SEMANTIC CALIBRATION BEYOND ANCHOR CORRESPONDENCE FOR MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering aims to uncover a shared partition from heterogeneous representations of the same underlying data. Most existing methods emphasize cross-view consistency, implicitly assuming that a global category exhibits compatible local semantics across views. In practice, however, differences in modalities, feature representations, anchor construction, missing or degraded observations, and other view-specific factors can induce class-level semantic shifts, making this assumption overly restrictive. We propose *Calibrated Anchor Semantic Transitions* (CAST), a probabilistic framework that explicitly models how global cluster semantics are expressed through view-local semantic states. CAST operates directly on heterogeneous sample-to-anchor assignments, allows different anchor cardinalities across views, and avoids requiring cross-view anchor correspondence. Learned class-level transitions use cross-view evidence to infer global semantics while identifying and calibrating systematic sample–view shifts. We further establish theoretical properties concerning correspondence-free semantic evidence, semantic discriminability, multi-view complementarity, and the convergence behavior of the proposed MAP-EM optimization. Extensive experiments on multiple public multi-view clustering benchmarks, together with controlled semantic-shift evaluations, demonstrate strong clustering performance across both naturally heterogeneous and deliberately shifted settings. Ablation studies on the semantic transition structure further verify that explicitly modeling class-level semantic transitions is important for capturing cross-view semantic heterogeneity.

1 INTRODUCTION

Multi-View Clustering (MVC) seeks a common partition from several representations of the same underlying population. A standard design principle is to extract the common structure across views while suppressing representation-specific variation. Consequently, existing methods often learn a consensus graph, common embedding, shared cluster indicator, or mutually regularized view-specific representations (Kang et al., 2020; Zhang et al., 2024; Chen et al., 2025a; Zhang et al., 2025). Anchor-based MVC makes this paradigm scalable by replacing dense sample–sample relations with sample-to-anchor representations. This compression is efficient, but it raises a more basic question: *what should actually be made consistent across heterogeneous views?*

A common implicit answer is that a global class has the same local semantic organization in every view. Let $Y_i \in [C]$ denote the global semantic identity of sample i, and let $L_i^{(v)} \in [C]$ denote the class-level local semantic state induced by view v. Identity coupling corresponds to $\Pr(L_i^{(v)} = k \mid Y_i = c) = \mathbb{I}[k = c]$. This is stronger than global semantic inference requires: a sample can have a coherent global identity while a particular view provides evidence for another class-level semantic state. We call this discrepancy a *view-specific semantic shift*. The term "shift" describes a relation between global and local class-level semantics; it does not imply that the true class label changes, nor that one view is a corrupted version of another.

Why can semantic shifts arise? Several mechanisms (See Appendix A for details) can alter the class-level semantic organization of a view before or during anchor representation. Different modalities expose different attributes of the same objects, while heterogeneous feature extractors can induce substantially different neighborhood geometries even within a common modality. Missing or degraded observations, imperfect observation association, and measurement variation or noise may further

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.

perturb such local structures. These effects act primarily on the view-specific feature representations $\mathbf{X}^{(v)}$. Anchor construction subsequently maps each $\mathbf{X}^{(v)}$ into a sample-to-anchor representation and introduces an additional source of heterogeneity through view-dependent anchor locations, selection procedures, and cardinalities. These factors can coexist and accumulate; CAST therefore focuses on their common statistical consequence—a discrepancy between shared global semantics and their view-local realizations—rather than on any individual physical cause.

This distinction remains important after the heterogeneous features are encoded by anchor assignments. For view v, let $\mathbf{Z}^{(v)} \in \mathbb{R}_{+}^{N\times m_v}$ denote a row-stochastic *sample-to-anchor probability assignment matrix* constructed from $\mathbf{X}^{(v)}$. The resulting anchor simplex inherits the semantic organization of its underlying view while also introducing view-specific anchor coordinates and potentially different cardinalities. Thus semantic shifts expressed in $\mathbf{Z}^{(v)}$ may reflect both upstream feature-level heterogeneity and additional anchor-representation heterogeneity. Cross-view anchor correspondence is therefore neither assumed nor the target of CAST: anchor identities describe the internal coordinates of individual representations, whereas the clustering objective concerns the class-level semantic information conveyed by those representations.

We therefore couple views at the semantic level. CAST introduces a categorical channel $T_{ck}^{(v)} = \Pr(L_i^{(v)} = k \mid Y_i = c)$ from a shared global class to a view-local class-level semantic state. A diagonal event denotes semantic preservation under a fixed component-label convention; off-diagonal mass denotes directional semantic shift. A local observation that appears inconsistent under identity coupling can thus remain informative when the shift is systematic. Using the same cardinality C gives a common class-semantic vocabulary in which $c \to k \neq c$ directly represents a view-specific shift; CAST calibrates these transitions rather than introducing arbitrary view-specific subclusters. Hence $\mathbf{T}^{(v)} \in \mathbb{R}^{C\times C}$ has a direct class-to-class interpretation.

The local observations are anchor assignments rather than raw features. We map each probability row to the positive orthant of the unit sphere via an elementwise square root and use a von Mises–Fisher (vMF) directional likelihood to model local semantic components (Banerjee et al., 2005). The transition matrix $\mathbf{T}^{(v)}$ is parameterized by preservation and conditional shift directions, while weak fixed Beta–Dirichlet priors stabilize low-count estimates. Exact latent-state posteriors yield a compact Maximum A Posteriori Expectation-Maximization (MAP-EM) algorithm with closed-form blocks and easily solvable finite scalar concentrations. Overall, we make four contributions:

- We formulate view-specific class-level semantic shift for heterogeneous multi-view anchor representations, replacing conventional identity coupling between global and view-local semantics with an explicit probabilistic global-to-local semantic relation.
- We introduce CAST, a semantic calibration framework operating on independent sample-to-anchor probability representations. CAST accommodates heterogeneous view-specific anchor cardinalities and learns class-level semantic transitions without requiring cross-view anchor correspondence.
- We derive exact latent-state inference and an efficient MAP-EM optimization procedure, and establish theoretical properties concerning correspondence-free semantic evidence, semantic discriminability and informativeness, multi-view distinguishability, and monotone convergence of the bounded MAP objective.
- We conduct extensive experiments on multiple public multi-view clustering benchmarks and controlled semantic-shift settings. Comparisons with representative methods, together with transition-structure ablations under both natural and deliberately controlled semantic heterogeneity, demonstrate strong clustering performance and support the motivation for explicitly modeling class-level semantic transitions.

2 Related work and positioning

Anchor-based multi-view clustering. Anchor representations reduce the quadratic cost of sample graphs by encoding sample-to-prototype relations. LMVSC uses view-specific reduced graphs for scalable clustering (Kang et al., 2020); CAMVC learns cluster-wise anchors (Zhang et al., 2024); ALPC introduces potential cluster constraints and a shared latent semantic module (Chen et al.,

2025a); MAGIC studies discriminative anchor design and guided multi-view representations (Zhang et al., 2025). Robust/consistent anchor methods such as RCAGL explicitly separate common and view-specific structure (Liu et al., 2024b), while AGLDR jointly learns anchors under prescribed feature-noise models (Chen et al., 2025b). CAST is complementary to these anchor-learning objectives: it receives independently constructed sample-to-anchor assignments and models the semantics induced by them rather than learning a common anchor coordinate system.

Probabilistic consensus clustering. Probabilistic formulations have been used to combine multiple sources or views into a shared clustering solution. For example, Bayesian consensus clustering couples source-specific assignments through a common latent partition (Lock & Dunson, 2013). Such methods remain centered on estimating a consensus partition from multiple sources, rather than modeling systematic class-level semantic shifts induced by heterogeneous representations. CAST addresses a different modeling problem: it introduces an explicit global-to-local semantic transition layer over independently constructed anchor simplices, so that view-local semantic evidence can depart systematically from a sample's global semantic state and be calibrated explicitly. Class-conditional transition matrices also appear in observer- or annotation-error models, but there they describe discrete reporting errors rather than latent view-local semantic realizations of heterogeneous representations (Dawid & Skene, 1979). Standard vMF mixtures provide the directional distributional foundation for our local observation model (Banerjee et al., 2005). Thus, CAST's central contribution is the class-level semantic calibration of heterogeneous anchor representations, together with inference and analysis tailored to this setting.

3 Calibrated Anchor Semantic Transitions

3.1 Problem setting

We use bold uppercase and lowercase symbols for matrices and vectors, respectively; plain symbols denote scalars or scalar entries. A parenthesized superscript (v) marks a view-specific quantity, while sample, class, and local-state indices appear as subscripts. View-dependent dimensions such as the anchor count retain the conventional subscript notation m_v.

For view $v \in [V]$, we observe a nonnegative row-stochastic sample-to-anchor probability assignment matrix $\mathbf{Z}^{(v)} \in \mathbb{R}_+^{N \times m_v}$ satisfying $\mathbf{Z}^{(v)}\mathbf{1} = \mathbf{1}$. The transposed ith row $\mathbf{z}_i^{(v)} \in \mathbb{R}^{m_v}$ lies on the simplex Δ^{m_v-1}. Anchor counts m_v may differ and no relation between columns of $\mathbf{Z}^{(u)}$ and $\mathbf{Z}^{(v)}$ is assumed. CAST does not prescribe how anchors are constructed; they may come from k-means prototypes, sampling, learned anchors, or another fixed upstream procedure.

Let $Y_i \in [C]$ denote the global semantic class. For each view, let $L_i^{(v)} \in [C]$ denote a view-local *class-level* semantic state. A shared C-state vocabulary makes a view-specific shift $c \to k \neq c$ directly representable. After fixing a component-label convention, $L_i^{(v)} = Y_i$ denotes semantic preservation and $L_i^{(v)} \neq Y_i$ a semantic shift. CAST jointly infers the global assignments Y_i and view-local states $L_i^{(v)}$, yielding sample–view semantic-shift posteriors $\tau_i^{(v)}$ associated with the event $L_i^{(v)} \neq Y_i$, while learning view-specific class-level transition matrices $\mathbf{T}^{(v)}$ that characterize the systematic global-to-local semantic relations used for calibration. Final labels follow from $\hat{y}_i = \arg\max_c q_{ic}$. Their formal definitions and estimation are given in Sections 3.3 and 3.5.

3.2 Spherical anchor evidence and local semantic states

The square-root map sends each anchor-assignment probability vector to

$$\mathbf{u}_i^{(v)} = \sqrt{\mathbf{z}_i^{(v)}} \in \mathbb{S}_+^{m_v-1} \subset \mathbb{S}^{m_v-1}, \qquad \|\mathbf{u}_i^{(v)}\|_2 = 1, \tag{1}$$

where the square root is elementwise. Because $\mathbf{z}_i^{(v)}$ lies on the probability simplex, this map preserves Hellinger geometry: for $\mathbf{z}, \mathbf{p} \in \Delta^{m_v-1}$, $\mathrm{H}^2(\mathbf{z}, \mathbf{p}) = \frac{1}{2}\|\sqrt{\mathbf{z}} - \sqrt{\mathbf{p}}\|_2^2 = 1 - (\sqrt{\mathbf{z}})^\mathsf{T}\sqrt{\mathbf{p}}$. Hence Euclidean distance and directional inner-product similarity on the sphere directly encode distributional similarity between the original anchor assignments. This gives the vMF model below a natural geometric basis. Each view keeps its own sphere dimension and anchor coordinate system.

Given $L_i^{(v)} = k$, we model the embedded observation with a standard von Mises–Fisher (vMF) directional working density, where $\|\boldsymbol{\mu}_k^{(v)}\|_2 = 1$:

$$f_k^{(v)}(\mathbf{u}) = \mathcal{C}_{m_v}(\kappa^{(v)}) \exp\Big(\kappa^{(v)} (\boldsymbol{\mu}_k^{(v)})^\top \mathbf{u}\Big), \qquad \mathcal{C}_D(\kappa) = \frac{\kappa^{D/2-1}}{(2\pi)^{D/2} I_{D/2-1}(\kappa)}. \tag{2}$$

Here $\mathcal{C}_D(\kappa)$ is the vMF normalizing constant in ambient dimension D, and $I_\nu(\kappa)$ denotes the modified Bessel function of the first kind of order ν. To prevent concentration degeneracy, Eq. (17) caps each view's attainable mean resultant length at $1 - \epsilon_\kappa$ with fixed $\epsilon_\kappa = 10^{-3}$, inducing a finite dimension-dependent bound on $\kappa^{(v)}$. The observations lie in the positive orthant, while standard vMF is normalized on the entire sphere. We therefore use it as an *ambient directional working likelihood*: the purpose is to exploit the directional geometry induced by anchor probabilities rather than to claim an exactly normalized density on the simplex.

3.3 Class-level semantic transition calibration

Instead of enforcing identity coupling, CAST learns a row-stochastic transition matrix $\mathbf{T}^{(v)} = [T_{ck}^{(v)}] \in \mathbb{R}_+^{C \times C}$ with

$$T_{ck}^{(v)} = \Pr(L_i^{(v)} = k \mid Y_i = c), \qquad \mathbf{T}^{(v)}\mathbf{1} = \mathbf{1}. \tag{3}$$

We parameterize each row as

$$T_{ck}^{(v)} = (1 - \rho_c^{(v)})\mathbb{I}[k = c] + \rho_c^{(v)} S_{ck}^{(v)}, \tag{4}$$

where $S_{cc}^{(v)} = 0$, $S_{ck}^{(v)} \geq 0$, and $\sum_{k \neq c} S_{ck}^{(v)} = 1$. Thus $\rho_c^{(v)}$ controls whether global class c departs from the matched local semantic state in view v, while $S_{ck}^{(v)}$ describes the destination probability conditional on that shift. This is a reparameterization of a row-stochastic transition matrix rather than a lower-dimensional model; its value lies in semantic interpretability and structured regularization.

The calibrated view-conditional density of global class c is

$$\psi_c^{(v)}(\mathbf{u}) = \sum_{k=1}^{C} T_{ck}^{(v)} f_k^{(v)}(\mathbf{u}) = (1 - \rho_c^{(v)}) f_c^{(v)}(\mathbf{u}) + \rho_c^{(v)} \sum_{k \neq c} S_{ck}^{(v)} f_k^{(v)}(\mathbf{u}). \tag{5}$$

A local state that would look inconsistent under identity coupling can therefore remain informative for global class c when the transition is systematic.

3.4 Generative model and MAP objective

For each sample i and view v, the hierarchical model is

$$Y_i \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad L_i^{(v)} \mid Y_i = c \sim \mathrm{Cat}(\mathbf{T}_{c:}^{(v)}), \quad \mathbf{u}_i^{(v)} \mid L_i^{(v)} = k \sim f_k^{(v)}. \tag{6}$$

Let $\mathbf{U}^{(v)} \in \mathbb{R}^{N \times m_v}$ collect the embedded observations in view v, with its ith row equal to $(\mathbf{u}_i^{(v)})^\top$. Define $\mathcal{U} := \{\mathbf{U}^{(v)}\}_{v=1}^V$ and $\mathcal{U}_i := \{\mathbf{u}_i^{(v)}\}_{v=1}^V$. We write $\mathbf{Y} := (Y_1, \ldots, Y_N)^\top \in [C]^N$ for the global latent assignments and $\mathbf{L}^{(v)} := (L_1^{(v)}, \ldots, L_N^{(v)})^\top \in [C]^N$ for the view-v local semantic states, with $\mathbf{L} := [\mathbf{L}^{(1)}, \ldots, \mathbf{L}^{(V)}] \in [C]^{N \times V}$. Let $\boldsymbol{\Theta}$ denote the parameter collection comprising $\boldsymbol{\pi}$, $\rho_c^{(v)}$, $\mathbf{S}^{(v)}$, $\boldsymbol{\mu}_k^{(v)}$, and $\kappa^{(v)}$. The complete-data joint mass/density is

$$p(\mathbf{Y}, \mathbf{L}, \mathcal{U} \mid \boldsymbol{\Theta}) = \prod_{i=1}^{N} \pi_{Y_i} \prod_{v=1}^{V} T_{Y_i L_i^{(v)}}^{(v)} f_{L_i^{(v)}}^{(v)}(\mathbf{u}_i^{(v)}). \tag{7}$$

Marginalizing the view-local states $\mathbf{L}$ and using Eq. (5) gives

$$p(\mathbf{Y}, \mathcal{U} \mid \boldsymbol{\Theta}) = \prod_{i=1}^{N} \pi_{Y_i} \prod_{v=1}^{V} \psi_{Y_i}^{(v)}(\mathbf{u}_i^{(v)}). \tag{8}$$

Further marginalizing the global assignments $\mathbf{Y}$ yields the observed-data log likelihood

$$\mathcal{L}(\mathbf{\Theta}) := \log p(\mathcal{U} \mid \mathbf{\Theta}) = \sum_{i=1}^{N} \log \left[\sum_{c=1}^{C} \pi_c \prod_{v=1}^{V} \psi_c^{(v)}(\mathbf{u}_i^{(v)}) \right]. \tag{9}$$

MAP-EM subsequently infers the posteriors of $\mathbf{Y}$ and $\mathbf{L}$. We use weak fixed priors

$$\rho_c^{(v)} \sim \mathrm{Beta}(a_\rho, b_\rho), \qquad \mathbf{S}_{-c}^{(v)} \sim \mathrm{Dir}\Big(\mathbf{1} + \gamma_S \boldsymbol{\omega}_{-c}^{(v)}\Big). \tag{10}$$

For a fixed source class c, the subscript $-c$ collects all destination classes except c. Explicitly,

$$\mathbf{S}_{-c}^{(v)} := \big(S_{c1}^{(v)}, \ldots, S_{c,c-1}^{(v)}, S_{c,c+1}^{(v)}, \ldots, S_{cC}^{(v)}\big)^\top, \quad \boldsymbol{\omega}_{-c}^{(v)} := \big(\omega_{c1}^{(v)}, \ldots, \omega_{c,c-1}^{(v)}, \omega_{c,c+1}^{(v)}, \ldots, \omega_{cC}^{(v)}\big)^\top,$$

with $\mathbf{S}_{-c}^{(v)}, \boldsymbol{\omega}_{-c}^{(v)} \in \Delta^{C-2}$. Thus $\mathbf{S}_{-c}^{(v)}$ is the conditional off-diagonal destination distribution and $\boldsymbol{\omega}_{-c}^{(v)}$ is its fixed prior profile. Defaults are $a_\rho = 1.1$, $b_\rho = 2$, and $\gamma_S = 2$. The weak Beta prior mildly favors preservation and avoids unstable endpoint MAP estimates; the Dirichlet prior contributes γ_S pseudo-counts, stabilizing low-count destination rows toward $\boldsymbol{\omega}_{-c}^{(v)}$ and vanishing in relative influence as counts grow. The profiles are initialized once and then fixed.

Up to constants, the MAP objective is

$$\mathcal{J}(\mathbf{\Theta}) = \mathcal{L}(\mathbf{\Theta}) + \sum_{v,c} \Big[(a_\rho - 1) \log \rho_c^{(v)} + (b_\rho - 1) \log(1 - \rho_c^{(v)}) \Big] + \sum_{v,c,k \neq c} \gamma_S \omega_{ck}^{(v)} \log S_{ck}^{(v)}. \tag{11}$$

3.5 MAP-EM INFERENCE

For compactness, let $f_{ik}^{(v)} := f_k^{(v)}(\mathbf{u}_i^{(v)})$ and $\psi_{ic}^{(v)} := \psi_c^{(v)}(\mathbf{u}_i^{(v)})$; i, c, k, v index sample, global class, local semantic state, and view. Conditioned on Y_i, the view-local branches factorize. Applying Bayes' rule gives the exact E-step posteriors below; Appendix B.1 provides the derivation.

E-step. The global semantic posterior is

$$q_{ic} := \Pr(Y_i = c \mid \mathcal{U}_i, \mathbf{\Theta}) = \frac{\pi_c \prod_v \psi_{ic}^{(v)}}{\sum_{c'=1}^{C} \pi_{c'} \prod_v \psi_{ic'}^{(v)}}. \tag{12}$$

The joint global/local responsibility in view v is

$$\xi_{ick}^{(v)} := \Pr(Y_i = c, L_i^{(v)} = k \mid \mathcal{U}_i, \mathbf{\Theta}) = q_{ic} \frac{T_{ck}^{(v)} f_{ik}^{(v)}}{\psi_{ic}^{(v)}}. \tag{13}$$

Hence $\sum_k \xi_{ick}^{(v)} = q_{ic}$. Marginalizing the global class gives $r_{ik}^{(v)} := \sum_c \xi_{ick}^{(v)} = \Pr(L_i^{(v)} = k \mid \mathcal{U}_i, \mathbf{\Theta})$, which we call the *marginal local-state posterior*. The sample–view semantic-shift posterior is $\tau_i^{(v)} := \Pr(L_i^{(v)} \neq Y_i \mid \mathcal{U}_i, \mathbf{\Theta}) = 1 - \sum_c \xi_{icc}^{(v)}$; the off-diagonal $\xi_{ick}^{(v)}$ further resolves its directional destination. Appendix B.1 distinguishes aggregate sample–view shift/preservation posteriors from the class/view-level transition parameters $\rho_c^{(v)}$, which need not have the same average. Define the aggregated posterior counts

$$n_c = \sum_i q_{ic}, \qquad n_{ck}^{(v)} = \sum_i \xi_{ick}^{(v)}, \qquad n_{-c}^{(v)} = \sum_{k \neq c} n_{ck}^{(v)}. \tag{14}$$

For fixed source class c, $n_{-c}^{(v)}$ is its total off-diagonal posterior mass in view v.

MAP M-step. Maximizing the MAP auxiliary objective blockwise gives Eqs. (15)–(17); Appendix B.2–B.7 provides the derivations. In particular,

$$\pi_c = n_c / N, \qquad \rho_c^{(v)} = \frac{n_{-c}^{(v)} + a_\rho - 1}{n_c + a_\rho + b_\rho - 2}, \qquad S_{ck}^{(v)} = \frac{n_{ck}^{(v)} + \gamma_S \omega_{ck}^{(v)}}{n_{-c}^{(v)} + \gamma_S}, \qquad k \neq c. \tag{15}$$

with $S_{cc}^{(v)} = 0$. For the emissions, set $\mathbf{s}_k^{(v)} = \sum_i r_{ik}^{(v)} \mathbf{u}_i^{(v)}$. Then

$$\boldsymbol{\mu}_k^{(v)} = \frac{\mathbf{s}_k^{(v)}}{\|\mathbf{s}_k^{(v)}\|_2}, \qquad \bar{R}^{(v)} = \frac{1}{N} \sum_k \|\mathbf{s}_k^{(v)}\|_2, \tag{16}$$

Define the mean-resultant function $A_D(\kappa) := I_{D/2}(\kappa)/I_{D/2-1}(\kappa)$ and the capped target $\widetilde{R}^{(v)} := \min\{\bar{R}^{(v)}, 1-\epsilon_\kappa\}$. Then $\kappa^{(v)}$ is the unique finite solution of

$$A_{m_v}(\kappa^{(v)}) = \widetilde{R}^{(v)}. \tag{17}$$

Thus all probability and direction blocks have closed-form MAP updates, and each view requires one scalar concentration solve. In addition, Appendices C and D provide initialization details and an algorithm overview, respectively.

4 THEORETICAL ANALYSIS

We organize the main analysis around four questions—whether heterogeneous anchor coordinates affect the global decision rule, when semantic transitions remain discriminative, how calibrated evidence combines across views, and whether MAP-EM optimizes a coherent objective—followed by computational complexity. Proofs and technical conditions are deferred to the appendix.

Proposition 4.1 (Correspondence-free semantic evidence; App. E.1)**.** *Define* $b_{ik}^{(v)} = f_{ik}^{(v)} / \sum_j f_{ij}^{(v)}$ *and* $\mathbf{b}_i^{(v)} = (b_{i1}^{(v)}, \dots, b_{iC}^{(v)})^\top$. *Then* $q_{ic} = \frac{\pi_c \prod_v [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_c}{\sum_{c'=1}^{C} \pi_{c'} \prod_v [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_{c'}}$. *Moreover, let* $\mathbf{P}^{(v)}$ *be any* $m_v \times m_v$ *permutation matrix on the anchor coordinates. Jointly replacing* $\mathbf{u}_i^{(v)}$ *and every* $\boldsymbol{\mu}_k^{(v)}$ *by* $\mathbf{P}^{(v)}\mathbf{u}_i^{(v)}$ *and* $\mathbf{P}^{(v)}\boldsymbol{\mu}_k^{(v)}$ *leaves all calibrated likelihoods and global posteriors unchanged.*

The proposition separates two invariances. Jointly relabeling the anchor coordinates of $\mathbf{u}_i^{(v)}$ and $\boldsymbol{\mu}_k^{(v)}$ leaves every local emission $f_{ik}^{(v)}$ unchanged; independently, a common positive rescaling $\widetilde{f}_{ik}^{(v)} = \lambda_i^{(v)} f_{ik}^{(v)}$ over k leaves $\mathbf{b}_i^{(v)}$ and hence q_{ic} unchanged. Thus global inference uses the relative C-state semantic evidence rather than anchor labels or its absolute likelihood scale. The result formalizes "beyond anchor correspondence": each view communicates through a C-dimensional semantic interface independent of its anchor cardinality. The permutation invariance concerns anchor-coordinate relabeling only—not semantic-state relabeling or changes in anchor locations, counts, or construction.

Proposition 4.2 (Calibrated semantic discriminability; Appendix E.2)**.** *Under the ambient spherical measure* $d\sigma(\mathbf{u})$*, define* $G_{k\ell}^{(v)} = \int f_k^{(v)}(\mathbf{u}) f_\ell^{(v)}(\mathbf{u})\, d\sigma(\mathbf{u})$*, the Gram matrix* $\mathbf{G}^{(v)} = [G_{k\ell}^{(v)}]$*, and* $\boldsymbol{\delta}_{cc'}^{(v)} = \mathbf{T}_{c:}^{(v)} - \mathbf{T}_{c':}^{(v)}$*. Then* $\|\psi_c^{(v)} - \psi_{c'}^{(v)}\|_{L_2}^2 = \boldsymbol{\delta}_{cc'}^{(v)} \mathbf{G}^{(v)} (\boldsymbol{\delta}_{cc'}^{(v)})^\top$*, where* $G_{k\ell}^{(v)} = \frac{\mathcal{C}_{m_v}(\kappa^{(v)})^2}{\mathcal{C}_{m_v}\left(\kappa^{(v)} \|\boldsymbol{\mu}_k^{(v)} + \boldsymbol{\mu}_\ell^{(v)}\|_2\right)}$*. If the local emissions are linearly independent in* $L_2(d\sigma)$*,* $\mathbf{G}^{(v)}$ *is positive definite and distinct transition rows induce distinct global-class conditional densities.*

The proposition characterizes calibrated semantic discriminability rather than claiming that semantic calibration universally guarantees class separation. The matrix $\mathbf{G}^{(v)}$ captures the intrinsic geometry of view-local semantic states through their emission overlap, while $\mathbf{T}^{(v)}$ determines how global classes are embedded into this local semantic geometry. Specifically, $\boldsymbol{\delta}_{cc'}^{(v)}$ measures the difference between transition profiles of two global classes, and the resulting separation is preserved only along directions supported by $\mathbf{G}^{(v)}$. Therefore, semantic shifts do not necessarily destroy discriminability: useful calibration is possible when local emissions remain distinguishable and different global classes induce different transition profiles. The closed-form overlap of the vMF emissions further reveals that this geometry is determined by the concentration parameters and angular relationships among semantic prototypes. Thus useful semantic evidence requires both a differentiated transition channel and locally distinguishable components; concentration alone is not a view-quality score.

Proposition 4.3 (Transition-induced semantic informativeness; Appendix E.3)**.** *Under the Markov chain* $Y_i \to L_i^{(v)} \to \mathbf{u}_i^{(v)}$*,* $I(Y_i; \mathbf{u}_i^{(v)}) \le I(Y_i; L_i^{(v)}) = \sum_{c,k} \pi_c T_{ck}^{(v)} \log \frac{T_{ck}^{(v)}}{\sum_c \pi_c T_{ck}^{(v)}}$*. If all rows of* $\mathbf{T}^{(v)}$ *are identical, view* v *carries no same-index global semantic information. Conversely, a permutation channel can retain full class information even when its diagonal preservation is zero under the chosen indexing.*

The proposition characterizes semantic informativeness through the transition channel rather than diagonal preservation alone. In particular, the retained information depends jointly on the dis-

tinguishability of transition profiles and the class prior distribution. Equivalently, $I(Y_i; L_i^{(v)}) = \sum_c \pi_c D_{\mathrm{KL}}\left(\mathbf{T}_{c:}^{(v)} \| \mathbf{p}^{(v)}\right)$, where $\mathbf{p}^{(v)} = \sum_c \pi_c \mathbf{T}_{c:}^{(v)}$. Thus, identical transition rows eliminate semantic information, while distinguishable transition profiles can preserve information even under nontrivial semantic shifts. Moreover, sufficiently represented classes (non-negligible π_c) are required for their transition patterns to contribute to the overall semantic information.

Theorem 4.4 (Multi-view Bhattacharyya error bound; Appendix E.4)**.** *Assume positive class priors and conditional independence under the working model. Let* $B_{cc'}^{(v)} = \int \sqrt{\psi_c^{(v)}(\mathbf{u})\psi_{c'}^{(v)}(\mathbf{u})}\, d\sigma(\mathbf{u})$. *The Bayes classification error satisfies* $R^\star \le \sum_{c<c'} \sqrt{\pi_c \pi_{c'}} \prod_v B_{cc'}^{(v)}$.

The product form characterizes how conditionally independent calibrated views can contribute complementary class evidence; it is a population-level bound and does not guarantee that every finite-sample EM run improves when another view is added.

Theorem 4.5 (MAP-EM monotonicity and objective convergence; Appendix E.5)**.** *Assume fixed priors with* $a_\rho, b_\rho > 1$ *and Dirichlet parameters* $1 + \gamma_S \omega_{ck}^{(v)} \ge 1$*, an exact E-step, and fixed* $\epsilon_\kappa \in (0, 1)$. *Then the MAP-EM iterations generate a non-decreasing and bounded objective-value sequence, i.e.,* $\{\mathcal{J}(\mathbf{\Theta}^t)\}$ *converges to a finite limit.*

Computational Complexity. For view v, evaluating vMF emissions and updating emission statistics require $O(NCm_v)$ operations, while transition mixing and posterior updates require $O(NC^2)$ operations. The M-step updates of transition parameters and mean directions cost $O(VC^2)$ and $O(NC\sum_v m_v)$, respectively, and the scalar updates of $\kappa^{(v)}$ introduce negligible overhead. Thus, one iteration costs $O(NC\sum_{v=1}^V m_v + NVC^2)$ operations. The required memory is $O(NC + VC^2 + C\sum_v m_v)$ for storing posteriors, sufficient statistics, and model parameters.

5 Empirical evaluation and semantic-shift diagnostics

5.1 Experimental setup

In our experiments, 8 publicly available datasets (EuroSAT, MSRC, RSICB128, YTF10, YTF20, YTF50, CIFAR10, CIFAR100; details in Appendix F.1), 11 representative algorithms (LMVSC (Kang et al., 2020), FPMVS (Wang et al., 2022), OMSC (Chen et al., 2022), EMVGC (Wen et al., 2023), FDAGF (Zhang et al., 2023), AEVC (Liu et al., 2024a), ESTMC (Ji & Feng, 2025), MVS4C (Lu et al., 2025), RTGD (Zou et al., 2025), TMC (Wang et al., 2026), SLWSE (Tang et al., 2026)) and 3 well-known metrics (Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI)) are used for empirical evaluation. Besides evaluations on the original datasets, we further conduct controlled semantic-shift experiments under two synthetic shift settings, namely Euclidean relocation and spherical rotation, to quantitatively assess the capability of different methods in handling view-specific semantic transitions. The construction details of these controlled shifts are provided in Appendix F.2. The input anchor graphs of our model are generated by the basic anchor subspace model $\min_{\mathbf{A},\mathbf{Z}} \|\mathbf{X} - \mathbf{Z}\mathbf{A}\|_F^2, \mathbf{Z}\mathbf{1} = \mathbf{1}, \mathbf{Z} \ge 0$. In particular, we denote the variant of CAST with heterogeneous anchor cardinalities by CAST-H. For all compared methods, hyperparameters other than the anchor number are searched according to the original papers. For the anchor number, we use a unified search range $m \in \{C, 2C, 3C\}$, where C is the number of clusters. We report the best mean performance over five runs. All experiments are conducted on a machine with an AMD 9950X CPU and 64GB RAM.

Ablation variants. We study the role of semantic transition modeling by progressively restricting the structure of $\mathbf{T}^{(v)}$. Identity removes transition modeling by fixing $\mathbf{T}^{(v)} = \mathbf{I}$. ViewLevel assumes a shared shift rate for each view and uniform transition destinations. Classwise allows class-dependent shift rates but keeps destination probabilities uniform. SharedDest further models class-dependent shift rates while constraining all source classes to share the same destination tendency. $\text{CAST}_{(\gamma_S=0)}$ uses the full class-level transition structure but removes the Dirichlet shrinkage prior. $\text{CAST}_{(\gamma_S=2)}$ adopts the complete transition parameterization with weak Dirichlet regularization. All variants share the same initialization, local emission model, and inference procedure, differing only in the transition structure.

5.2 EXPERIMENTAL ANALYSIS

Effectiveness under natural and controlled semantic shifts. Table 1 reports the clustering performance on both original datasets and datasets with controlled semantic shifts. Several observations are noteworthy. First, on the original datasets without manually introduced shifts, CAST already achieves highly competitive performance and consistently ranks among the best-performing methods. This indicates that CAST does not rely on artificially created semantic transitions, but can effectively exploit naturally existing semantic discrepancies among views. Second, under both Euclidean relocation and spherical rotation settings, CAST maintains superior performance compared with existing methods, while the performance degradation is generally smaller. This demonstrates that explicitly modeling view-specific semantic transitions provides better adaptability when the semantic correspondence between views becomes heterogeneous. Third, on shifted YTF10, YTF20, and YTF50 datasets, most comparison methods suffer from noticeable performance drops, whereas CAST occasionally achieves improvements over the original setting. This observation supports our motivation that semantic transition does not necessarily imply semantic information loss. When the systematic transition pattern is correctly identified and calibrated, the transformed semantic structures can provide additional discriminative information for clustering.

Effect of semantic transition modeling. Table 3 investigates the necessity and effectiveness of the proposed transition matrices $\mathbf{T}^{(v)}$ through ablation studies. First, on original datasets, the Identity variant performs competitively on EuroSAT, RSICB128, CIFAR10, and CIFAR100. This is consistent with the characteristics of these datasets, where different views are extracted from the same images using different neural networks with similar architectures and pretraining objectives. Consequently, the semantic discrepancy among views is relatively limited, and an identity transition assumption is already reasonable. In contrast, CAST substantially outperforms Identity on MSRC and YTF10 even without manually introduced shifts. These datasets contain views generated from heterogeneous handcrafted feature descriptors designed from different perspectives, resulting in stronger semantic heterogeneity. The improvement demonstrates that semantic transition is not merely an artificial phenomenon, but naturally exists in multi-view data. Furthermore, under controlled semantic-shift settings, all simplified variants of $\mathbf{T}^{(v)}$ are inferior to CAST, and some even perform worse than Identity. This verifies that the proposed class-level transition modeling achieves an appropriate balance between flexibility and constraint: it is expressive enough to capture systematic semantic changes while avoiding excessive freedom that may introduce unreliable transitions.

Effect of heterogeneous anchor numbers. Table 4 evaluates CAST when each view is allowed to use different numbers of anchors. Compared with the default setting where all views share the same anchor number, heterogeneous anchor graphs provide better compatibility with independently constructed view representations. Moreover, allowing view-specific m_v can further improve clustering performance in several cases. This suggests that CAST is not limited by artificial anchor correspondence assumptions and can naturally adapt to heterogeneous anchor structures.

Additional analysis. Figures 2 and 3 report the view-wise ACC, mean semantic preservation, and normalized semantic information under the two controlled semantic-shift settings. The consistent trends among these measures indicate that CAST effectively captures view-specific semantic variations. Their differences are also reasonable, since semantic transition does not necessarily imply semantic information loss. Figure 1 shows the convergence behavior of CAST, demonstrating stable optimization. Figures 4 and 5 visualize the learned transition matrices $\mathbf{T}^{(v)}$, providing intuitive evidence of view-specific semantic calibration.

6 CONCLUSION

CAST addresses semantic heterogeneity in multi-view clustering by explicitly modeling view-specific class-level semantic transitions rather than assuming fixed cross-view semantic correspondence. By jointly inferring global assignments, view-local semantic states, and transition relations, CAST can identify systematic semantic shifts and calibrate heterogeneous views without requiring anchor correspondence. Extensive experiments on original benchmarks and controlled semantic-shift settings demonstrate that semantic transitions naturally exist in multi-view data and can be effectively exploited to improve clustering performance.

Table 1: Clustering performance (%) under the standard setting and Euclidean and spherical shifts. Higher values are better. Bold and underlined values denote the best and second-best results for each setting, dataset, and metric. A dash indicates an out of memory error.

Setting	Method	EuroSAT			MSRC-v6			RSICB128			YTF10		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Standard	LMVSC	73.60	59.84	52.52	63.33	54.53	41.36	68.21	78.02	62.37	71.27	76.89	64.06
	FPMVS	85.94	78.41	75.81	86.19	78.09	72.32	73.72	80.99	55.22	74.96	78.69	67.06
	OMSC	84.28	77.91	75.22	83.81	72.95	68.51	77.99	86.53	72.96	79.22	81.07	72.11
	EMVGCLG	80.40	81.00	75.31	39.05	31.29	15.15	75.92	86.60	73.86	69.35	69.76	50.55
	FDAGF	89.35	79.88	78.85	87.62	79.02	74.44	74.07	87.58	73.38	73.04	77.84	65.42
	AEVC	81.12	69.98	67.01	85.24	75.83	70.45	74.39	85.69	71.32	75.53	78.17	68.63
	ESTMC	77.00	73.12	67.84	87.62	77.80	75.21	73.20	85.32	71.29	77.48	76.87	66.05
	MVS4C	87.24	76.71	75.03	75.40	62.34	55.16	69.82	79.30	65.01	73.57	75.47	64.98
	RTGD	68.07	69.30	60.78	57.14	50.26	38.05	75.51	83.67	70.95	77.92	82.18	69.77
	TMC	81.96	77.19	72.75	83.52	73.10	66.79	72.07	86.55	70.16	78.77	81.39	71.62
	SLWSE	83.58	78.45	73.83	77.14	74.86	65.80	77.10	87.75	71.93	81.32	81.91	72.15
	CAST	**89.77**	**81.01**	**79.26**	**95.71**	**90.97**	**90.12**	**79.16**	**88.40**	**78.64**	**88.41**	**84.18**	**76.74**
Euclidean	LMVSC	40.96	40.76	23.95	31.90	25.13	12.61	41.98	71.02	38.97	43.69	56.00	33.23
	FPMVS	67.18	56.46	46.44	43.81	29.26	16.71	65.98	80.11	62.86	68.07	69.68	56.92
	OMSC	61.96	57.95	44.27	45.71	30.81	19.61	69.08	79.54	65.79	64.55	64.00	54.93
	EMVGCLG	85.48	75.00	72.25	48.57	39.72	25.31	67.99	83.63	63.52	70.38	76.26	64.79
	FDAGF	89.81	80.00	79.18	70.48	60.50	52.45	77.74	89.64	76.98	71.87	76.79	66.49
	AEVC	59.00	46.17	37.51	52.38	34.65	25.91	59.29	78.87	58.53	82.90	78.97	74.80
	ESTMC	51.35	48.69	34.36	44.29	32.97	18.95	46.69	74.03	45.22	68.97	73.09	62.93
	MVS4C	53.19	45.53	33.05	40.00	28.83	16.47	58.37	70.31	51.77	75.48	73.05	63.53
	RTGD	58.53	57.64	43.43	43.33	33.38	21.49	60.53	84.24	65.15	70.90	80.15	68.91
	TMC	66.90	61.64	50.20	42.45	26.24	15.72	59.05	81.60	58.77	65.71	62.63	53.65
	SLWSE	61.57	53.22	40.31	48.10	32.49	22.12	67.33	81.63	65.63	71.09	73.73	62.44
	CAST	**93.81**	**86.82**	**86.98**	**77.14**	**66.26**	**57.11**	**83.34**	**91.56**	**81.94**	**97.59**	**94.94**	**95.38**
Spherical	LMVSC	42.96	43.13	26.00	43.81	34.96	20.81	41.99	70.87	38.40	43.83	52.97	30.26
	FPMVS	64.14	57.92	47.33	43.81	26.66	14.44	66.18	80.02	62.68	67.14	67.37	55.67
	OMSC	61.92	58.06	44.51	41.43	26.92	15.82	70.98	82.08	65.97	66.49	69.60	58.84
	EMVGCLG	86.12	75.54	73.20	49.05	36.86	21.04	65.36	81.69	65.60	69.23	71.62	61.68
	FDAGF	89.38	79.17	78.41	70.00	58.73	48.92	73.79	86.67	71.56	85.56	84.45	77.94
	AEVC	51.14	46.73	34.89	50.00	31.84	23.11	64.47	80.44	64.33	78.92	81.16	73.13
	ESTMC	46.39	47.24	31.81	41.90	31.92	17.80	45.12	72.20	43.23	69.31	73.37	63.33
	MVS4C	49.12	44.50	30.57	37.62	27.53	14.70	56.78	70.05	49.71	75.57	73.08	63.51
	RTGD	48.23	47.67	30.64	58.10	40.65	29.30	61.15	83.33	65.84	70.75	78.95	67.52
	TMC	52.71	52.78	39.75	41.76	29.31	17.32	58.03	80.79	56.66	63.65	69.84	54.02
	SLWSE	62.63	56.05	44.33	41.90	31.09	18.09	68.33	79.86	62.77	64.61	71.25	60.13
	CAST	**93.26**	**85.80**	**85.80**	**75.71**	**63.93**	**55.72**	**81.03**	**92.73**	**83.39**	**96.87**	**93.67**	**93.71**

Setting	Method	YTF20			YTF50			CIFAR10			CIFAR100		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Standard	LMVSC	65.10	74.76	52.78	66.89	80.59	55.20	88.98	78.91	76.93	76.81	93.67	73.38
	FPMVS	68.66	75.15	61.16	–	–	–	98.99	97.30	97.78	–	–	–
	OMSC	69.86	76.18	61.22	–	–	–	97.99	95.37	95.68	–	–	–
	EMVGCLG	66.67	76.95	55.13	69.09	83.22	61.73	99.01	97.40	97.83	91.61	98.15	92.77
	FDAGF	65.62	77.31	53.78	67.87	80.56	56.57	98.72	96.58	97.20	85.42	96.72	86.56
	AEVC	70.34	76.24	60.53	71.91	82.82	61.69	96.27	92.17	92.18	92.18	98.14	93.20
	ESTMC	75.30	78.47	64.59	73.62	83.99	62.35	**99.34**	**98.19**	**98.55**	93.50	98.46	94.20
	MVS4C	65.33	71.66	55.99	68.89	77.10	60.80	97.53	96.20	95.51	96.43	98.39	95.65
	RTGD	66.68	77.37	60.99	64.25	82.27	56.53	85.79	93.81	86.97	76.67	90.94	55.55
	TMC	71.38	78.74	59.33	71.50	82.89	60.08	98.97	97.13	97.74	91.61	98.63	92.81
	SLWSE	67.98	77.23	60.92	–	–	–	98.98	97.42	97.78	96.57	**99.39**	96.73
	CAST	**76.49**	**80.66**	**65.88**	**76.56**	**85.98**	**70.32**	99.07	97.48	97.97	**96.97**	99.05	**96.99**
Euclidean	LMVSC	43.25	62.52	33.99	41.31	70.66	35.87	46.73	55.63	33.88	43.79	80.08	40.49
	FPMVS	64.77	69.78	54.77	–	–	–	75.10	68.26	59.03	–	–	–
	OMSC	68.07	75.44	61.01	–	–	–	70.90	64.21	51.76	–	–	–
	EMVGCLG	66.82	74.12	49.93	69.24	81.99	58.90	76.64	74.80	64.47	51.65	82.52	44.98
	FDAGF	71.91	79.49	63.58	67.37	81.73	59.07	95.23	89.09	89.81	80.47	93.51	70.54
	AEVC	72.01	76.57	63.82	75.33	83.84	65.66	69.21	62.99	53.08	75.17	91.42	72.19
	ESTMC	71.49	75.43	61.87	72.38	82.85	62.06	69.02	75.93	64.29	81.23	94.34	80.74
	MVS4C	59.91	64.02	47.19	63.84	72.84	53.58	80.25	82.19	77.06	87.55	92.39	82.99
	RTGD	70.53	80.81	64.41	70.85	85.15	65.17	96.56	92.08	92.56	82.84	93.99	75.09
	TMC	62.90	74.09	51.74	65.30	79.28	55.72	75.16	76.78	66.66	64.02	88.85	61.09
	SLWSE	65.08	75.00	59.54	–	–	–	87.59	80.29	75.21	70.63	89.74	67.14
	CAST	**84.90**	**89.54**	**82.33**	**82.86**	**90.83**	**79.74**	**97.95**	**94.87**	**95.56**	**95.99**	**98.77**	**96.07**
Spherical	LMVSC	41.93	61.18	32.00	40.61	69.87	33.99	46.58	53.14	32.83	45.51	80.36	40.99
	FPMVS	63.73	71.74	48.53	–	–	–	71.56	64.26	49.70	–	–	–
	OMSC	64.04	71.29	55.30	–	–	–	63.84	62.56	49.43	–	–	–
	EMVGCLG	61.45	72.51	53.57	72.41	83.57	62.20	77.84	80.10	73.74	50.27	83.03	46.28
	FDAGF	72.52	79.73	65.64	69.58	82.04	57.81	91.95	82.72	83.18	80.59	94.06	74.77
	AEVC	70.03	74.63	59.87	73.85	82.65	64.89	77.18	67.97	60.94	84.16	95.11	83.57
	ESTMC	74.53	76.32	62.39	71.42	82.56	62.02	68.23	74.74	63.06	81.48	94.35	80.81
	MVS4C	63.65	65.15	51.86	64.05	72.85	53.75	86.59	80.80	74.85	89.31	94.96	86.88
	RTGD	74.07	83.06	69.09	69.39	81.37	50.94	96.96	93.18	93.44	81.67	93.67	74.31
	TMC	61.88	74.45	51.55	65.23	79.26	54.73	73.03	75.39	64.38	63.23	88.61	60.23
	SLWSE	63.62	72.69	56.63	–	–	–	80.56	75.43	66.45	72.91	90.24	68.13
	CAST	**86.02**	**91.45**	**85.24**	**84.12**	**91.70**	**82.35**	**97.77**	**94.40**	**95.17**	**95.28**	**98.75**	**95.62**

AI USE STATEMENT

In this work, we used generative AI tools for assistance with manuscript preparation, including language polishing, improving clarity and organization, and literature discovery. We also used generative AI tools for discussing experimental design and evaluation settings, as well as for checking the consistency of mathematical derivations. These uses were limited to feedback and verification; the core research ideas, methodology, theoretical developments, experimental results, and conclusions were developed by the authors.

We have not used generative AI tools to generate research contributions, prove mathematical claims, create datasets, or produce experimental results. All AI-assisted suggestions were critically reviewed by the authors. The mathematical derivations were developed by the authors and AI-assisted feedback was only used to identify potential inconsistencies or errors. All implementations and experimental results were validated through author-conducted experiments. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

Man-Sheng Chen, Chang-Dong Wang, Dong Huang, Jian-Huang Lai, and Philip S Yu. Efficient orthogonal multi-view subspace clustering. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 127–135, 2022.

Yawei Chen, Huibing Wang, Jinjia Peng, and Yang Wang. Anchor learning with potential cluster constraints for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(15), pp. 15939–15947, 2025a.

Zhe Chen, Mingzhi Zhu, Hui Li, and Tianyang Xu. Anchor graph learning with double noise removal for multi-view clustering. *Neural Networks*, 191:107779, 2025b.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28 (1):20–28, 1979.

Jintian Ji and Songhe Feng. Anchors crash tensor: Efficient and scalable tensorial multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2660–2675, 2025.

Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(04), pp. 4412–4419, 2020.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26151–26161, 2024a.

Suyuan Liu, Qing Liao, Siwei Wang, Xinwang Liu, and En Zhu. Robust and consistent anchor graph learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36 (8):4207–4219, 2024b.

Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, 2013.

Zhoumin Lu, Feiping Nie, Linru Ma, Rong Wang, and Xuelong Li. Simplifying scalable subspace clustering and its multi-view extension by anchor-to-sample kernel. *IEEE Transactions on Image Processing*, 34:5084–5098, 2025.

Chuan Tang, Miaomiao Li, Jun Wang, Chang Tang, En Zhu, and Xinwang Liu. Sample-level weighted and structure-enhanced anchor graph learning for scalable multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Jun Wang, Miaomiao Li, Zhenglai Li, Hao Yu, Suyuan Liu, Dayu Hu, Chang Tang, and Xinwang Liu. Tensor multi-rank constraint guided anchor-wise adaptive alignment for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

Siwei Wang, Xinwang Liu, Xinzhong Zhu, Pei Zhang, Yi Zhang, Feng Gao, and En Zhu. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. *IEEE Transactions on Image Processing*, 31:556–568, 2022.

Yi Wen, Suyuan Liu, Xinhang Wan, Siwei Wang, Ke Liang, Xinwang Liu, Xihong Yang, and Pei Zhang. Efficient multi-view graph clustering with local and global structure preservation. In *Proceedings of the ACM International Conference on Multimedia*, pp. 3021–3030, 2023.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, pp. 95–103, 1983.

Chao Zhang, Xiuyi Jia, Zechao Li, Chunlin Chen, and Huaxiong Li. Learning cluster-wise anchors for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(15), pp. 16696–16704, 2024.

Pei Zhang, Siwei Wang, Liang Li, Changwang Zhang, Xinwang Liu, En Zhu, Zhe Liu, Lu Zhou, and Lei Luo. Let the data choose: Flexible and diverse anchor graph fusion for scalable multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(9), pp. 11262–11269, 2023.

Pei Zhang, Yuangang Pan, Siwei Wang, Shengju Yu, Huiying Xu, En Zhu, Xinwang Liu, and Ivor Tsang. Max-mahalanobis anchors guidance for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(21), pp. 22488–22496, 2025.

Jiale Zou, Yan Chen, Bingbing Jiang, Peng Zhou, Liang Du, Lei Duan, and Yuhua Qian. Robust tensor learning with graph diffusion for scalable multi-view graph clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2207–2215, 2025.

A SOURCES AND INTERPRETATION OF SEMANTIC SHIFT

Modality-dependent local semantics. Let an object O_i be observed through $\mathbf{x}_i^{(v)} = \Phi^{(v)}(O_i)$. When $\Phi^{(v)}$ measures different physical attributes, class pairs that are separable in one view may overlap in another. Appearance, depth, motion, text, audio, and sensor measurements can therefore induce different class-conditional neighborhoods without any view being "wrong".

Representation and feature-extractor variation. Even within a common modality, handcrafted descriptors, convolutional networks, transformers, or self-supervised encoders impose different metrics. Since anchor assignments inherit these geometries, the same global class can occupy different local semantic regions across feature spaces.

Anchor-construction heterogeneity. Anchor assignments are secondary representations of the features. View v may use a distinct selector, optimization rule, neighborhood width, sparsity level, or anchor budget m_v. Consequently, two views may have unrelated anchor identities and even different simplex dimensions. Unequal anchor counts alone do not imply a semantic shift: a faithful representation can preserve class semantics despite different coordinates. The transition is learned from the induced class-level evidence, not inferred from m_v.

Missing, degraded, and imperfectly associated observations. Missing feature coordinates, masking, imputation, reconstruction, or degraded acquisition can alter local neighborhoods. Imperfect observation association can likewise modify the effective class-conditional relation when nominally paired observations belong to different categories. These mechanisms are mentioned as possible upstream causes only; CAST does not estimate a missingness mechanism or an instance matching matrix.

Measurement variation and noise. Illumination, compression, sensor noise, stochastic feature extraction, or acquisition conditions may move an observation within its representation space. Perturbation magnitude and semantic effect are not equivalent: a large change may remain in the same class-level region, while a small change near a boundary can alter the local semantic state. Noise is therefore one possible generator of semantic shift rather than the defining task.

B EXACT POSTERIOR AND MAP-EM DERIVATION

B.1 EXACT E-STEP POSTERIOR AND DIAGNOSTICS

For fixed parameters, Eq. (5) is the class-conditional likelihood of one view after marginalizing its local semantic state:

$$p(\mathbf{u}_i^{(v)} \mid Y_i = c, \mathbf{\Theta}) = \sum_{k=1}^{C} \Pr(L_i^{(v)} = k \mid Y_i = c, \mathbf{\Theta}) p(\mathbf{u}_i^{(v)} \mid L_i^{(v)} = k, \mathbf{\Theta}) = \sum_{k=1}^{C} T_{ck}^{(v)} f_{ik}^{(v)} = \psi_{ic}^{(v)}. \tag{18}$$

Conditioned on $Y_i = c$, the view-local branches are independent under Eq. (6), so

$$p(\mathcal{U}_i \mid Y_i = c, \mathbf{\Theta}) = \prod_{v=1}^{V} \psi_{ic}^{(v)}. \tag{19}$$

Bayes' rule therefore gives

$$\Pr(Y_i = c \mid \mathcal{U}_i, \mathbf{\Theta}) = \frac{\Pr(Y_i = c \mid \mathbf{\Theta}) p(\mathcal{U}_i \mid Y_i = c, \mathbf{\Theta})}{\sum_{c'=1}^{C} \Pr(Y_i = c' \mid \mathbf{\Theta}) p(\mathcal{U}_i \mid Y_i = c', \mathbf{\Theta})} = \frac{\pi_c \prod_v \psi_{ic}^{(v)}}{\sum_{c'=1}^{C} \pi_{c'} \prod_v \psi_{ic'}^{(v)}} = q_{ic}, \tag{20}$$

which proves Eq. (12).

For the joint global/local posterior in view v, the chain rule gives

$$\Pr(Y_i = c, L_i^{(v)} = k \mid \mathcal{U}_i, \mathbf{\Theta}) = q_{ic} \Pr(L_i^{(v)} = k \mid Y_i = c, \mathcal{U}_i, \mathbf{\Theta}). \tag{21}$$

Given $Y_i = c$, the state $L_i^{(v)}$ is conditionally independent of observations in the other views, hence

$$\Pr(L_i^{(v)} = k \mid Y_i = c, \mathcal{U}_i, \mathbf{\Theta}) = \Pr(L_i^{(v)} = k \mid Y_i = c, \mathbf{u}_i^{(v)}, \mathbf{\Theta}) = \frac{T_{ck}^{(v)} f_{ik}^{(v)}}{\sum_{j=1}^{C} T_{cj}^{(v)} f_{ij}^{(v)}} = \frac{T_{ck}^{(v)} f_{ik}^{(v)}}{\psi_{ic}^{(v)}}. \tag{22}$$

Substitution into Eq. (21) yields

$$\Pr(Y_i = c, L_i^{(v)} = k \mid \mathcal{U}_i, \mathbf{\Theta}) = q_{ic} \frac{T_{ck}^{(v)} f_{ik}^{(v)}}{\psi_{ic}^{(v)}} = \xi_{ick}^{(v)}, \tag{23}$$

which proves Eq. (13). Since $\sum_k T_{ck}^{(v)} f_{ik}^{(v)} = \psi_{ic}^{(v)}$,

$$\sum_k \xi_{ick}^{(v)} = q_{ic}, \qquad r_{ik}^{(v)} := \sum_c \xi_{ick}^{(v)} = \Pr(L_i^{(v)} = k \mid \mathcal{U}_i, \mathbf{\Theta}). \tag{24}$$

The same joint posterior yields the sample–view semantic-shift posterior

$$\tau_i^{(v)} = \sum_c \sum_{k \neq c} \xi_{ick}^{(v)} = 1 - \sum_c \xi_{icc}^{(v)}. \tag{25}$$

Its complement $1 - \tau_i^{(v)} = \sum_c \xi_{icc}^{(v)}$ is the sample–view semantic-preservation posterior under the chosen local component indexing. At the aggregate level, $\overline{\tau} := (NV)^{-1} \sum_{i,v} \tau_i^{(v)}$ summarizes posterior sample–view shift mass, whereas $\overline{\rho} := (VC)^{-1} \sum_{v,c} \rho_c^{(v)}$ summarizes class/view-level transition parameters. In general $\overline{\tau} \neq \overline{\rho}$ because classes need not have equal posterior sizes and the Beta prior modifies the MAP rates. Hence sample-level preservation/shift diagnostics should not be interpreted as direct averages of the learned transition parameters.

Thus the pairwise posteriors q_{ic} and $\xi_{ick}^{(v)}$ are sufficient for all E-step statistics and diagnostics; no enumeration of the C^V joint local-state configurations is required.

B.2 MAP AUXILIARY OBJECTIVE

At a fixed E-step, let q_{ic} and $\xi_{ick}^{(v)}$ be the posteriors in Eqs. (12)–(13), let $r_{ik}^{(v)} = \sum_c \xi_{ick}^{(v)}$, and use the counts in Eq. (14). For a realized pair $(Y_i = c, L_i^{(v)} = k)$, the transition log term from Eq. (4) is

$$\log T_{ck}^{(v)} = \mathbb{I}[k = c] \log(1 - \rho_c^{(v)}) + \mathbb{I}[k \neq c] \left[\log \rho_c^{(v)} + \log S_{ck}^{(v)} \right]. \tag{26}$$

Taking expectation under $\xi_{ick}^{(v)}$ converts the three indicator-weighted terms in Eq. (26) into $n_{cc}^{(v)}$, $n_{-c}^{(v)}$, and $n_{ck}^{(v)}$, respectively. Likewise, summing the emission responsibility $\xi_{ick}^{(v)}$ over the global class gives $r_{ik}^{(v)}$. Therefore, taking the posterior expectation of the complete-data log density in Eq. (7) gives

$$\begin{aligned} \mathcal{Q}_{\text{like}}(\mathbf{\Theta}) = & \sum_c n_c \log \pi_c + \sum_{v,c} \left[n_{cc}^{(v)} \log(1 - \rho_c^{(v)}) + n_{-c}^{(v)} \log \rho_c^{(v)} \right] \\ & + \sum_{v,c} \sum_{k \neq c} n_{ck}^{(v)} \log S_{ck}^{(v)} \\ & + \sum_{v,i,k} r_{ik}^{(v)} \left[\log \mathcal{C}_{m_v}(\kappa^{(v)}) + \kappa^{(v)} (\boldsymbol{\mu}_k^{(v)})^{\mathsf{T}} \mathbf{u}_i^{(v)} \right]. \end{aligned} \tag{27}$$

The Beta and Dirichlet priors in Eq. (10) contribute, up to constants independent of the optimized parameters,

$$\begin{aligned} \log p\Big(\{\rho_c^{(v)}\}_{v,c}, \{\mathbf{S}_{-c}^{(v)}\}_{v,c} \Big) = & \sum_{v,c} \left[(a_\rho - 1) \log \rho_c^{(v)} + (b_\rho - 1) \log(1 - \rho_c^{(v)}) \right] \\ & + \sum_{v,c} \sum_{k \neq c} \gamma_S \omega_{ck}^{(v)} \log S_{ck}^{(v)}. \end{aligned} \tag{28}$$

Hence the MAP auxiliary objective is

$$\begin{aligned}\mathcal{Q}_{\mathrm{MAP}}(\boldsymbol{\Theta}) = &\sum_c n_c \log \pi_c \\ &+ \sum_{v,c} \Big[(n_{-c}^{(v)} + a_\rho - 1) \log \rho_c^{(v)} + (n_{cc}^{(v)} + b_\rho - 1) \log(1 - \rho_c^{(v)}) \Big] \\ &+ \sum_{v,c} \sum_{k \neq c} (n_{ck}^{(v)} + \gamma_S \omega_{ck}^{(v)}) \log S_{ck}^{(v)} \\ &+ \sum_{v,i,k} r_{ik}^{(v)} \Big[\log \mathcal{C}_{m_v}(\kappa^{(v)}) + \kappa^{(v)} (\boldsymbol{\mu}_k^{(v)})^\mathsf{T} \mathbf{u}_i^{(v)} \Big].\end{aligned} \tag{29}$$

For brevity, in the blockwise derivations below, $\mathcal{Q}(\cdot)$ denotes the terms of $\mathcal{Q}_{\mathrm{MAP}}$ that depend on the displayed parameter block, with all remaining quantities held fixed. Given the E-step posteriors, the parameter families $\boldsymbol{\pi}$, $\{\rho_c^{(v)}\}_{v,c}$, $\{\mathbf{S}_{-c}^{(v)}\}_{v,c}$, and the view-specific emission parameters separate as shown below.

B.3 Global proportions

The $\boldsymbol{\pi}$ block is

$$\mathcal{Q}(\boldsymbol{\pi}) = \sum_c n_c \log \pi_c, \qquad \pi_c \geq 0, \quad \sum_c \pi_c = 1. \tag{30}$$

Using a multiplier λ for $1 - \sum_c \pi_c$, the Lagrangian is

$$\mathscr{L}_\pi = \sum_c n_c \log \pi_c + \lambda \left(1 - \sum_c \pi_c \right). \tag{31}$$

For an occupied class, stationarity gives $n_c/\pi_c - \lambda = 0$, so $\pi_c = n_c/\lambda$. Because

$$\sum_c n_c = \sum_i \sum_c q_{ic} = N, \tag{32}$$

the simplex constraint gives $\lambda = N$ and therefore

$$\pi_c = \frac{n_c}{N}. \tag{33}$$

The same expression extends to $n_c = 0$ by the simplex optimum at the boundary.

B.4 Semantic-shift probabilities

For fixed (v, c), the $\rho_c^{(v)}$ block is

$$\mathcal{Q}(\rho_c^{(v)}) = (n_{-c}^{(v)} + a_\rho - 1) \log \rho_c^{(v)} + (n_{cc}^{(v)} + b_\rho - 1) \log(1 - \rho_c^{(v)}). \tag{34}$$

Its derivative is

$$\frac{\partial \mathcal{Q}(\rho_c^{(v)})}{\partial \rho_c^{(v)}} = \frac{n_{-c}^{(v)} + a_\rho - 1}{\rho_c^{(v)}} - \frac{n_{cc}^{(v)} + b_\rho - 1}{1 - \rho_c^{(v)}}. \tag{35}$$

Setting it to zero yields

$$\rho_c^{(v)} = \frac{n_{-c}^{(v)} + a_\rho - 1}{n_{-c}^{(v)} + n_{cc}^{(v)} + a_\rho + b_\rho - 2}. \tag{36}$$

Moreover,

$$n_{cc}^{(v)} + n_{-c}^{(v)} = \sum_k n_{ck}^{(v)} = \sum_i \sum_k \xi_{ick}^{(v)} = \sum_i q_{ic} = n_c, \tag{37}$$

where Eq. (24) is used in the penultimate equality. Hence Eq. (36) reduces exactly to Eq. (15). For $a_\rho, b_\rho > 1$, the second derivative is strictly negative on $(0, 1)$, so this stationary point is the unique interior maximizer.

B.5 DIRECTIONAL DESTINATIONS

For fixed (v, c), define $a_k = n_{ck}^{(v)} + \gamma_S \omega_{ck}^{(v)}$ for $k \neq c$. The directional block is

$$\mathcal{Q}(\mathbf{S}_{-c}^{(v)}) = \sum_{k \neq c} a_k \log S_{ck}^{(v)}, \qquad S_{ck}^{(v)} \geq 0, \quad \sum_{k \neq c} S_{ck}^{(v)} = 1. \tag{38}$$

For an interior solution, use

$$\mathscr{L}_S = \sum_{k \neq c} a_k \log S_{ck}^{(v)} + \lambda \left(1 - \sum_{k \neq c} S_{ck}^{(v)} \right). \tag{39}$$

Stationarity gives $a_k / S_{ck}^{(v)} - \lambda = 0$, hence $S_{ck}^{(v)} = a_k / \lambda$. Since $\sum_{k \neq c} \omega_{ck}^{(v)} = 1$,

$$\lambda = \sum_{k \neq c} a_k = n_{-c}^{(v)} + \gamma_S, \tag{40}$$

and therefore

$$S_{ck}^{(v)} = \frac{n_{ck}^{(v)} + \gamma_S \omega_{ck}^{(v)}}{n_{-c}^{(v)} + \gamma_S}, \qquad k \neq c. \tag{41}$$

When $\gamma_S > 0$ and $\boldsymbol{\omega}_{-c}^{(v)}$ has positive support, all $a_k > 0$ and the maximizer is interior. If $\gamma_S = 0$ and $n_{-c}^{(v)} = 0$, the block is flat over the off-diagonal simplex; retaining the previous row is an admissible tie-breaking rule.

B.6 MEAN DIRECTIONS

For view v, define the resultant vectors

$$\mathbf{s}_k^{(v)} := \sum_i r_{ik}^{(v)} \mathbf{u}_i^{(v)}. \tag{42}$$

Because $\sum_k r_{ik}^{(v)} = 1$ for every sample,

$$\sum_{i,k} r_{ik}^{(v)} = N. \tag{43}$$

The emission contribution of view v in Eq. (29) can therefore be written as

$$N \log \mathcal{C}_{m_v}(\kappa^{(v)}) + \kappa^{(v)} \sum_k (\boldsymbol{\mu}_k^{(v)})^\mathsf{T} \mathbf{s}_k^{(v)}. \tag{44}$$

For fixed $\kappa^{(v)} > 0$, the block for each mean direction is

$$\mathcal{Q}(\boldsymbol{\mu}_k^{(v)}) = \kappa^{(v)} (\boldsymbol{\mu}_k^{(v)})^\mathsf{T} \mathbf{s}_k^{(v)}, \qquad \|\boldsymbol{\mu}_k^{(v)}\|_2 = 1. \tag{45}$$

Because $\kappa^{(v)} > 0$, Cauchy–Schwarz gives $(\boldsymbol{\mu}_k^{(v)})^\mathsf{T} \mathbf{s}_k^{(v)} \leq \|\mathbf{s}_k^{(v)}\|_2$, with equality at

$$\boldsymbol{\mu}_k^{(v)} = \frac{\mathbf{s}_k^{(v)}}{\|\mathbf{s}_k^{(v)}\|_2}, \tag{46}$$

which proves the first part of Eq. (16) whenever $\mathbf{s}_k^{(v)} \neq \mathbf{0}$. If a resultant is exactly zero, that component's direction is not identified by the current auxiliary objective and any unit vector is optimal; an implementation may retain the previous direction as a deterministic tie rule. At $\kappa^{(v)} = 0$, all mean directions are likewise immaterial.

Substituting the maximizing directions into Eq. (44) gives

$$N \log \mathcal{C}_{m_v}(\kappa^{(v)}) + \kappa^{(v)} \sum_k \|\mathbf{s}_k^{(v)}\|_2, \tag{47}$$

which motivates the normalized resultant

$$\bar{R}^{(v)} := \frac{1}{N} \sum_k \|\mathbf{s}_k^{(v)}\|_2. \tag{48}$$

This is the second quantity in Eq. (16).

B.7 CONCENTRATION

After maximizing the mean directions, the concentration block for view v is

$$\mathcal{Q}(\kappa^{(v)}) = N\Big[\log \mathcal{C}_{m_v}(\kappa^{(v)}) + \kappa^{(v)}\bar{R}^{(v)}\Big], \qquad \kappa^{(v)} \geq 0. \tag{49}$$

For the vMF normalizer,

$$\frac{d}{d\kappa}\log \mathcal{C}_D(\kappa) = -A_D(\kappa), \qquad A_D(\kappa) = \frac{I_{D/2}(\kappa)}{I_{D/2-1}(\kappa)}. \tag{50}$$

For example, this follows by differentiating Eq. (2) and using $I'_\nu(\kappa) = I_{\nu+1}(\kappa) + \nu I_\nu(\kappa)/\kappa$ with $\nu = D/2 - 1$. Hence

$$\frac{\partial \mathcal{Q}(\kappa^{(v)})}{\partial \kappa^{(v)}} = N\big[-A_{m_v}(\kappa^{(v)}) + \bar{R}^{(v)}\big]. \tag{51}$$

The vMF mean-resultant function is continuous and strictly increasing with $A_D(0) = 0$ and $A_D(\kappa) \to 1$ as $\kappa \to \infty$. Instead of imposing the same raw concentration cap in every dimension, CAST fixes $\epsilon_\kappa \in (0, 1)$ (default 10^{-3}) and defines the dimension-dependent finite bound

$$K_v(\epsilon_\kappa) := A_{m_v}^{-1}(1 - \epsilon_\kappa) < \infty. \tag{52}$$

Maximizing Eq. (49) over $0 \leq \kappa^{(v)} \leq K_v(\epsilon_\kappa)$ has two cases. If $\bar{R}^{(v)} \leq 1 - \epsilon_\kappa$, the derivative changes sign at the unique interior root $A_{m_v}(\kappa^{(v)}) = \bar{R}^{(v)}$. If $\bar{R}^{(v)} > 1 - \epsilon_\kappa$, the derivative remains positive at the upper boundary and the constrained maximizer is $K_v(\epsilon_\kappa)$. Both cases are represented by

$$\widetilde{R}^{(v)} := \min\{\bar{R}^{(v)}, 1 - \epsilon_\kappa\}, \qquad A_{m_v}(\kappa^{(v)}) = \widetilde{R}^{(v)}, \tag{53}$$

which is Eq. (17). Thus the ϵ_κ formulation is exactly equivalent to a view-specific bounded concentration block, but applies the same maximum mean-resultant level across heterogeneous dimensions m_v. Because $\widetilde{R}^{(v)} < 1$, the root is finite and unique. Numerically, we evaluate the Bessel ratio with scaled special functions, expand a positive upper bracket until $A_{m_v}(\kappa) \geq \widetilde{R}^{(v)}$, and then use safeguarded bisection or Newton refinement; no common raw $\kappa_{\max}$ is required.

C INITIALIZATION

Before MAP-EM, independent spherical K-means provides initial view-local components. We align only the C component labels, form a provisional cross-view global-semantic posterior, and derive leave-one-view-out overlaps for initializing the transition parameters.

C.1 INDEPENDENT LOCAL GEOMETRY

For each view v, spherical K-means with exactly C components is run on the rows of $\mathbf{U}^{(v)}$. Multiple random starts are used and the solution with the largest within-cluster cosine objective is retained. Let $h_i^{(v)} \in [C]$ denote the resulting hard local component index. The corresponding initial mean directions are the normalized hard-cluster resultants,

$$\boldsymbol{\mu}_k^{(v,0)} = \frac{\sum_{i:h_i^{(v)}=k} \mathbf{u}_i^{(v)}}{\left\|\sum_{i:h_i^{(v)}=k} \mathbf{u}_i^{(v)}\right\|_2}, \tag{54}$$

whenever the numerator is nonzero. Empty or zero-resultant components are resolved by the same deterministic restart/tie rule used by the implementation rather than by consulting ground-truth labels. The initial shared concentration for view v is estimated from these hard-cluster resultants. Define

$$\bar{R}^{(v,0)} := \frac{1}{N}\sum_{k=1}^{C}\left\|\sum_{i:h_i^{(v)}=k} \mathbf{u}_i^{(v)}\right\|_2, \qquad \widetilde{R}^{(v,0)} := \min\{\bar{R}^{(v,0)}, 1 - \epsilon_\kappa\}. \tag{55}$$

The initial concentration is then obtained by the same capped mean-resultant solve used in the MAP update:

$$A_{m_v}\left(\kappa^{(v,0)}\right) = \widetilde{R}^{(v,0)}. \tag{56}$$

Thus $\kappa^{(v,0)}$ is the unique finite solution determined by the initial spherical K-means partition.

The hard partition is then converted to a row-stochastic soft local-evidence matrix $\mathbf{H}^{(v)} \in \mathbb{R}_+^{N\times C}$ by the local vMF scores

$$H_{ik}^{(v)} = \frac{\exp\Big(\kappa^{(v,0)}(\boldsymbol{\mu}_k^{(v,0)})^\top \mathbf{u}_i^{(v)}\Big)}{\sum_{\ell=1}^{C}\exp\Big(\kappa^{(v,0)}(\boldsymbol{\mu}_\ell^{(v,0)})^\top \mathbf{u}_i^{(v)}\Big)}. \tag{57}$$

Thus $\sum_k H_{ik}^{(v)} = 1$. These soft scores are provisional view-local semantic evidence, not final global assignments.

C.2 Class-level label calibration

The numbering of independently fitted local components is arbitrary. To compare two views u and v, form the $C \times C$ soft-overlap matrix

$$\mathbf{M}^{(uv)} = \mathbf{H}^{(u)\top}\mathbf{H}^{(v)}, \qquad M_{ck}^{(uv)} = \sum_{i=1}^{N} H_{ic}^{(u)} H_{ik}^{(v)}. \tag{58}$$

Because both $\mathbf{H}^{(u)}$ and $\mathbf{H}^{(v)}$ are row-stochastic,

$$\sum_{c,k} M_{ck}^{(uv)} = \sum_i \left(\sum_c H_{ic}^{(u)}\right)\left(\sum_k H_{ik}^{(v)}\right) = N. \tag{59}$$

Let $\mathcal{P}_C$ denote the set of $C \times C$ permutation matrices. For $u \neq v$, we measure the best component-level agreement by

$$s_{uv} = \frac{1}{N}\max_{\mathbf{P}\in\mathcal{P}_C} \operatorname{tr}\Big(\mathbf{M}^{(uv)}\mathbf{P}\Big). \tag{60}$$

The maximization is a $C \times C$ linear assignment problem and is solved by the Hungarian algorithm. Define the mean cross-view agreement score and the medoid view by

$$a^{(v)} := \frac{1}{V-1}\sum_{u\neq v} s_{uv}, \qquad v_\star \in \arg\max_v a^{(v)}. \tag{61}$$

For each $v \neq v_\star$, choose

$$\mathbf{P}^{(v)} \in \arg\max_{\mathbf{P}\in\mathcal{P}_C} \operatorname{tr}\Big(\mathbf{M}^{(v_\star v)}\mathbf{P}\Big), \tag{62}$$

and set $\bar{\mathbf{H}}^{(v)} = \mathbf{H}^{(v)}\mathbf{P}^{(v)}$; for the medoid, $\mathbf{P}^{(v_\star)} = \mathbf{I}$ and $\bar{\mathbf{H}}^{(v_\star)} = \mathbf{H}^{(v_\star)}$. This operation relabels only the C local semantic components. It does not align anchors, alter sample rows, or force sample-level semantic agreement. Its purpose is solely to establish the component-index convention under which diagonal versus off-diagonal transitions are interpretable.

C.3 Global-semantic initialization and leave-one-view-out transition counts

Use the mean cross-view agreement scores from Eq. (61) only for initialization,

$$\alpha^{(v)} := \frac{a^{(v)}}{\sum_{u=1}^{V} a^{(u)}}. \tag{63}$$

Since every aligned $\bar{\mathbf{H}}^{(v)}$ is row-stochastic, their convex combination

$$\mathbf{Q}^{(0)} = \sum_{v=1}^{V} \alpha^{(v)}\bar{\mathbf{H}}^{(v)} \tag{64}$$

is also row-stochastic and serves only as a provisional global-semantic posterior for initialization. The initial class proportions are therefore

$$\pi_c^{(0)} = \frac{1}{N}\sum_i Q_{ic}^{(0)}. \tag{65}$$

These $\alpha^{(v)}$ are not learned view weights in the CAST likelihood and disappear after initialization.

To avoid allowing view v to certify its own diagonal transition structure, construct a leave-one-view-out (LOO) initializer from the remaining views,

$$\mathbf{Q}_{-v}^{(0)} = \frac{\sum_{u \neq v} \alpha^{(u)} \bar{\mathbf{H}}^{(u)}}{\sum_{u \neq v} \alpha^{(u)}}, \qquad V \geq 2, \tag{66}$$

with a uniform average over $u \neq v$ as a numerical fallback if the denominator vanishes. The soft global-to-local overlap count matrix is then

$$\mathbf{N}^{(v,0)} = (\mathbf{Q}_{-v}^{(0)})^{\top} \bar{\mathbf{H}}^{(v)}, \qquad N_{ck}^{(v,0)} = \sum_i Q_{-v,ic}^{(0)} \bar{H}_{ik}^{(v)}. \tag{67}$$

The row total and off-diagonal total are

$$n_c^{(v,0)} := \sum_k N_{ck}^{(v,0)}, \qquad d_c^{(v,0)} := \sum_{k \neq c} N_{ck}^{(v,0)}. \tag{68}$$

They are initialization overlaps rather than exact EM sufficient statistics. Applying the same MAP forms as Eq. (15) gives

$$\rho_c^{(v,0)} = \frac{d_c^{(v,0)} + a_\rho - 1}{n_c^{(v,0)} + a_\rho + b_\rho - 2}, \tag{69}$$

and, once the fixed directional profile below is specified,

$$S_{ck}^{(v,0)} = \frac{N_{ck}^{(v,0)} + \gamma_S \omega_{ck}^{(v)}}{d_c^{(v,0)} + \gamma_S}, \qquad k \neq c, \quad S_{cc}^{(v,0)} = 0. \tag{70}$$

The directional prior profile is constructed from pooled off-diagonal destination mass. Define

$$\zeta_k^{(v)} := \sum_{c \neq k} N_{ck}^{(v,0)}. \tag{71}$$

We then set

$$\omega_{ck}^{(v)} = \frac{\zeta_k^{(v)}}{\sum_{j \neq c} \zeta_j^{(v)}}, \qquad k \neq c, \qquad \omega_{cc}^{(v)} = 0. \tag{72}$$

If the denominator vanishes, the corresponding off-diagonal profile is taken to be uniform; hence $\boldsymbol{\omega}_{-c}^{(v)} \in \Delta^{C-2}$.

D ALGORITHM OVERVIEW

Algorithm 1 CAST: MAP-EM on independent anchor assignments

Require: Row-stochastic $\{\mathbf{Z}^{(v)}\}_{v=1}^{V}$, class count C, fixed weak priors.

1: Embed each assignment row by Eq. (1), stack the rows as $\mathbf{U}^{(v)}$, and initialize $\boldsymbol{\pi}$, the local emissions, transition parameters, and fixed prior profiles.
2: **repeat**
3: Evaluate local vMF emissions and calibrated $\psi_c^{(v)}$ in the log domain.
4: Compute q_{ic} and $\xi_{ick}^{(v)}$ via Eqs. (12)–(13).
5: Update π_c, $\rho_c^{(v)}$, and $S_{ck}^{(v)}$ by Eqs. (14)–(15).
6: Update $\boldsymbol{\mu}_k^{(v)}$ and solve ϵ_κ-capped $\kappa^{(v)}$ by Eqs. (16)–(17).
7: Monitor the fixed-prior MAP objective $\mathcal{J}$.
8: **until** the relative objective change is below tolerance.
9: **return** $\hat{y}_i = \arg\max_c q_{ic}$, $\{\mathbf{T}^{(v)}\}_{v=1}^{V}$, and $\{\tau_i^{(v)}\}_{i,v}$.

E Proofs of Theoretical Results

E.1 Proof of Proposition 4.1

For fixed sample i and view v, define

$$a_i^{(v)} := \sum_{j=1}^{C} f_{ij}^{(v)}, \qquad b_{ik}^{(v)} := \frac{f_{ik}^{(v)}}{a_i^{(v)}}, \qquad \mathbf{b}_i^{(v)} := (b_{i1}^{(v)}, \dots, b_{iC}^{(v)})^\top. \tag{73}$$

The vMF working densities are strictly positive on the ambient sphere, hence $a_i^{(v)} > 0$, and $\sum_k b_{ik}^{(v)} = 1$. Using Eq. (5),

$$\psi_{ic}^{(v)} = \sum_{k=1}^{C} T_{ck}^{(v)} f_{ik}^{(v)} = a_i^{(v)} \sum_{k=1}^{C} T_{ck}^{(v)} b_{ik}^{(v)} = a_i^{(v)} [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_c. \tag{74}$$

Substituting this expression into Eq. (12) yields

$$\begin{aligned} q_{ic} &= \frac{\pi_c \prod_v \left(a_i^{(v)} [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_c \right)}{\sum_{c'} \pi_{c'} \prod_v \left(a_i^{(v)} [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_{c'} \right)} \\ &= \frac{\pi_c \left(\prod_v a_i^{(v)} \right) \prod_v [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_c}{\left(\prod_v a_i^{(v)} \right) \sum_{c'} \pi_{c'} \prod_v [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_{c'}} \\ &= \frac{\pi_c \prod_v [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_c}{\sum_{c'} \pi_{c'} \prod_v [\mathbf{T}^{(v)} \mathbf{b}_i^{(v)}]_{c'}}. \end{aligned} \tag{75}$$

where the common factor $\prod_v a_i^{(v)}$ is independent of the candidate global class c and therefore can be factored out from the denominator and canceled during posterior normalization. The two invariance statements in Proposition 4.1 act at different levels.

Anchor-coordinate permutation invariance. Let $\mathbf{P}^{(v)}$ be an $m_v \times m_v$ permutation matrix on the anchor coordinates. Since a permutation matrix satisfies $(\mathbf{P}^{(v)})^\top \mathbf{P}^{(v)} = \mathbf{I}_{m_v}$, it is orthogonal. With the semantic-state index k fixed, jointly relabel $\widetilde{\mathbf{u}} = \mathbf{P}^{(v)} \mathbf{u}$ and $\widetilde{\boldsymbol{\mu}}_k^{(v)} = \mathbf{P}^{(v)} \boldsymbol{\mu}_k^{(v)}$. Their directional score is unchanged:

$$(\widetilde{\boldsymbol{\mu}}_k^{(v)})^\top \widetilde{\mathbf{u}} = (\mathbf{P}^{(v)} \boldsymbol{\mu}_k^{(v)})^\top (\mathbf{P}^{(v)} \mathbf{u}) = (\boldsymbol{\mu}_k^{(v)})^\top (\mathbf{P}^{(v)})^\top \mathbf{P}^{(v)} \mathbf{u} = (\boldsymbol{\mu}_k^{(v)})^\top \mathbf{u}. \tag{76}$$

The dimension m_v, concentration $\kappa^{(v)}$, and vMF normalizer are unchanged, so the numerical value of every local emission $f_{ik}^{(v)}$ is exactly preserved under $\mathbf{P}^{(v)}$. Consequently $\psi_{ic}^{(v)}$ and q_{ic} are unchanged as well. Independent coordinate permutations may be applied to different views. This invariance is only to coordinate relabeling; it does not cover changes in anchor locations, anchor counts, or the upstream anchor-construction rule.

Likelihood-scale invariance and sufficiency. Now allow the local emissions themselves to change by any common positive sample–view scale, $\widetilde{f}_{ik}^{(v)} = \lambda_i^{(v)} f_{ik}^{(v)}$ with $\lambda_i^{(v)} > 0$ independent of k. The corresponding normalization factor becomes $\widetilde{a}_i^{(v)} = \sum_j \widetilde{f}_{ij}^{(v)} = \lambda_i^{(v)} a_i^{(v)}$, and therefore $\widetilde{b}_{ik}^{(v)} = \widetilde{f}_{ik}^{(v)} / \widetilde{a}_i^{(v)} = b_{ik}^{(v)}$. Consequently, the resulting global posterior remains unchanged, i.e., $\widetilde{q}_{ic} = q_{ic}$. Thus the global posterior is insensitive to the absolute scale of a view's local likelihood vector even though the individual $f_{ik}^{(v)}$ values may change. For fixed fitted parameters, its contribution to global inference is completely summarized by the relative C-state evidence $\mathbf{b}_i^{(v)}$. This is the predictive sufficiency claimed here; it is not classical parameter sufficiency for $\boldsymbol{\Theta}$.

E.2 Proof of Proposition 4.2

Fix a view and suppress the superscript (v) for readability. Let

$$\boldsymbol{\delta} := \mathbf{T}_{c:} - \mathbf{T}_{c':}. \tag{77}$$

By Eq. (5),

$$\psi_c(\mathbf{u}) - \psi_{c'}(\mathbf{u}) = \sum_{k=1}^{C} \delta_k f_k(\mathbf{u}). \tag{78}$$

Therefore

$$\begin{aligned} \|\psi_c - \psi_{c'}\|_{L_2}^2 &= \int \left(\sum_k \delta_k f_k(\mathbf{u})\right)\left(\sum_\ell \delta_\ell f_\ell(\mathbf{u})\right) d\sigma(\mathbf{u}) \\ &= \sum_{k,\ell} \delta_k \delta_\ell \int f_k(\mathbf{u}) f_\ell(\mathbf{u}) d\sigma(\mathbf{u}) \\ &= \boldsymbol{\delta} \mathbf{G} \boldsymbol{\delta}^\top. \end{aligned} \tag{79}$$

For the equal-concentration vMF case,

$$f_k(\mathbf{u}) f_\ell(\mathbf{u}) = \mathcal{C}_m(\kappa)^2 \exp\left(\kappa \boldsymbol{\mu}_k^\top \mathbf{u} + \kappa \boldsymbol{\mu}_\ell^\top \mathbf{u}\right) = \mathcal{C}_m(\kappa)^2 \exp\left(\kappa(\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell)^\top \mathbf{u}\right). \tag{80}$$

Let

$$\mathbf{a} := \kappa(\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell), \qquad \|\mathbf{a}\|_2 = \kappa \|\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell\|_2. \tag{81}$$

By the defining normalization identity of the vMF family,

$$\int_{\mathbb{S}^{m-1}} \exp(\mathbf{a}^\top \mathbf{u}) d\sigma(\mathbf{u}) = \frac{1}{\mathcal{C}_m(\|\mathbf{a}\|_2)}, \tag{82}$$

with the continuous extension at $\|\mathbf{a}\|_2 = 0$. Combining Eqs. (80)–(82) gives

$$G_{k\ell} = \frac{\mathcal{C}_m(\kappa)^2}{\mathcal{C}_m(\kappa \|\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell\|_2)}, \tag{83}$$

Finally, for any coefficient vector $\mathbf{a} \in \mathbb{R}^C$,

$$\mathbf{a}^\top \mathbf{G} \mathbf{a} = \sum_{k,\ell} a_k a_\ell \int f_k f_\ell d\sigma = \int \left(\sum_k a_k f_k(\mathbf{u})\right)^2 d\sigma(\mathbf{u}) \geq 0. \tag{84}$$

Thus $\mathbf{G}$ is always positive semidefinite. It is positive definite exactly when no nonzero coefficient vector produces an L_2-zero linear combination of the emissions, i.e., exactly when $\{f_k\}_{k=1}^C$ is linearly independent in $L_2(d\sigma)$. Under that condition, distinct transition rows have $\boldsymbol{\delta}_{cc'}^{(v)} \neq \mathbf{0}$ and therefore induce distinct effective class-conditional densities.

The identity $\|\psi_c^{(v)} - \psi_{c'}^{(v)}\|_{L_2}^2 = \boldsymbol{\delta}_{cc'}^{(v)} \mathbf{G}^{(v)} (\boldsymbol{\delta}_{cc'}^{(v)})^\top$ also clarifies the role of the transition matrix. The matrix $\mathbf{G}^{(v)}$ describes the intrinsic geometry of the view-local semantic states, whereas $\mathbf{T}^{(v)}$ determines how global classes are embedded into this local semantic geometry. Consequently, a semantic shift represented by a non-identity transition does not imply information destruction. If the emission components remain distinguishable and the transition rows are different, the induced global-class conditional densities remain distinguishable. This is the sense in which CAST calibrates semantic transitions rather than merely averaging or smoothing view-specific evidence.

E.3 Proof of Proposition 4.3

Fix view v. The joint law of the global class and view-local semantic state is

$$\Pr(Y_i = c, L_i^{(v)} = k) = \pi_c T_{ck}^{(v)}. \tag{85}$$

The marginal probability of local state k is therefore

$$p_k^{(v)} := \Pr(L_i^{(v)} = k) = \sum_{c=1}^{C} \pi_c T_{ck}^{(v)}. \tag{86}$$

Substituting these two quantities into the definition of mutual information gives

$$\begin{aligned} I(Y_i; L_i^{(v)}) &= \sum_{c,k} \Pr(Y_i = c, L_i^{(v)} = k) \log \frac{\Pr(Y_i = c, L_i^{(v)} = k)}{\Pr(Y_i = c)\Pr(L_i^{(v)} = k)} \\ &= \sum_{c,k} \pi_c T_{ck}^{(v)} \log \frac{\pi_c T_{ck}^{(v)}}{\pi_c p_k^{(v)}} = \sum_{c,k} \pi_c T_{ck}^{(v)} \log \frac{T_{ck}^{(v)}}{p_k^{(v)}}, \end{aligned} \tag{87}$$

Equivalently, the retained semantic information can be written as

$$I(Y_i; L_i^{(v)}) = \sum_c \pi_c D_{\mathrm{KL}}\left(\mathbf{T}_{c:}^{(v)} \| \mathbf{p}^{(v)}\right), \tag{88}$$

where $\mathbf{p}^{(v)} = \sum_c \pi_c \mathbf{T}_{c:}^{(v)}$. This form highlights that informativeness is determined by the distinguishability of transition profiles and their statistical support under the class prior distribution. Terms with zero joint mass are interpreted by the standard convention $0 \log 0 = 0$.

Under the working model, $\mathbf{u}_i^{(v)}$ is generated from $L_i^{(v)}$ and has no additional dependence on Y_i once $L_i^{(v)}$ is given. Hence

$$Y_i \longrightarrow L_i^{(v)} \longrightarrow \mathbf{u}_i^{(v)} \tag{89}$$

is a Markov chain. The data-processing inequality then gives

$$I(Y_i; \mathbf{u}_i^{(v)}) \le I(Y_i; L_i^{(v)}). \tag{90}$$

This upper bound separates the information that can in principle pass through the semantic channel from the information retained by the particular emission geometry.

If every row of $\mathbf{T}^{(v)}$ equals the same probability vector $\mathbf{t}$, then $T_{ck}^{(v)} = t_k$ for all c, so $p_k^{(v)} = t_k$ and every nonzero logarithmic ratio above equals one. Thus $I(Y_i; L_i^{(v)}) = 0$. The same fact is visible directly in the effective density:

$$\psi_c^{(v)}(\mathbf{u}) = \sum_k t_k f_k^{(v)}(\mathbf{u}), \tag{91}$$

which no longer depends on c. In Eq. (12), the factor from view v is then identical for all candidate global classes and cancels from the posterior normalization.

Conversely, if $\mathbf{T}^{(v)}$ is a permutation matrix, then there exists a permutation σ such that $L^{(v)} = \sigma(Y)$ deterministically. Because σ is invertible,

$$H(L_i^{(v)} \mid Y_i) = 0, \qquad H(Y_i \mid L_i^{(v)}) = 0, \tag{92}$$

and therefore

$$I(Y_i; L_i^{(v)}) = H(Y_i). \tag{93}$$

A derangement permutation may have zero diagonal while still preserving all class information. This is why diagonal preservation and informativeness are not equivalent notions.

E.4 Proof of Theorem 4.4

For the joint observation $\mathcal{U} := \{\mathbf{U}^{(v)}\}_{v=1}^V$, define the class-conditional density

$$p_c(\mathcal{U}) := p(\mathcal{U} \mid Y = c) = \prod_{v=1}^V \psi_c^{(v)}(\mathbf{u}^{(v)}) \tag{94}$$

under the conditional-independence assumption. Define

$$a_c(\mathcal{U}) := \pi_c p_c(\mathcal{U}). \tag{95}$$

The Bayes classifier selects a class attaining $\max_c a_c(\mathcal{U})$. Since $\sum_c \int a_c(\mathcal{U}) d\mathcal{U} = 1$, the Bayes error can be written as

$$R^\star = \int \left[\sum_c a_c(\mathcal{U}) - \max_c a_c(\mathcal{U})\right] d\mathcal{U}. \tag{96}$$

For a fixed $\mathcal{U}$, let j be any maximizing class. Then $a_c \le a_j$ for every $c \neq j$, and therefore

$$a_c \le \sqrt{a_c a_j}, \qquad c \neq j. \tag{97}$$

Consequently,

$$\sum_c a_c - \max_c a_c = \sum_{c \neq j} a_c \le \sum_{c \neq j} \sqrt{a_c a_j} \le \sum_{c<c'} \sqrt{a_c a_{c'}}. \tag{98}$$

Integrating Eq. (98) and using nonnegativity yields

$$R^\star \le \sum_{c<c'} \sqrt{\pi_c \pi_{c'}} \int \sqrt{p_c(\mathcal{U}) p_{c'}(\mathcal{U})} d\mathcal{U}. \tag{99}$$

By Eq. (94), the integrand factorizes across views. Tonelli's theorem therefore gives

$$\begin{aligned} \int \sqrt{p_c(\mathcal{U}) p_{c'}(\mathcal{U})} d\mathcal{U} &= \int \prod_v \sqrt{\psi_c^{(v)}(\mathbf{u}^{(v)}) \psi_{c'}^{(v)}(\mathbf{u}^{(v)})} \prod_v d\sigma_v(\mathbf{u}^{(v)}) \\ &= \prod_v \int \sqrt{\psi_c^{(v)}(\mathbf{u}) \psi_{c'}^{(v)}(\mathbf{u})} d\sigma_v(\mathbf{u}) \\ &= \prod_v B_{cc'}^{(v)}. \end{aligned} \tag{100}$$

Substituting Eq. (100) into Eq. (99) gives the multi-view Bhattacharyya error bound. The product structure follows from the working conditional-independence model; it is not a guarantee that adding an empirically estimated view must improve a finite-sample clustering result.

E.5 Proof of Theorem 4.5

Let

$$q^t(\mathbf{Y}, \mathbf{L}) := p(\mathbf{Y}, \mathbf{L} \mid \mathcal{U}, \mathbf{\Theta}^t) \tag{101}$$

be the exact E-step posterior at iteration t. For any normalized distribution $q(\mathbf{Y}, \mathbf{L})$ over the latent variables, define

$$\mathcal{F}(q, \mathbf{\Theta}) := \mathbb{E}_q[\log p(\mathcal{U}, \mathbf{Y}, \mathbf{L} \mid \mathbf{\Theta})] + \mathcal{H}(q) + \log p(\mathbf{\Theta}), \tag{102}$$

where

$$\mathcal{H}(q) := -\mathbb{E}_q[\log q(\mathbf{Y}, \mathbf{L})] \tag{103}$$

and $\log p(\mathbf{\Theta})$ denotes the fixed Beta–Dirichlet prior terms, including their normalizing constants if desired. The posterior identity

$$\log p(\mathbf{Y}, \mathbf{L} \mid \mathcal{U}, \mathbf{\Theta}) = \log p(\mathcal{U}, \mathbf{Y}, \mathbf{L} \mid \mathbf{\Theta}) - \log p(\mathcal{U} \mid \mathbf{\Theta}) \tag{104}$$

implies

$$\mathrm{KL}(q \,\|\, p(\mathbf{Y}, \mathbf{L} \mid \mathcal{U}, \mathbf{\Theta})) = \mathbb{E}_q[\log q] - \mathbb{E}_q[\log p(\mathcal{U}, \mathbf{Y}, \mathbf{L} \mid \mathbf{\Theta})] + \log p(\mathcal{U} \mid \mathbf{\Theta}). \tag{105}$$

Rearranging and adding the log prior gives the exact decomposition

$$\mathcal{J}(\mathbf{\Theta}) = \mathcal{F}(q, \mathbf{\Theta}) + \mathrm{KL}(q \,\|\, p(\mathbf{Y}, \mathbf{L} \mid \mathcal{U}, \mathbf{\Theta})), \tag{106}$$

where $\mathcal{J}(\mathbf{\Theta}) = \log p(\mathcal{U} \mid \mathbf{\Theta}) + \log p(\mathbf{\Theta})$, up to any fixed prior constants omitted consistently in both expressions.

At $q = q^t$ and $\mathbf{\Theta} = \mathbf{\Theta}^t$, the KL term in Eq. (106) is exactly zero, so

$$\mathcal{J}(\mathbf{\Theta}^t) = \mathcal{F}(q^t, \mathbf{\Theta}^t). \tag{107}$$

For any candidate $\mathbf{\Theta}^{t+1}$, nonnegativity of KL gives

$$\mathcal{J}(\mathbf{\Theta}^{t+1}) \ge \mathcal{F}(q^t, \mathbf{\Theta}^{t+1}). \tag{108}$$

Therefore, if the M-step is exact or merely satisfies

$$\mathcal{F}(q^t, \mathbf{\Theta}^{t+1}) \ge \mathcal{F}(q^t, \mathbf{\Theta}^t), \tag{109}$$

then Eqs. (107)–(109) imply

$$\mathcal{J}(\mathbf{\Theta}^{t+1}) \geq \mathcal{F}(q^t, \mathbf{\Theta}^{t+1}) \geq \mathcal{F}(q^t, \mathbf{\Theta}^t) = \mathcal{J}(\mathbf{\Theta}^t). \tag{110}$$

This is the MAP analogue of the standard EM ascent argument.

For CAST, Appendix B shows that $\boldsymbol{\pi}$, each $\rho_c^{(v)}$, each off-diagonal row $\mathbf{S}_{-c}^{(v)}$, and the mean directions are exact maximizers of their mutually separable blocks in the MAP auxiliary objective. After optimizing the directions, the concentration profile is

$$N\left[\log \mathcal{C}_{m_v}(\kappa) + \kappa \bar{R}^{(v)}\right]. \tag{111}$$

With fixed $\epsilon_\kappa \in (0,1)$, define $K_v(\epsilon_\kappa) = A_{m_v}^{-1}(1-\epsilon_\kappa)$. Appendix B.7 shows that Eq. (17) is exactly the maximizer on the compact interval $[0, K_v(\epsilon_\kappa)]$: it is the unconstrained root when $\bar{R}^{(v)} \leq 1-\epsilon_\kappa$ and the upper boundary otherwise. Solving the capped-target scalar equation to numerical tolerance therefore implements the exact constrained block update up to that tolerance.

The monotonicity result alone does not imply convergence unless the objective is also bounded above. For any unit vectors $\boldsymbol{\mu}$ and $\mathbf{u}$,

$$\boldsymbol{\mu}^\top \mathbf{u} \leq 1. \tag{112}$$

Hence, for every local emission in view v,

$$f_k^{(v)}(\mathbf{u}) = \mathcal{C}_{m_v}(\kappa^{(v)}) \exp\left(\kappa^{(v)} (\boldsymbol{\mu}_k^{(v)})^\top \mathbf{u}\right) \leq \mathcal{C}_{m_v}(\kappa^{(v)}) e^{\kappa^{(v)}}. \tag{113}$$

Using the continuous extension of the vMF normalizer at $\kappa = 0$, define

$$M_v := \max_{0 \leq \kappa \leq K_v(\epsilon_\kappa)} \mathcal{C}_{m_v}(\kappa) e^{\kappa}. \tag{114}$$

The maximand is continuous on a compact interval, so $M_v < \infty$. Therefore

$$f_k^{(v)}(\mathbf{u}) \leq M_v \quad \text{for all } k, \mathbf{u}. \tag{115}$$

Because $\mathbf{T}_{c:}^{(v)}$ is row-stochastic, $\psi_c^{(v)}$ is a convex combination of the local emissions and satisfies

$$\psi_c^{(v)}(\mathbf{u}) = \sum_k T_{ck}^{(v)} f_k^{(v)}(\mathbf{u}) \leq \sum_k T_{ck}^{(v)} M_v = M_v. \tag{116}$$

For one sample,

$$p(\mathcal{U}_i \mid \mathbf{\Theta}) = \sum_c \pi_c \prod_v \psi_c^{(v)}(\mathbf{u}_i^{(v)}) \leq \sum_c \pi_c \prod_v M_v = \prod_v M_v. \tag{117}$$

Thus

$$\log p(\mathcal{U} \mid \mathbf{\Theta}) \leq N \sum_{v=1}^{V} \log M_v. \tag{118}$$

The prior terms are also bounded above for the fixed hyperparameters used by CAST. For $a_\rho, b_\rho > 1$, the Beta contribution

$$(a_\rho - 1)\log \rho + (b_\rho - 1)\log(1-\rho) \tag{119}$$

has a finite interior maximum on $[0,1]$ and tends to $-\infty$ at either endpoint. For one directional row, the parameter-dependent Dirichlet contribution is

$$\gamma_S \sum_{k \neq c} \omega_{ck}^{(v)} \log S_{ck}^{(v)} \leq 0, \tag{120}$$

because $0 \leq S_{ck}^{(v)} \leq 1$ and every coefficient is nonnegative; when a positive coefficient multiplies $S_{ck}^{(v)} = 0$, the contribution is $-\infty$, never $+\infty$. Fixed prior normalizing constants merely add a finite constant. Consequently, there exists a finite $J_{\max}$ such that

$$\mathcal{J}(\mathbf{\Theta}) \leq J_{\max} \tag{121}$$

throughout the feasible parameter set. Combining this upper bound with Eq. (110), the monotone scalar sequence $\{\mathcal{J}(\mathbf{\Theta}^t)\}$ converges to a finite limit. The fixed tolerance $\epsilon_\kappa > 0$ is essential to this

direct compactness argument because it induces the finite view-specific bounds $K_v(\epsilon_\kappa)$. Without any concentration control (equivalently, as $\epsilon_\kappa \downarrow 0$), a directional mixture may approach arbitrarily sharp densities, and a separate argument would be required to rule out an unbounded likelihood sequence.

A stronger stationary-point statement requires additional regularity. Let Ω denote the constrained parameter space and consider the exact-update sequence. Assume:

1. the iterates remain in a compact subset of the finite-objective region of Ω;
2. the posterior $p(\mathbf{Y}, \mathbf{L} \mid \mathcal{U}, \boldsymbol{\Theta})$ and the MAP auxiliary objective are continuous on that set, and $\mathcal{J}$ is continuously differentiable on each relevant smooth feasible stratum;
3. every active local component has nonzero resultant, each ϵ_κ-bounded concentration block is solved to its global maximizer, and the M-step update map is closed, with a fixed admissible tie rule for nonunique blocks; and
4. a standard constraint qualification holds at each accumulation point so that first-order feasible-direction optimality is equivalent to the corresponding KKT conditions.

These assumptions explicitly exclude the degenerate cases in which a component has zero occupancy/resultant or a parameter lies on a nonsmooth change of active constraints without a compatible limiting argument.

The key local identity is the EM score/tangency relation. Fix a parameter value $\boldsymbol{\Theta}^\star$ and set

$$q^\star(\mathbf{Y}, \mathbf{L}) = p(\mathbf{Y}, \mathbf{L} \mid \mathcal{U}, \boldsymbol{\Theta}^\star). \tag{122}$$

Differentiating the observed log likelihood under the finite latent sum gives

$$\nabla_{\boldsymbol{\Theta}} \log p(\mathcal{U} \mid \boldsymbol{\Theta})\big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^\star} = \sum_{\mathbf{Y},\mathbf{L}} q^\star(\mathbf{Y}, \mathbf{L}) \nabla_{\boldsymbol{\Theta}} \log p(\mathcal{U}, \mathbf{Y}, \mathbf{L} \mid \boldsymbol{\Theta})\big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^\star}. \tag{123}$$

Adding the prior gradient yields

$$\nabla \mathcal{J}(\boldsymbol{\Theta}^\star) = \nabla_{\boldsymbol{\Theta}} \mathcal{F}(q^\star, \boldsymbol{\Theta})\big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}^\star}, \tag{124}$$

because the entropy of the fixed distribution $q^\star$ does not depend on $\boldsymbol{\Theta}$. Eq. (124) is the first-order form of the tangency used by EM convergence arguments.

Now let $\boldsymbol{\Theta}^{t_j} \to \boldsymbol{\Theta}^\star$ be any convergent subsequence. Closedness and continuity of the exact update map imply that any corresponding limit of $\boldsymbol{\Theta}^{t_j+1}$ is an admissible blockwise maximizer of $\mathcal{F}(q^\star, \cdot)$. Suppose, for contradiction, that $\boldsymbol{\Theta}^\star$ is not constrained stationary. By the constraint qualification, there then exists a feasible direction with strictly positive directional derivative of $\mathcal{J}$. By Eq. (124), the same direction has strictly positive derivative for the auxiliary objective with $q^\star$ fixed. Continuity therefore provides a neighborhood of $\boldsymbol{\Theta}^\star$ in which an admissible update can improve the auxiliary objective by a positive amount bounded away from zero. Exact block maximization would achieve at least that improvement. Repeated sufficiently late visits of the convergent subsequence to this neighborhood would then force a fixed positive increase in $\mathcal{J}$, contradicting convergence of the objective values. Hence every accumulation point is constrained stationary. Under the stated constraint qualification, this is equivalent to the first-order/KKT conditions on Ω.

This argument is consistent with the classical EM convergence framework of Wu (1983). The conclusion is deliberately limited: it establishes stationarity of accumulation points under the additional assumptions above, not convergence of the entire parameter sequence, uniqueness of the stationary point, or global optimality. Label-switching symmetries can also generate multiple equivalent stationary representations.

F Detailed experimental protocols

F.1 Datasets

The evaluated datasets cover diverse scales, with the number of views ranging from 3 to 6, the number of samples from 210 to 126,054, and the number of classes from 7 to 100. Detailed statistical information is shown in Table 2.

Table 2: Dataset descriptions

Dataset	#View	#Instance	#Class	#Feature
EuroSAT	4	18900	10	512, 512, 512, 768
MSRC	6	210	7	1302, 48, 512, 100, 256, 210
RSICB128	4	25696	45	512, 512, 512, 768
YTF10	4	38654	10	944, 576, 512, 640
YTF20	4	63896	20	944, 576, 512, 640
YTF50	4	126054	50	944, 576, 512, 640
CIFAR10	3	50000	10	512, 2048, 1024
CIFAR100	3	50000	100	512, 2048, 1024

F.2 Controlled semantic-shift setup

Let $\mathbf{a}_c^{(v)}$ denote the original class-centroid vector in feature space and let $\phi^{(v)}(c) \neq c$ be a target class, chosen by nearest competing class unless otherwise specified. For each source class, select a fixed random ordering of its observations. At requested prevalence δ, use the first $\mathrm{round}(\delta n_c)$ observations; this makes selected subsets nested as δ increases. The same generation seed is shared across all clustering methods.

Euclidean semantic relocation. For selected i with $Y_i = c$,

$$\widetilde{\mathbf{x}}_i^{(v)} = \mathbf{x}_i^{(v)} + \eta\big(\mathbf{a}_{\phi^{(v)}(c)}^{(v)} - \mathbf{a}_c^{(v)}\big). \tag{125}$$

All selected samples in the same source/view subgroup receive the same displacement. Hence $\widetilde{\mathbf{x}}_i^{(v)} - \widetilde{\mathbf{x}}_j^{(v)} = \mathbf{x}_i^{(v)} - \mathbf{x}_j^{(v)}$ for selected i, j from that subgroup, preserving pairwise Euclidean geometry and centered covariance.

Spherical semantic rotation. When samples are ℓ_2 normalized, define class mean directions

$$\mathbf{e}_c^{(v)} = \frac{\sum_{i:Y_i=c} \mathbf{x}_i^{(v)}}{\|\sum_{i:Y_i=c} \mathbf{x}_i^{(v)}\|_2}. \tag{126}$$

Within view v, define the source and target directions and their angular separation by

$$\mathbf{u}^{(v)} = \mathbf{e}_c^{(v)}, \qquad \mathbf{w}^{(v)} = \mathbf{e}_{\phi^{(v)}(c)}^{(v)}, \qquad \theta^{(v)} = \arccos\Big((\mathbf{u}^{(v)})^\mathsf{T}\mathbf{w}^{(v)}\Big).$$

The in-plane orthogonal direction is

$$\mathbf{b}^{(v)} = \frac{\mathbf{w}^{(v)} - ((\mathbf{u}^{(v)})^\mathsf{T}\mathbf{w}^{(v)})\mathbf{u}^{(v)}}{\|\mathbf{w}^{(v)} - ((\mathbf{u}^{(v)})^\mathsf{T}\mathbf{w}^{(v)})\mathbf{u}^{(v)}\|_2}.$$

Rotate $\mathrm{span}\{\mathbf{u}^{(v)}, \mathbf{b}^{(v)}\}$ by $\varphi^{(v)} = \eta\theta^{(v)}$ and leave the orthogonal complement unchanged. The resulting matrix $\mathbf{R}^{(v)}$ is orthogonal, so

$$\|\mathbf{R}^{(v)}\mathbf{x}_i^{(v)}\|_2 = \|\mathbf{x}_i^{(v)}\|_2, \qquad (\mathbf{R}^{(v)}\mathbf{x}_i^{(v)})^\mathsf{T}(\mathbf{R}^{(v)}\mathbf{x}_j^{(v)}) = (\mathbf{x}_i^{(v)})^\mathsf{T}\mathbf{x}_j^{(v)} \tag{127}$$

for all selected i, j receiving the same rotation in view v.

F.3 SUPPLEMENTARY EXPERIMENTAL RESULTS

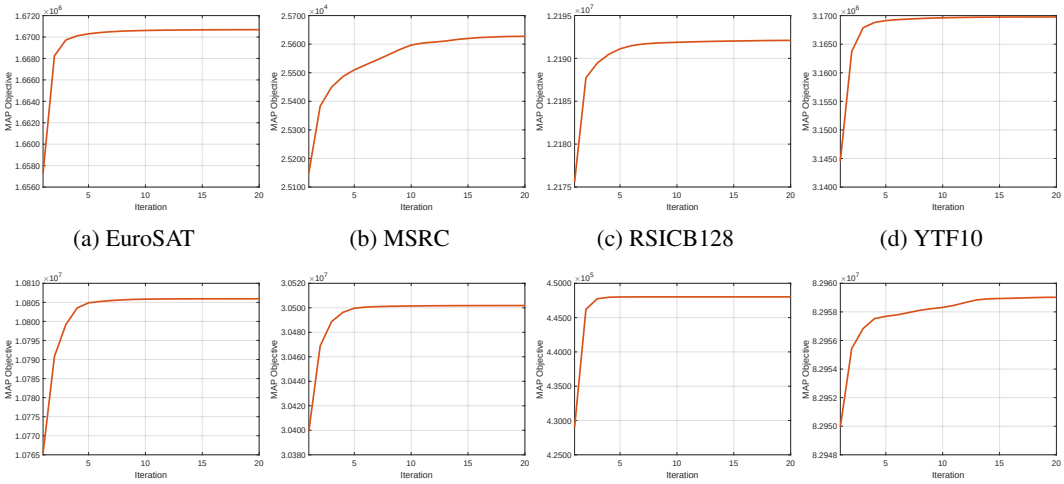
Table 3: Ablation study under the original setting and two semantic shifts. ACC, NMI, and ARI (%, higher is better) are grouped by dataset. Within each setting, the best result in each column is bold and the second-best distinct result is underlined; ties share the same mark. CAST uses prior strengths $\gamma_S = 0$ or $\gamma_S = 2$.

Setting	Method	EuroSAT			MSRC-v6			RSICB128			YTF10			YTF20			YTF50			CIFAR10			CIFAR100		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Ori.	Identity	**90.3**	**81.2**	**79.8**	72.4	72.6	63.7	**79.6**	**89.4**	**79.2**	79.7	80.3	68.2	**78.8**	**82.1**	**68.3**	76.4	84.2	63.0	**99.4**	**98.3**	**98.7**	**98.0**	**99.4**	**98.0**
	ViewLevel	88.3	78.3	76.0	91.0	83.6	79.6	77.7	86.4	76.7	85.3	81.7	73.3	73.9	77.3	62.0	75.2	84.4	66.6	99.1	97.5	98.1	96.0	98.6	96.0
	Classwise	88.5	79.1	76.5	94.8	89.9	87.6	77.2	86.9	75.5	85.2	82.0	72.9	74.1	77.4	59.4	74.8	84.0	65.1	99.1	97.6	98.1	**98.0**	**99.4**	**98.0**
	SharedDest	89.0	80.2	77.5	94.3	88.4	86.7	79.0	88.0	78.2	85.6	81.0	72.6	73.7	78.1	60.6	76.3	84.5	65.7	99.1	97.5	98.0	**98.0**	**99.4**	**98.0**
	CAST$_{(\gamma_S=0)}$	89.7	81.0	79.2	94.8	88.9	88.1	79.2	88.4	78.6	**88.4**	**84.3**	**76.8**	76.5	80.7	65.9	**76.6**	**86.0**	**70.3**	99.1	97.5	98.0	**98.0**	**99.4**	**98.0**
	CAST$_{(\gamma_S=2)}$	89.8	81.0	79.3	**95.7**	**91.0**	**90.1**	79.2	88.4	78.6	**88.4**	84.2	76.7	76.5	80.7	65.9	**76.6**	**86.0**	**70.3**	99.1	97.5	98.0	97.0	99.1	97.0
Euc.	Identity	71.7	64.0	54.7	45.2	33.0	21.4	65.7	80.2	62.1	74.6	79.4	69.2	77.4	83.2	70.8	72.8	84.1	66.2	96.9	92.5	93.3	95.5	**99.1**	**96.2**
	ViewLevel	74.0	70.2	62.2	43.3	38.6	23.3	57.7	66.9	46.8	78.9	78.7	69.4	69.6	75.4	59.6	73.7	82.5	66.1	92.7	84.2	84.8	94.1	98.4	94.5
	Classwise	55.3	52.5	34.3	45.2	41.2	25.0	56.5	67.5	44.2	75.0	75.9	64.3	64.1	70.1	49.5	73.1	81.6	64.1	59.3	59.8	38.4	88.7	95.1	83.9
	SharedDest	69.0	66.9	57.3	46.7	39.0	24.8	59.7	69.4	49.5	77.2	80.3	70.7	72.4	77.2	58.2	73.6	82.5	65.8	84.7	79.7	71.0	91.0	96.1	87.9
	CAST$_{(\gamma_S=0)}$	82.1	81.9	76.0	**79.5**	**66.6**	**60.6**	81.1	90.7	**82.6**	87.4	92.9	89.3	82.9	88.1	78.6	82.7	90.7	78.8	**98.0**	**94.9**	**95.6**	95.5	98.7	95.8
	CAST$_{(\gamma_S=2)}$	**93.8**	**86.8**	**87.0**	77.1	66.3	57.1	**83.3**	**91.6**	81.9	**97.6**	**94.9**	**95.4**	**84.9**	**89.5**	**82.3**	**82.9**	**90.8**	**79.7**	**98.0**	**94.9**	**95.6**	**96.0**	98.8	96.1
Sph.	Identity	71.8	64.2	55.2	55.7	40.2	26.9	73.4	85.7	71.6	89.6	87.5	80.6	75.4	79.3	65.4	72.6	83.9	64.3	96.7	92.0	92.8	94.2	98.7	95.1
	ViewLevel	76.4	68.8	62.3	50.5	36.2	20.7	73.4	82.9	68.8	86.2	81.0	75.4	75.1	79.2	67.6	75.8	83.3	68.1	92.7	84.1	84.7	94.1	98.4	94.3
	Classwise	69.2	60.9	50.4	50.0	38.3	21.5	71.1	82.0	65.6	79.3	73.9	63.7	70.2	74.7	56.9	73.7	81.8	64.9	59.7	59.6	38.6	86.8	94.1	80.9
	SharedDest	78.4	69.0	64.2	50.5	36.8	25.2	72.7	83.0	68.1	78.9	75.2	67.1	74.7	80.3	68.5	74.1	82.2	66.3	84.6	79.4	70.8	90.6	96.7	89.1
	CAST$_{(\gamma_S=0)}$	**93.3**	**85.9**	**85.9**	72.4	**67.6**	**61.4**	**82.6**	**92.7**	**84.8**	**96.9**	**93.7**	**93.7**	**86.6**	**91.8**	**85.8**	**84.3**	**92.1**	**83.1**	**97.8**	**94.4**	**95.2**	**95.3**	98.7	**95.6**
	CAST$_{(\gamma_S=2)}$	**93.3**	85.8	85.8	**75.7**	63.9	55.7	81.0	**92.7**	83.4	**96.9**	**93.7**	**93.7**	86.0	91.5	85.2	84.1	91.7	82.4	**97.8**	**94.4**	**95.2**	**95.3**	**98.8**	**95.6**

Table 4: Effect of heterogeneous anchor cardinalities across views.

Dataset	ACC ↑			NMI ↑			ARI ↑		
	CAST	CAST-H	Gain	CAST	CAST-H	Gain	CAST	CAST-H	Gain
EuroSAT	89.77	90.37	+0.60	81.01	81.64	+0.63	79.26	80.26	+1.00
MSRC-v6	95.71	95.71	+0.00	90.97	90.97	+0.00	90.12	90.12	+0.00
RSICB128	79.16	80.57	+1.41	88.40	90.17	+1.76	78.64	78.73	+0.09
YTF10	88.41	91.06	+2.64	84.18	87.33	+3.15	76.74	84.86	+8.12
YTF20	76.49	79.76	+3.27	80.66	83.15	+2.49	65.88	72.71	+6.83
YTF50	76.56	78.18	+1.62	85.98	86.71	+0.74	70.32	72.66	+2.33
CIFAR10	99.07	99.07	+0.00	97.48	97.48	+0.00	97.97	97.97	+0.00
CIFAR100	96.97	97.77	+0.80	99.05	99.17	+0.12	96.99	97.57	+0.57

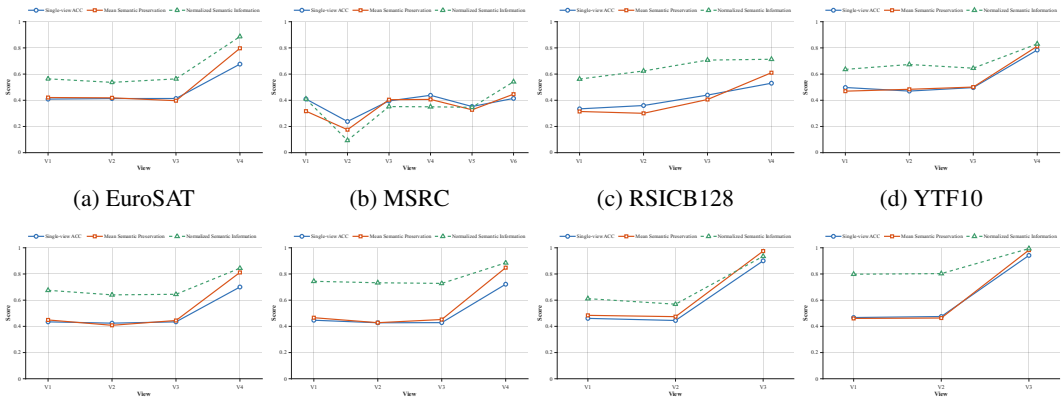
Note: Values are percentages. CAST uses the same number of anchors in all views; CAST-H allows view-specific anchor numbers. Gain is CAST-H minus CAST in percentage points, computed before rounding.



(a) EuroSAT (b) MSRC (c) RSICB128 (d) YTF10

(e) YTF20 (f) YTF50 (g) CIFAR10 (h) CIFAR100

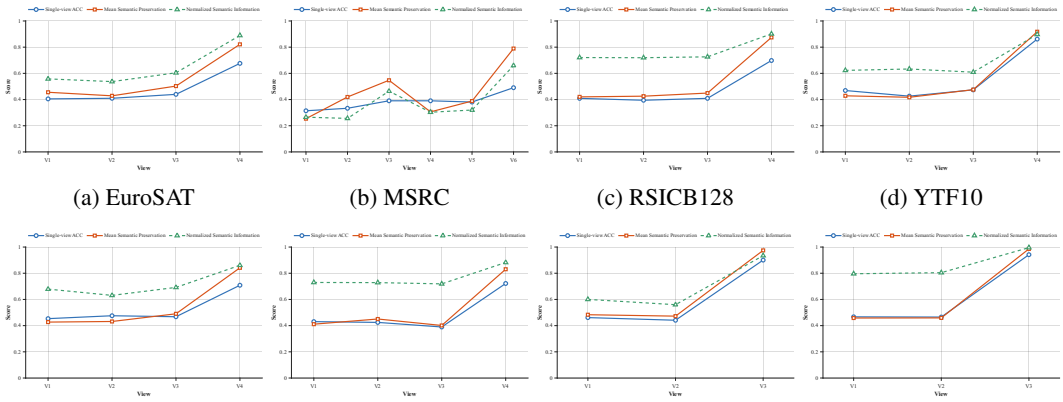
Figure 1: MAP objective curves of our algorithm on benchmark datasets.



(a) EuroSAT (b) MSRC (c) RSICB128 (d) YTF10

(e) YTF20 (f) YTF50 (g) CIFAR10 (h) CIFAR100

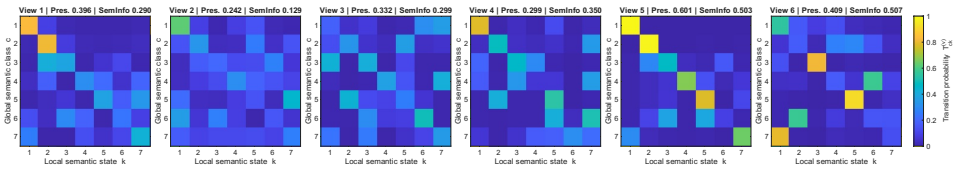
Figure 2: View-wise semantic diagnostics under Euclidean relocation.



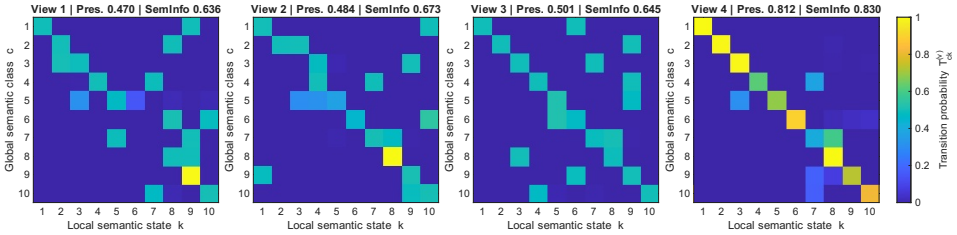
(a) EuroSAT (b) MSRC (c) RSICB128 (d) YTF10

(e) YTF20 (f) YTF50 (g) CIFAR10 (h) CIFAR100

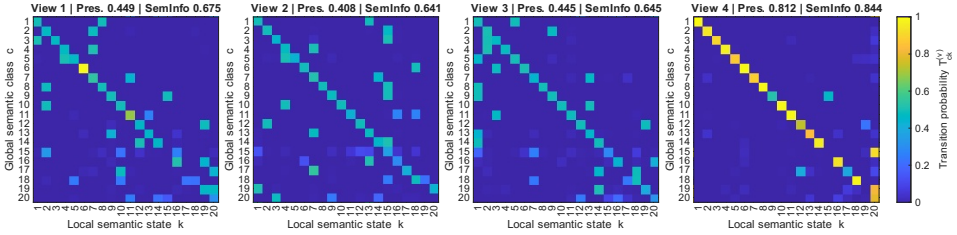
Figure 3: View-wise semantic diagnostics under spherical semantic rotation.



(a) MSRC

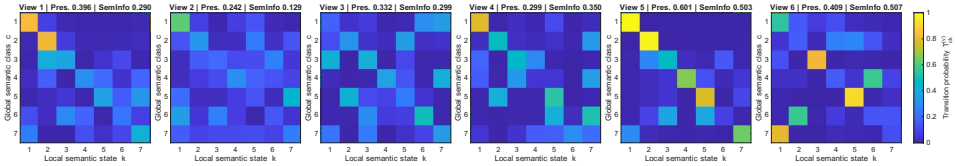


(b) YTF10

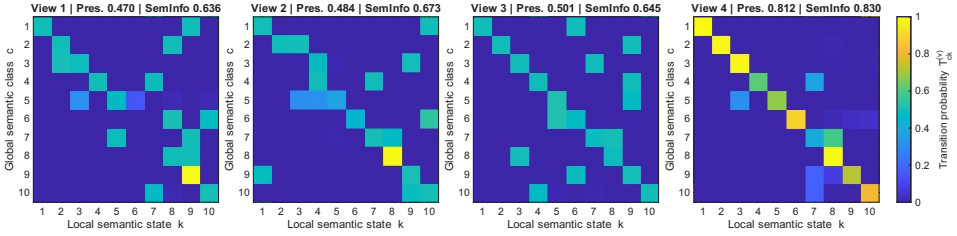


(c) YTF20

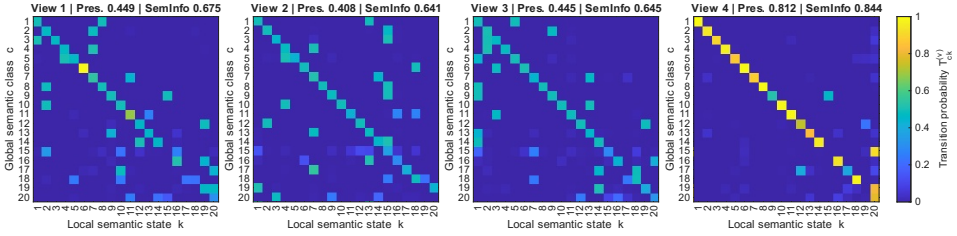
Figure 4: Visualization of class-level semantic transition matrices under Euclidean relocation.



(a) MSRC



(b) YTF10



(c) YTF20

Figure 5: Visualization of class-level semantic transition matrices under spherical rotation.