^{(v)}$, where $d_{i,(1)}^{(v)}$ and $d_{i,(2)}^{(v)}$ are the smallest and second-smallest cosine distances from $z_i^{(v)}$ to the semantic prototypes. Margins are converted into within-view ranks $r_i^{(v)} = \frac{\text{rank}_\uparrow(\delta_i^{(v)})-1}{|\mathcal{C}|-1}$, where $\text{rank}_\uparrow$ ranks margins from the smallest to the largest value, and the reliability score is $\rho_i = \frac{1}{V}\sum_{v=1}^{V} r_i^{(v)}$. Let $\rho_{(1)} \le \cdots \le \rho_{(m)}$ denote the sorted scores in $\mathcal{C}$, where $m = |\mathcal{C}|$. For a candidate split after the t-th score, Otsu's criterion maximizes the between-group variance:

$$t^\star = \arg\max_{1 \le t < m} \frac{t(m-t)}{m^2}\left(\bar{\rho}_{1:t} - \bar{\rho}_{t+1:m}\right)^2, \tag{15}$$

where $\bar{\rho}_{1:t}$ and $\bar{\rho}_{t+1:m}$ are the means of the two score groups. The corresponding threshold is $\theta_{\text{Otsu}} = \left(\rho_{(t^\star)} + \rho_{(t^\star+1)}\right)/2$. This split minimizes the within-group score variance and requires no manually specified confidence threshold or retention ratio.

A.5 MLLM PARAMETERS AND PROMPTS

Model and inference. All semantic queries use Qwen3-Omni (Xu et al., 2025), a multimodal thinking model with native image, audio, and video understanding. The model processes raw modality inputs (images, audio waveforms, video frames, and text transcripts) without requiring clustering feature extractors; the MLLM encoder is independent of the multi-view feature pipeline. Inference uses greedy decoding (temperature 0, top-p 1.0) to ensure deterministic responses. Each query expects structured JSON output, and malformed or ambiguous responses trigger fail-closed behavior: the candidate or assignment is rejected rather than guessed.

Query operations. The method employs six core MLLM operations: (1) OP1: single-sample semantic summary, producing a two-sentence description (core content, observable details) of one anonymous sample from all its available raw modalities in a single request; (2) OP1B: instance fusion, merging aligned per-view summaries into one multi-view description; (3) OP2: local cluster semantics, identifying the largest evidence-supported subset sharing one value under the supplied criterion and returning a concise canonical label; (4) OP5A: high-recall semantic routing, partitioning provisional groups into candidate-comparison blocks (max 30 members per block) organized by inferred semantic families; (5) OP5B: conservative canonical merge, assigning one shared label only when summaries evidence the same category under the criterion; (6) OP5B-REVIEW: candidate discovery, globally identifying nontrivial sets of groups that merit exact semantic equivalence rechecking. Construction uses OP1, OP1B, and OP2 sequentially. Routing and merging use OP5A followed by OP5B. Audits reuse OP1 for the interpretation query and a canonical comparison against the accepted inventory for the matching query; explicit semantic relations are established by OP5B queries over cluster pairs. All operations receive the frozen criterion card and enforce its include/ignore constraints. No operation sees cluster IDs, metadata, filenames, or target group counts.

Prompt design. Every prompt is role-neutral and criterion-driven. The MLLM is instructed to treat the supplied criterion as authoritative and not to infer a dataset, taxonomy, label inventory, or grouping rule. OP2 prompts state: "Find the largest subset whose samples share one value under the criterion's comparison axis and working granularity. Ignore only the variations listed by the criterion and preserve every distinction it lists." OP5B prompts emphasize: "Assign the same label exactly when groups denote the SAME category under the criterion. Spelling variants, singular/plural, and genuine synonyms MUST receive one shared label when summaries evidence the same concrete category." This design prevents the MLLM from inventing labels or forcing all samples into one subset.

Criterion cards. Each run receives a frozen criterion card, shown below verbatim.

COCO: *The class is the primary object or entity identity depicted in the image, at the ordinary object-category level rather than an instance, part, or attribute.* **Include:** animal type, vehicle type, object type, food item. **Ignore:** color, location, background elements, quantity, material, action.
Caltech101-20: *The class is the primary object or entity identity depicted in the image, including living beings, depicted characters, manufactured objects, and graphic symbols, at ordinary category level rather than an instance or a scene.* **Include:** primary object identity, animal/person/character identity, symbol or sign identity, identity-defining visual structure. **Ignore:** background and scene, color/pose/viewpoint, lighting and image quality, incidental clutter around the object.
Fluent Speech Commands: *The complete spoken command intent, including every explicit argument that changes the requested outcome. Presence versus absence of a room, language, or setting is part of that identity.* **Include:** command action, target device or object, destination or room, setting/language or other explicit argument value. **Ignore:** speaker characteristics and voice quality, background noise, politeness and redundant wording, recording conditions, non-command conversational context.
UCF50: *The primary physical action or event identity performed in the video, named at the finest observable action rather than a sport family, person, or scene.* **Include:** observable action or sport-specific movement, equipment interaction required by the action, body motion and temporal sequence. **Ignore:** clothing color, environment type and background scenery, person appearance, camera angle and recording quality.

No card lists class names. The same card accompanies every construction, ownership, relation, and audit request, so the include/ignore constraint shapes cluster descriptions and audit decisions. For the UCF50 dual-criterion experiment (Appendix A.7), only the criterion card, the evaluation labels, and

the cluster count change; features remain identical, and that experiment uses its own activity-domain card.

Candidate routing and merging. Routing uses a block size of 30: OP5A partitions provisional groups into candidate-comparison blocks, each covering one semantic family. Within each block, OP5B performs conservative canonical mapping. The three-round cap limits iterative refinement: each OP5A/OP5B pass is one consolidation round, and up to three rounds are allowed per semantic construction or merge phase. This cap prevents unbounded query cost while allowing sufficient refinement for complex semantic spaces.

Hardware and cost. Experiments run on a single NVIDIA A100 GPU (80GB). The MLLM endpoint handles raw modality inputs.

A.6 MLLM CALL COMPLEXITY

SECMVC's query cost is governed by the cluster count, not by the dataset size. Each semantic round issues three kinds of MLLM requests. Construction draws up to $n_a = 16$ members per cluster: each member receives one summary query and one fusion query, and the cluster adds one cluster-level summary query, for at most $K(2n_a + 1)$ requests. Relation queries compare cluster pairs, at most $K(K-1)/2$ in the worst case (one per pair). Auditing issues one identification and one matching query for each of the $B_{\text{audit}} = 120$ audited samples, for $2B_{\text{audit}}$ requests. A round therefore uses at most

$$Q_{\text{round}}(K) = K(2n_a + 1) + \frac{K(K-1)}{2} + 2B_{\text{audit}} \tag{16}$$

requests. With $T_{\text{rounds}} = 10$ rounds, $n_a = 16$, and $B_{\text{audit}} = 120$, the total bound over a full run is

$$T(K) = T_{\text{rounds}}\, Q_{\text{round}}(K) = 10\left[K(2n_a + 1) + \frac{K(K-1)}{2} + 2B_{\text{audit}}\right] = 5K^2 + 325K + 2400. \tag{17}$$

The sample count N does not enter the bound: n_a and B_{audit} are fixed hyperparameters, and the reliable set learns from prototypes without any query. Eq. (17) is therefore an upper bound: a cluster draws at most n_a members, and N only imposes feasibility in the sense that the dataset must contain enough samples overall to support the per-cluster caps and the audit budget. All four recorded runs satisfy these constraints.

Table 2 compares the bound with the measured real-API calls of the ten recorded rounds.

Table 2: MLLM request counts over the 10 recorded rounds ($n_a = 16$, $B_{\text{audit}} = 120$). The bound is Eq. (17); remote requests are real API calls.

Dataset	N	K	Bound $T(K)$	Remote	Remote / bound
COCO	24,278	44	26,380	9,494	36.0%
Caltech101-20	2,386	20	10,900	2,481	22.8%
Fluent	30,043	31	17,280	4,862	28.1%
UCF50	6,681	50	31,150	12,370	39.7%

The measured totals follow the K-scaling of the bound—Caltech101-20 at $K = 20$ is the cheapest run and UCF50 at $K = 50$ the most expensive—and stay at 22.8% to 39.7% of the worst case, for two reasons. Relation queries are merged into batches, costing about $2K$ calls per round in the recorded runs instead of $K(K-1)/2$, and the audit matching query is only issued for samples whose interpretation is judged uncertain, so not every audited sample costs two requests. Dataset size plays no visible role: Fluent, the largest dataset with $N = 30{,}043$, costs about half of UCF50's remote calls at $N = 6{,}681$. Real-API consumption averages 949, 248, 486, and 1,237 requests per round on COCO, Caltech101-20, Fluent, and UCF50, respectively. The query cost of SECMVC therefore scales with the cluster count and the fixed semantic budget, not with the number of samples.

A.7 RECONFIGURABLE SEMANTIC CRITERION (Q3)

To verify that SECMVC resolves geometry-criterion misalignment rather than simply improving feature quality, we evaluate on UCF50 under two different semantic criteria using the same pretrained

features. The default criterion groups videos by fine-grained physical action (*e.g.*, basketball shooting vs. basketball dribbling). The alternative criterion groups by coarse activity domain (*e.g.*, all basketball actions together, all diving actions together). The pretrained input features and encoder initialization stay the same; the criterion card, the target partition, and the cluster count change.

Table 3 reports results. SECMVC adapts to both criteria by querying the MLLM with the corresponding card, achieving the best ACC, NMI, and ARI under both partitions. The input features and encoder initialization are unchanged; the cluster count follows the target partition ($K = 50$ versus $K = 10$), and only the criterion card and the evaluation partition change. This controlled change shows that the semantic signal is reconfigurable, rather than being a fixed improvement to the feature extractor.

Table 3: Reconfigurable criterion on UCF50 (%). Same input features and encoder initialization; the cluster count follows each criterion ($K = 50$ vs. $K = 10$). SECMVC adapts by changing the criterion card; baselines do not receive the card. Best per criterion in bold.

Method	Fine action ($K = 50$)			Activity domain ($K = 10$)		
	ACC	NMI	ARI	ACC	NMI	ARI
SDMVC	27.26	42.04	14.37	30.43	19.58	10.49
DEMVC	28.36	47.38	17.63	33.83	23.40	12.95
ROLL	54.29	67.09	40.94	46.62	39.23	24.24
DCMVC	36.55	47.90	22.81	30.04	19.08	12.04
STCMC-UR	24.34	44.93	10.86	42.06	33.94	23.74
ALPC	46.46	57.83	30.36	19.28	3.26	0.43
SECMVC (ours)	**62.88**	**68.57**	**48.80**	**79.34**	**61.95**	**62.11**

The ACC gaps over the strongest baseline on the two criteria are 8.59 and 32.72 points, and after the activity-domain coarsening the feature-based methods shown stay within a 19.28–46.62 ACC band while SECMVC reaches 79.34. This pattern is consistent with the intended role of SECMVC: criterion-conditioned evidence provides semantic targets for relationships that are coherent in feature space but misaligned with the task axis, and those targets follow the supplied criterion.

A.8 MLLM Budget Efficiency (Q4)

On COCO we vary the audit budget over $\{40, 80, 120, 160\}$ using the recorded path weights. Smaller budgets keep prefixes of the recorded audit ranking, while 160 adds the next forty candidates.

Figure 4 shows diminishing returns in the COCO replay. ACC rises by 0.35 points between 40 and 80 audited samples, then by only 0.05 points between 80 and 120. At 160 audited samples, ACC is 0.13 points below the recorded 120-audit result; NMI and ARI show the same non-monotonic pattern.

The shape of the curve follows from how the audited samples are chosen. Each round ranks candidates by neighborhood centrality and spends the budget from the top of that ranking, so the earliest queries always target the most representative assignments. Once the head of the ranking is exhausted, the remaining candidates carry little additional correctable evidence, and the curve flattens by construction. The 0.13-point dip at 160 reflects the same mechanism: the extra forty lowest-ranked candidates raise the number of confirmed matches from 40 to 56, but their assignments reuse reference sets that already exist for their semantics, so the added supervision brings no gain and the difference sits within round-to-round fluctuation. Thus 120 is a measured operating point on this nested sweep, not a claim that additional queries always improve clustering.

A.9 Parameter Sensitivity (Q5)

The reliability partition uses the parameter-free Otsu split. In this experiment only the two semantic path weights vary; all other settings stay fixed. Figure 5 shows their joint effect on COCO and UCF50: the center cell of each grid is the recorded live run, and the surrounding eight cells replay the same frozen semantic snapshot without new MLLM calls.

The results show that SECMVC needs no delicate tuning. On COCO the plane is essentially flat: last-round ACC spans only 0.11 points (87.14–87.25) and the best round-level ACC of each run only

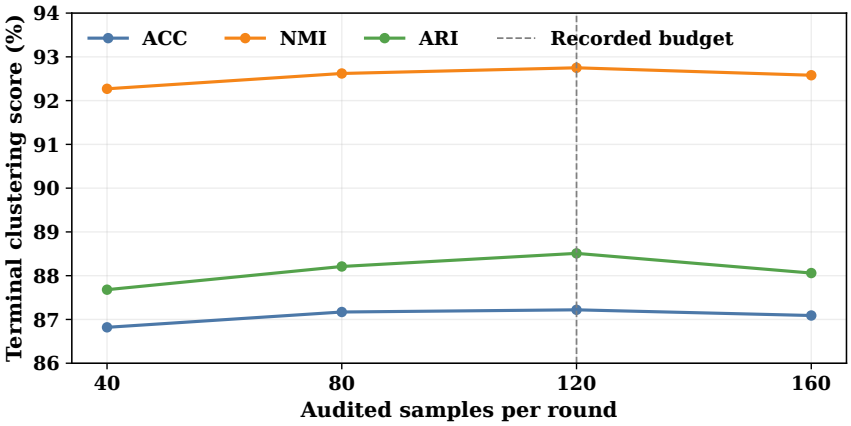


Figure 4: Q4 audit-budget sensitivity on COCO. Terminal ACC, NMI, and ARI use the nested audit rankings recorded in the experiment ledger.

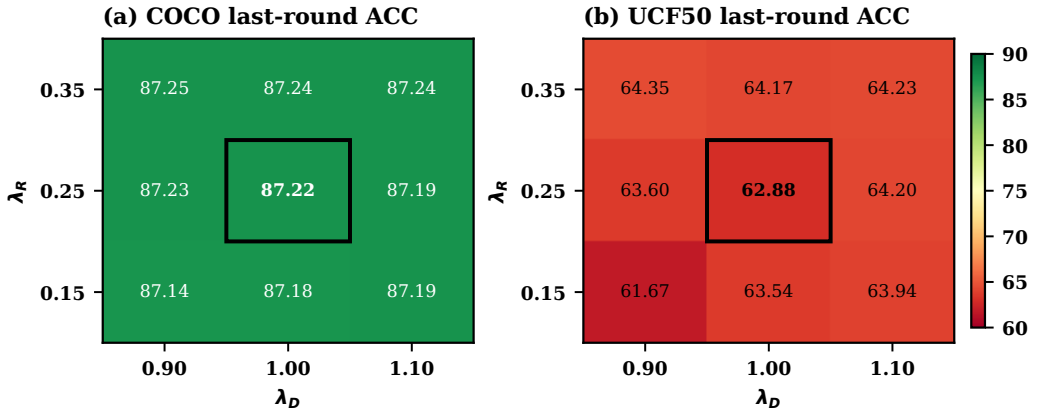


Figure 5: Joint (λ_R, λ_D) sensitivity of last-round ACC. Both panels share one color scale, so the same color means the same ACC level; the boxed center cell is the recorded live run, and the surrounding cells reuse its frozen semantic evidence.

0.34 points (87.16–87.50) across the entire grid, with the recorded pair in the middle of the plateau. On UCF50 last-round ACC rises gently with λ_R and shows no collapse anywhere: the grid spans 61.67–64.35, with the best round-level ACC of each run spanning 63.31–65.17; the recorded live pair $(0.25, 1.00)$ at 62.88 lies within 1.5 points of the best cell $(0.35, 0.90)$ at 64.35, and the small offset between the live center and its own ring neighbors reflects the live-versus-replay protocol difference. The two recorded weight pairs therefore require no further dataset-specific tuning.

A.10 Semantic-State Coverage and Convergence

Table 4 reports semantic-state counts at export for all four datasets. K_s is the number of accepted semantics after fail-closed construction and deduplication. H contains construction supports, R the reliable samples, U the unreliable samples, and D the double-confirmed audit matches. The reliable set covers 29.0% to 45.1% of each dataset, so this part of the prototype supervision costs no query beyond construction.

On four datasets, both paths are populated, so semantic evidence flows through both propagation and targeted guidance. When an audit yields no confirmed match, the method degrades to prototype supervision without injecting a guessed assignment.

Table 4: Semantic-state counts at export (round 10) with audit budget 120. $H + R + U = N$; $D \subseteq U$.

Dataset	K_s	H	R	U	D	R/N (%)
COCO	41	620	10,960	12,698	40	45.1
Caltech101-20	16	256	976	1,154	95	40.9
Fluent	19	295	8,701	21,047	9	29.0
UCF50	43	614	2,215	3,852	24	33.2

Within each round, the semantic targets are frozen, so the inner optimization does not rewrite the anchors. Across rounds, every semantic state is rebuilt from a fresh snapshot, and the exported value is the last round.

Figure 6 traces ACC and NMI over the ten rounds on Caltech101-20 and UCF50, the two datasets with the clearest progressions. Caltech101-20 gains most of its ACC after round 7, while UCF50 rises overall with small mid-run dips; on both datasets the two metrics move together, indicating that the semantic guidance improves the partition as a whole rather than a single metric. The round-to-round variation reflects the fresh semantic-state rebuild at each round, not a loss of supervision, and the last-round export is what we report.

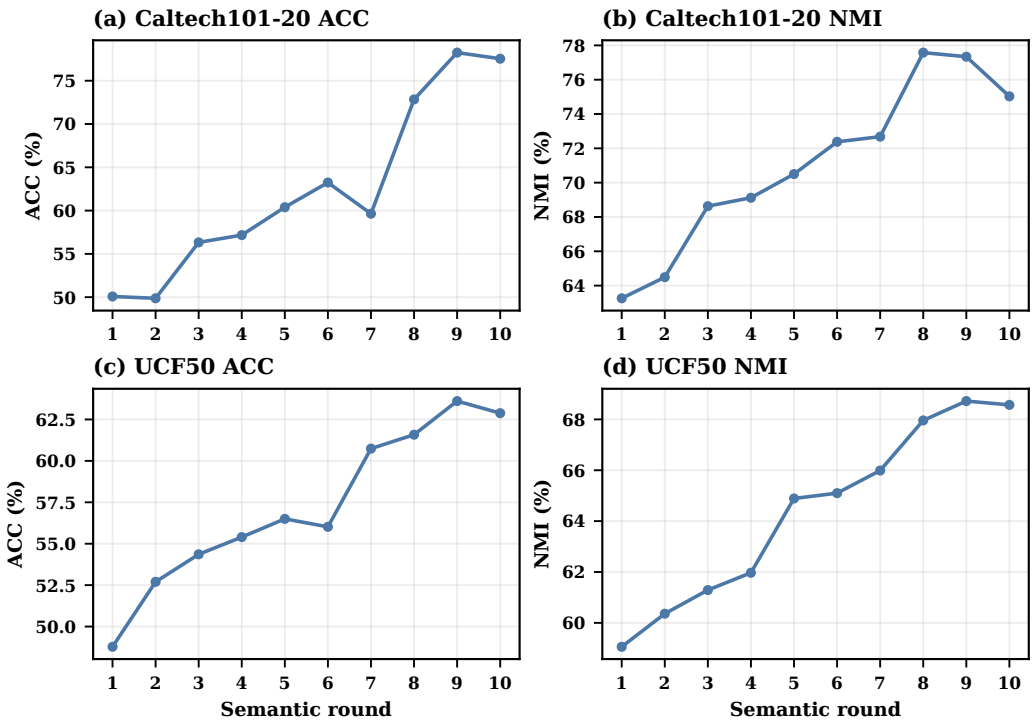


Figure 6: Round-wise convergence on Caltech101-20 and UCF50. Each panel traces one metric across the ten semantic rounds; the exported value is the last round.

A.11 ROUND PSEUDOCODE

1	**Warm-up** (once). 20 epochs of reconstruction and same-item cross-view alignment. The temporary assignment gives every item its own ID.
2	**Provisional partition**. Fixed-K spherical K-means on fused $\bar{z}_i$ yields $\{c_i\}$.
3	**Construction**. Per cluster, select up to n_a members by cosine similarity to the fused center. The MLLM summarizes their raw modalities and, under the criterion κ, returns a specific shared meaning with supports S_k, or abstains. Accept only if the description is nonempty and specific, $\lvert S_k\rvert \geq n_{\min}$, and $\pi_k \geq \pi_{\min}$.
4	**Semantic identity**. Process candidates in descending $\lvert C_k\rvert\pi_k$ and compare them with accepted meanings. New candidates receive a semantic ID; uniquely matched duplicates contribute their nonconflicting supports to the existing meaning; unresolved matches are discarded. $K_s \leq K$.
5	**Relations**. Query every active-semantic pair as DIFFERENT/UNKNOWN, then freeze the symmetric table. Only DIFFERENT creates negative edges.
6	**Freeze**. Prototypes $p_u^{(v)}$ and stop-gradient reference bank $b_j^{(v)}$ from H.
7	**Reliability partition**. Per-view nearest prototypes $q_i^{(v)}$ and the consensus score g_i (Eq. (5)) form $\mathcal{C}$. Within-view margin ranks give ρ_i, and the Otsu threshold θ_{Otsu} splits $\mathcal{C}$ into the reliable set R and the unreliable set U.
8	**Audit**. Within each cluster, rank previously unaudited U samples by local centrality and select the top a_k per cluster, spending the global budget B_{audit}. Each selected sample receives at most two MLLM requests: interpret under κ, then match to an accepted semantic only when the interpretation is uncertain. Verified matches form D; the rest receive no semantic target.
9	**Train**. 20 epochs on $\mathcal{L} = \lambda_{\mathrm{cross}}\mathcal{L}_{\mathrm{cross}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_R\mathcal{L}_R + \lambda_D\mathcal{L}_D$ with separate samplers per path.
10	**Round end**. Recompute z_i and $\bar{z}_i$, run full-N fixed-K spherical K-means, persist all artifacts. The next round rebuilds semantic state from a fresh snapshot.
11	**Final export**. Round 10 clustering is exported. Evaluate ACC/NMI/ARI offline.