CROSS-DOMAIN MICRO-ACTION LEARNING VIA INTER-SPACE ALIGNMENT AND PROTOTYPE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prototype distances learned independently on heterogeneous datasets are defined relative to dataset-specific geometries, so the same class-relative score need not carry the same meaning across datasets. We study *cross-dataset anchor generality*: whether one fixed class-reference system can provide a common class-reference coordinate for SMG and iMiGUE under a prespecified binary polarity ontology while retaining recognition utility. This is a reference-comparability problem rather than a prototype-storage or unseen-domain generalization claim. We realize it through three coupled requirements: a domain-shared class geometry is constructed, frozen source features are aligned to that reference geometry, and residual sample-level deviations are calibrated with the same fixed anchors. Cross-domain Class-conditional Shared Geometry (CCSG), a soft-gated source-to-reference mapper, and Prototype-Guided Anchor-Pair Calibration (PG-APC) implement these roles. On SMG, dataset-specific PG-APC obtains $69.00 \pm 2.79\%/68.69 \pm 3.12\%$ Accuracy/Macro-F1, while CCSG-PG-APC obtains $68.00 \pm 4.31\%/67.65 \pm 4.48\%$. On iMiGUE, the same five seeds fixed a priori give $70.38 \pm 1.58\%/69.73 \pm 2.18\%$ for CCSG-PG-APC and $68.65 \pm 1.46\%/68.09 \pm 1.92\%$ for PG-APC, a descriptive $+1.73$-point Accuracy gap. Naive pooling also degrades substantially, consistent with reference-geometry mismatch.

1 INTRODUCTION

Class prototypes are meaningful only relative to the feature geometry in which their distances are measured. Equal-dimensional representations from heterogeneous datasets therefore do not imply interchangeable class references: in our controlled anchor-swap experiment, substituting the other dataset's anchors yields only 61.33% Accuracy on SMG and 59.81% on iMiGUE. This motivates *cross-dataset anchor generality*: whether, under a prespecified shared polarity ontology, one fixed class-reference system can be reused across observed datasets while preserving the interpretation and recognition utility of class-relative distances. The object being shared is thus not two anchor vectors, but a reference coordinate in which the sign and magnitude of $\|f - t_0\|_2 - \|f - t_1\|_2$ admit the same polarity interpretation. We retain dataset-specific downstream predictors to isolate reference reuse from joint-classifier training; no unseen-domain generalization is claimed. The exact label mapping is given in Sec. 5.1.

Achieving this goal exposes three coupled geometric mismatches. First, samples assigned to the same shared emotion state in SMG and iMiGUE can follow different class-conditional distributions, so averaging prototypes can place a reference between incompatible modes rather than define a shared class-reference center. Second, the transported CCSG reference changes the first- and second-order geometry seen by the frozen features; sharing the ambient dimension D does not make their reference statistics interchangeable. Third, after distribution-level alignment, individual samples can remain locally inconsistent with the shared anchors. We therefore operationalize anchor generality through *reference consistency*, *source-to-reference consistency*, and *prototype-relative consistency*. These requirements are induced by the shared-reference objective, rather than introduced as interchangeable post-hoc modules.

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.

Existing skeleton- and RGB-based methods mainly strengthen upstream video representations through structural, temporal, and trajectory-aware modeling Yan et al. (2018); Liu et al. (2020); Li et al. (2024); Ho & Madden (2025); Vandeghen et al. (2026). Our focus is complementary: the encoder remains frozen, and we ask whether two heterogeneous datasets can use one reusable class-reference system rather than two independently estimated prototype systems. Directly averaging independent prototypes can be inadequate when their class-conditional geometries differ.

We therefore construct Cross-domain Class-conditional Shared Geometry (CCSG) from the two training domains, use a supervised train-only soft-gated *source-to-reference mapper* to match frozen-feature statistics to the learned CCSG reference, and apply Prototype-Guided Anchor-Pair Calibration (PG-APC) against the resulting shared anchor pair. Figure 1 labels this block "Cross-Space Alignment" to mean source-to-reference alignment between different reference geometries: CCSG and the frozen features occupy the same ambient $\mathbb{R}^D$ rather than two independently pretrained embedding spaces. CCSG establishes the shared reference, the mapper makes its class-conditional statistics accessible from the source representation, and PG-APC constrains local correction to the same fixed anchors. Validation and test samples are excluded from barycenter construction, mapper statistics, gate fitting, and anchor estimation.

Our main contributions are:

- We formulate heterogeneous micro-action learning as a cross-dataset anchor-generality problem in which one fixed reference must preserve the class-relative interpretation of distances across heterogeneous observed datasets. This separates reference comparability from ordinary accuracy improvement, prototype compression, and unseen-domain generalization.
- We develop a reference-consistent geometric pipeline in which CCSG constructs the common class geometry and a soft-gated source-to-reference mapper matches frozen TrackMAE features to that geometry. The two stages are coupled by one reference rather than acting as independent feature adapters.
- PG-APC completes the same reference system by performing sample-adaptive calibration only with the fixed CCSG anchors. Repeated-run comparisons and fixed-pipeline diagnostics then measure both the recognition cost of replacing dataset-specific reference systems and the failure of naive pooling.

2 Related Context and Preliminaries

Transport-based shared reference. For class-k domain distributions $\mathcal{P}_{d,k}$, a common reference can be defined by the Wasserstein barycenter

$$\mathcal{P}^{\star}_{C,k} = \arg\min_{\mathcal{Q}} \sum_d \omega_d W_2^2(\mathcal{Q}, \mathcal{P}_{d,k}), \quad \sum_d \omega_d = 1.$$

Schrödinger-bridge methods further connect shifted distributions through noisy interpolants such as $\xi_t = (1-t)\xi_0 + t\xi_1 + \sigma\sqrt{t(1-t)}\,\zeta, \zeta \sim \mathcal{N}(0, I)$, with learned forward/reverse transport fields Zheng et al. (2026); Mishra et al. (2026). CCSG adopts this class-conditionally so that the transported root, rather than a pooled mean, becomes the shared reference.

Source-to-reference moment alignment. Given source/reference statistics (μ_X, Σ_X) and (μ_C, Σ_C), a whitening–coloring transform has the form

$$T(x) = (x - \mu_X)\Sigma_X^{-1/2}\Sigma_C^{1/2} + \mu_C.$$

This provides a compact second-order alignment template for heterogeneous representations Chen et al. (2025); Jalali et al. (2025); Xu et al. (2025); our mapper makes it class-conditional and routes samples without test labels.

Prototype-relative prediction. Prototype methods convert distances to class references into posterior weights Snell et al. (2017); Wen et al. (2016); Schroff et al. (2015), e.g.,

$$p(y = k \mid f) = \frac{\exp[-d(f, t_k)/\tau]}{\sum_j \exp[-d(f, t_j)/\tau]}.$$

The posterior is only comparable when the reference geometry is comparable. Our formulation therefore couples these three ingredients: a shared transported reference, source-to-reference alignment, and calibration against the same fixed anchors.

Relation to prototype-based adaptation. Prototype and metric-learning methods use class references for prediction or embedding shaping Snell et al. (2017); Wen et al. (2016); Zhou et al. (2023); Ren et al. (2024). Here the issue is different: independently estimated references across heterogeneous datasets need not be comparable. We therefore study fixed-reference comparability; CCSG and the mapper establish the common reference and PG-APC provides local calibration.

3 Problem Formulation

Let $d \in \{\mathrm{S}, \mathrm{I}\}$ denote SMG or iMiGUE, $y_i \in \{0, 1\}$ the shared negative/positive polarity, and E a frozen video encoder. For sample $v_i^{(d)}$,

$$x_i^{(d)} = E\Big(v_i^{(d)}\Big) \in \mathbb{R}^D, \tag{1}$$

and $\mathcal{P}_{d,k}$ denotes the class-k training-feature distribution. Our target is a single class-reference system $\mathcal{T} = \{t_0, t_1\}$ that is estimated from the training domains and then reused unchanged by both dataset branches. CCSG constructs reference features $z_j^C \in \mathbb{R}^D$ from which $\mathcal{T}$ is formed. Although source and CCSG features share the same ambient coordinates, their class-conditional statistics need not match; the source-to-reference mapper therefore produces

$$f_i = T_{\mathrm{soft}}\Big(x_i^{(d)}\Big) \in \mathbb{R}^D, \tag{2}$$

after which PG-APC yields the calibrated feature $\tilde{f}_i$ using fixed CCSG anchors. All geometric statistics, mapper parameters, and anchors are estimated from training data only.

What anchor generality requires. Dataset-specific anchors can be locally meaningful yet cross-dataset incomparable. We fix the same $\mathcal{T}$, polarity convention, and downstream use across observed domains, aligning only source features. Anchor generality is measured by the recognition cost of replacing native reference systems; direct exchange and naive pooling are negative controls, with no shared classifier or statistical equivalence assumed.

4 Methodology

The methodology follows the three consistency requirements induced by anchor generality. CCSG enforces *reference consistency* by constructing one class-conditional geometry from both training domains. The soft-gated source-to-reference mapper enforces *source-to-reference consistency* by matching frozen features to that reference using training-derived statistics, without validation/test labels for routing. PG-APC enforces *prototype-relative consistency* by restricting sample-level correction to the same fixed CCSG anchors. The stages therefore share one geometric object and serve distinct roles rather than forming a sequence of unrelated adapters.

Why the stages are coupled. CCSG first defines the shared class geometry, the mapper then matches frozen features to that reference, and PG-APC restricts local correction to its anchors. Direct pooling omits the first requirement, while downstream calibration cannot repair an inconsistent reference; Supplementary ablations test these simplifications.

4.1 Class-Conditional Cross-domain CCSG Construction

Barycentric tree-structured bidirectional bridge. For domain $d \in \{\mathrm{S}, \mathrm{I}\}$ and class $k \in \{0, 1\}$, let $\widehat{\mathcal{P}}_{d,k}$ denote the empirical distribution of class-k training features. We first pool the same-class SMG and iMiGUE training features to initialize the support of a common reference. The class-conditional root is then computed as a weighted Wasserstein barycenter

$$\mathcal{P}_{C,k}^{(0)} = \arg\min_{\mathcal{Q}} \sum_{d \in \{\mathrm{S},\mathrm{I}\}} \omega_d W_2^2\Big(\mathcal{Q}, \widehat{\mathcal{P}}_{d,k}\Big), \tag{3}$$

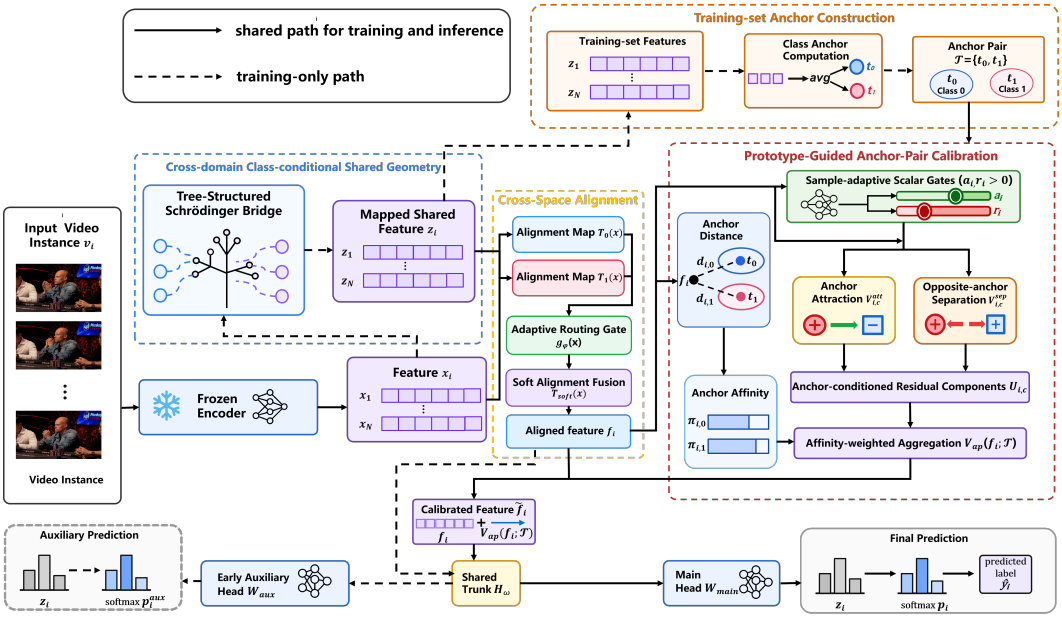


Figure 1: Overall framework. CCSG builds the shared reference, the source-to-reference mapper aligns frozen features, and PG-APC calibrates them with the same fixed anchors.

where $\omega_d \geq 0$, $\sum_d \omega_d = 1$, are fixed domain weights and only training samples are used. For each class, the barycenter is the root of a three-node topology whose SMG and iMiGUE distributions are connected to it by independent bidirectional bridges; no direct SMG–iMiGUE edge is used. We write $\mathcal{P}_{d,k} = \widehat{\mathcal{P}}_{d,k}$ and $\mathcal{P}_{C,k}$ for the corresponding child and root distributions; additional CCSG settings are summarized in Supplementary Sec. A. With $\xi_0 \sim \mathcal{P}_{d,k}$ and $\xi_1 \sim \mathcal{P}_{C,k}$, the bridge state is

$$\xi_t = (1-t)\xi_0 + t\xi_1 + \sigma_{d,k}\sqrt{t(1-t)}\,\zeta. \tag{4}$$

Here $t \in [0,1]$ is bridge time, $\sigma_{d,k} > 0$ is the noise scale, $\zeta \sim \mathcal{N}(0, I_D)$ is Gaussian noise, and I_D is the $D \times D$ identity matrix.

Let $v^+_{\theta^+}$ and $v^-_{\theta^-}$ be forward/reverse velocity fields shared across domain/class groups, with separate parameters and targets $u^+ = (\xi_1 - \xi_t)/(1 - t + \varepsilon)$ and $u^- = (\xi_0 - \xi_t)/(t + \varepsilon)$. We use the bidirectional bridge-matching objective

$$\mathcal{L}_{\text{bridge}} = \frac{1}{2} \sum_{d \in \{\text{S,I}\}} \sum_{k=0}^{1} \mathbb{E}\left[\|v^+_{\theta^+}(\xi_t, t) - u^+\|_2^2 + \|v^-_{\theta^-}(\xi_t, 1-t) - u^-\|_2^2\right]. \tag{5}$$

Here $\varepsilon > 0$ stabilizes the targets, and the learned forward field is integrated with R Euler steps. Supplementary Sec. A summarizes the reported CCSG settings.

Class-conditional geometry constraints. Let $Z^C_{\text{S},k}$ and $Z^C_{\text{I},k}$ be class-k CCSG features, $Z^C_k = Z^C_{\text{S},k} \cup Z^C_{\text{I},k}$, and $\mu^C_k = |Z^C_k|^{-1} \sum_{z \in Z^C_k} z$. We use the RBF kernel $\kappa_h(u,v) = \exp(-\|u - v\|_2^2/(2h^2))$.

$$\mathcal{D}^{(k)}_{\text{CS}} = \log \mathbb{E}_{u,u' \sim Z^C_{\text{S},k}} \kappa_h(u,u') + \log \mathbb{E}_{v,v' \sim Z^C_{\text{I},k}} \kappa_h(v,v') - 2 \log \mathbb{E}_{u \sim Z^C_{\text{S},k},\, v \sim Z^C_{\text{I},k}} \kappa_h(u,v). \tag{6}$$

Its class average $\mathcal{L}_{\text{CCS}} = \frac{1}{2}\sum_{k=0}^{1} \mathcal{D}^{(k)}_{\text{CS}}$ is the explicit same-class cross-domain alignment loss.

$$\mathcal{C}_{\text{intra}} = \frac{1}{2} \sum_{k=0}^{1} \frac{|Z^C_k|^{-1} \sum_{z \in Z^C_k} \|z - \mu^C_k\|_2^2}{|Z^C_k|^{-1} \sum_{z \in Z^C_k} \|z\|_2^2 + \varepsilon}. \tag{7}$$

This normalized dispersion is used only as an auxiliary compactness stabilizer (weight λ_{comp}; Supplementary Sec. A), not as a core loss.

$$\mathcal{L}_{\text{sep}} = -\operatorname{clip}\left(\frac{\|\mu_1^C - \mu_0^C\|_2^2}{\frac{1}{2}(\|\mu_0^C\|_2^2 + \|\mu_1^C\|_2^2) + \varepsilon}, 0, \delta_{\max}\right). \tag{8}$$

Minimizing $\mathcal{L}_{\text{sep}}$ separates the class centers; clipping prevents unbounded expansion.

The core CCSG objective therefore has the bridge term and two explicit geometric losses:

$$\mathcal{L}_{\text{CCSG}} = \mathcal{L}_{\text{bridge}} + \lambda_{\text{CCS}}\mathcal{L}_{\text{CCS}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}}. \tag{9}$$

Reported training additionally applies $\lambda_{\text{comp}}\mathcal{C}_{\text{intra}}$ as the auxiliary stabilizer. Only $\mathcal{L}_{\text{bridge}}$ is active during the initial η_{warm} fraction; afterwards all terms are enabled and geometry statistics are refreshed every K_{geo} steps.

4.2 Soft-Gated Source-to-Reference Mapper

Class-conditional affine maps. For class k, let $(\mu_{X,k}, \Sigma_{X,k})$ and $(\mu_{C,k}, \Sigma_{C,k})$ be source- and CCSG-reference mean/covariance pairs, respectively, estimated from training features over the two observed domains. For $A \in \{X, C\}$, regularize the covariance as

$$\bar{\Sigma}_{A,k} = (1-\alpha)\Sigma_{A,k} + \alpha \operatorname{Diag}(\Sigma_{A,k}) + \rho_{A,k} I_D. \tag{10}$$

Here A indexes source (X) or CCSG (C) space, $\bar{\Sigma}_{A,k}$ is the regularized covariance, $\operatorname{Diag}(\Sigma_{A,k})$ keeps diagonal entries, $\alpha \in [0,1]$ is the shrinkage coefficient, and $\rho_{A,k} > 0$ is the ridge coefficient.

The class-k source-to-CCSG transform is

$$T_k(x) = (x - \mu_{X,k})\bar{\Sigma}_{X,k}^{-\frac{1}{2}}\bar{\Sigma}_{C,k}^{\frac{1}{2}} + \mu_{C,k}. \tag{11}$$

Here $x \in \mathbb{R}^D$ is a frozen source feature; $\bar{\Sigma}_{X,k}^{-1/2}$ whitens source covariance and $\bar{\Sigma}_{C,k}^{1/2}$ restores CCSG covariance geometry. Thus T_k is a class-conditional moment transform inside the same ambient feature space, not a mapping between separately pretrained encoders.

Soft routing. Standardize x as $\tilde{x} = (x - \mu_{\text{train}})/(\sigma_{\text{train}} + \varepsilon)$, where μ_{train} and σ_{train} are coordinate-wise training statistics. A binary gate with parameters $\phi = \{W, b\}$ uses

$$g_\phi(\tilde{x}) = W\tilde{x} + b. \tag{12}$$

Here g_ϕ is the gate, with weight W and bias b. Its routing weight is

$$p_k(x) = \operatorname{softmax}\left(\frac{g_\phi(\tilde{x})}{\tau_g}\right)_k. \tag{13}$$

Here $p_k(x)$ weights T_k, $\tau_g > 0$ is the routing temperature, and $p_0(x) + p_1(x) = 1$.

The gate is fitted on the training split with class supervision:

$$\mathcal{L}_{\text{gate}} = -\frac{1}{N_{\text{tr}}}\sum_{i=1}^{N_{\text{tr}}}\sum_{k=0}^{1}\mathbf{1}[y_i = k]\log p_k(x_i). \tag{14}$$

After this fit, ϕ is fixed. Routing is supervised during training, but validation/test inference uses only the input feature. Importantly, g_ϕ is not the task classifier: it only mixes T_0 and T_1, no validation/test label selects a transform, and Supplementary Table 6 reports a Gate-Only control.

For sample i from domain d, the mapper output is

$$f_i = T_{\text{soft}}(x_i^{(d)}) = \sum_{k=0}^{1} p_k(x_i^{(d)})T_k(x_i^{(d)}). \tag{15}$$

Here T_{soft} denotes the soft-gated mapper and f_i its CCSG-reference output; $x_i^{(d)}$ is defined in the problem formulation.

The fixed class-k CCSG anchor is

$$t_k = \frac{1}{N_{C,k}} \sum_{j:\, y_j = k} z_j^C. \tag{16}$$

Here z_j^C is a class-k CCSG training feature, $N_{C,k}$ its sample count, t_k the anchor, and $\mathcal{T} = \{t_0, t_1\}$ the anchor pair. The index j ranges over transported training features from both SMG and iMiGUE; hence this equation produces one unique pair rather than one pair per dataset. The same $\mathcal{T}$ is fixed and used by both downstream domain branches.

4.3 Prototype-Guided Anchor-Pair Calibration

Anchor-conditioned calibration. For mapped feature f_i, let $d_{i,k} = \|f_i - t_k\|_2$ be the distance to anchor t_k. The affinity is

$$\pi_{i,k} = \frac{\exp(-d_{i,k}/\tau_p)}{\sum_{\ell=0}^{1} \exp(-d_{i,\ell}/\tau_p)}. \tag{17}$$

Here $\pi_{i,k}$ is the anchor affinity, $\tau_p > 0$ its temperature, and ℓ the summation class index.

Two lightweight scalar-gating branches predict sample-adaptive strengths,

$$a_i = \mathrm{Softplus}(g_{\theta_a}(f_i)), \qquad r_i = \mathrm{Softplus}(g_{\theta_r}(f_i)). \tag{18}$$

Here g_{θ_a} and g_{θ_r} are the attraction and separation gates, while $a_i \geq 0$ and $r_i \geq 0$ are their non-negative magnitudes.

For class hypothesis k, let $\bar{k} = 1 - k$ be the opposite class. The attraction displacement is

$$V_{i,k}^{\mathrm{att}} = a_i(t_k - f_i). \tag{19}$$

The separation displacement is

$$V_{i,k}^{\mathrm{sep}} = r_i[\mu_{\mathrm{sep}} - d_{i,\bar{k}}]_+ \frac{f_i - t_{\bar{k}}}{\|f_i - t_{\bar{k}}\|_2 + \varepsilon}. \tag{20}$$

Here $V_{i,k}^{\mathrm{att}}$ and $V_{i,k}^{\mathrm{sep}}$ denote attraction and separation displacements, $\mu_{\mathrm{sep}} > 0$ is the separation radius, and $[a]_+ = \max(0, a)$ is the positive-part operator.

The affinity-weighted residual is

$$V_{\mathrm{ap}}(f_i; \mathcal{T}) = \sum_{k=0}^{1} \pi_{i,k} \left(V_{i,k}^{\mathrm{att}} + V_{i,k}^{\mathrm{sep}} \right). \tag{21}$$

Here $V_{\mathrm{ap}}(f_i; \mathcal{T})$ is the anchor-pair residual. The calibrated feature is

$$\tilde{f}_i = f_i + V_{\mathrm{ap}}(f_i; \mathcal{T}). \tag{22}$$

Here $\tilde{f}_i$ is passed to its prediction trunk and classifier. The anchors constrain calibration directions but do not define the final prediction rule.

Training objective. Let $\mathcal{L}_{\mathrm{CE}}$ denote the classification cross-entropy used in training. Anchor ordering is regularized by

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{N} \sum_{i=1}^{N} \max\Big(0, m + \|\tilde{f}_i - t_{y_i}\|_2 - \|\tilde{f}_i - t_{1-y_i}\|_2\Big). \tag{23}$$

Here N is mini-batch size, $m > 0$ is the ranking margin, y_i is the class label defined in the problem formulation, and t_{1-y_i} is the opposite anchor.

Residual magnitude is controlled by

$$\mathcal{L}_{\mathrm{step}} = \frac{1}{ND} \sum_{i=1}^{N} \|V_{\mathrm{ap}}(f_i; \mathcal{T})\|_2^2. \tag{24}$$

Here $\mathcal{L}_{\text{step}}$ is the residual-step penalty and D is the feature dimension defined in the problem formulation.

The complete objective is

$$\mathcal{L}_{\text{PG-APC}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{AP}}\mathcal{L}_{\text{rank}} + \beta_{\text{step}}\mathcal{L}_{\text{step}} + \mathcal{L}_{\text{aux}}. \tag{25}$$

Here $\lambda_{\text{AP}} \geq 0$ weights ranking, $\beta_{\text{step}} \geq 0$ weights residual-step control, and unit-weight $\mathcal{L}_{\text{aux}}$ is auxiliary cross-entropy used only during early training; its head is removed for validation and testing.

4.4 Multiclass Extension

Let $\mathcal{K} = \{0, \ldots, K-1\}$ with $K > 2$ and let $\mathcal{D}$ denote the observed-domain set. CCSG instantiates the same class-conditional bridge for every $(d, k) \in \mathcal{D} \times \mathcal{K}$, with $\sum_{d\in\mathcal{D}} \omega_d = 1$; the explicit geometry terms are averaged over K classes (and all unordered class pairs for separation). The mapper becomes K-way:

$$p_k^{(K)}(x) = \text{softmax}\left(\frac{g_\phi^{(K)}(\tilde{x})}{\tau_g}\right)_k, \qquad f_i^{(K)} = \sum_{k\in\mathcal{K}} p_k^{(K)}(x_i^{(d)}) T_k(x_i^{(d)}), \tag{26}$$

with one shared anchor t_k per class. The binary opposite anchor is replaced by the nearest competing anchor

$$c_i(k) = \arg\min_{j\in\mathcal{K}, j\neq k} \|f_i^{(K)} - t_j\|_2, \tag{27}$$

which is used in the separation direction. The ranking objective compares the ground-truth anchor with the closest negative anchor,

$$\mathcal{L}_{\text{rank}}^{(K)} = \frac{1}{N}\sum_{i=1}^{N} \max\left(0,\, m + \|\tilde{f}_i^{(K)} - t_{y_i}\|_2 - \min_{j\in\mathcal{K}, j\neq y_i} \|\tilde{f}_i^{(K)} - t_j\|_2\right). \tag{28}$$

The affinity normalization and classifiers are extended to K classes in the same way. The complete K-class CCSG objective, residual definition, and HEROES evaluation protocol are provided in Supplementary Sec. E.

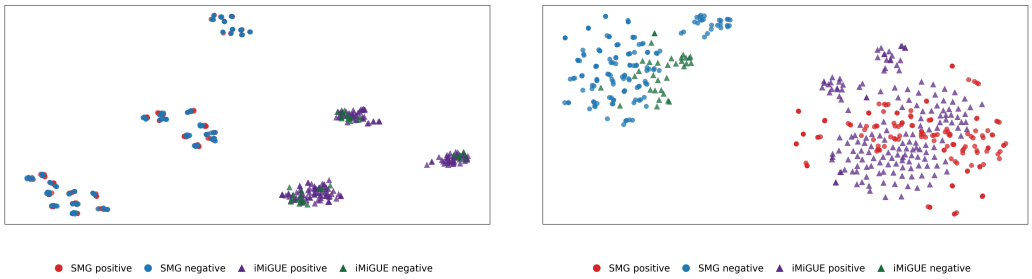
5 Experiments

5.1 Experimental Setup

Datasets and labels. SMG Chen et al. (2023) uses $30/5/5$ recordings for training/validation/testing, yielding $293/60/60$ clips after segmentation; iMiGUE Liu et al. (2021) follows its official $205/50/104$ video split. Following the binary emotion-state convention in Xia et al. (2025), we use NES/Win as positive and SES/Loss as negative. This mapping is an operational task correspondence for reference comparability, not a claim of psychological equivalence between the datasets; CCSG addresses the remaining within-polarity geometric mismatch. For HEROES Mannocchi et al. (2022), we use an actor-disjoint $10/3/3$ split ($160/48/48$ videos) only as a four-class proof-of-concept because it contains conventional body actions rather than micro-actions.

Evaluation protocol. We report Accuracy and Macro-F1. SMG and iMiGUE use the same five random seeds fixed a priori; within each seed, checkpoints are selected on the validation split and the test split is evaluated only after model selection. Component ablations are single-run mechanism diagnostics using a prespecified seed fixed independently of test performance. "Shared" means that one CCSG anchor pair is fixed and reused by both dataset branches, while downstream predictors remain dataset-specific. Hyperparameters are selected on SMG and reused unchanged on iMiGUE. For SMG and iMiGUE, TrackMAE remains frozen ($D = 768$).

Controls. Anchor controls include dataset-specific PG-APC, direct cross-dataset anchor swapping, naive pooled anchors, and CCSG-PG-APC. Adapter controls use $\tilde{f}_i^{\text{ada}} = f_i + h_\psi(f_i)$; Add. Adapter + Same Losses matches PG-APC supervision and is near parameter matched (0.2993M versus 0.3004M trainable parameters active at inference; Supplementary Table 8). AdaSP Zhou et al. (2023) and DADA Ren et al. (2024) retain their native objectives.



(a) Before CCSG (b) After CCSG

Figure 2: t-SNE visualization of cross-domain geometry before and after CCSG. After CCSG, samples with the same polarity exhibit stronger cross-domain overlap while the two polarity groups remain visually distinguishable.

Method	Acc. (%)	Macro-F1 (%)
SMG		
Add. Adapter + Same Losses	62.67 ± 2.24	62.27 ± 1.98
AdaSP	65.67 ± 0.91	65.63 ± 0.91
PG-APC (w/o CCSG)	69.00 ± 2.79	68.69 ± 3.12
CCSG-PG-APC	68.00 ± 4.31	67.65 ± 4.48
iMiGUE		
Add. Adapter + Same Losses	59.42 ± 4.05	57.63 ± 4.12
AdaSP	64.62 ± 1.58	62.71 ± 1.24
PG-APC (w/o CCSG)	68.65 ± 1.46	68.09 ± 1.92
CCSG-PG-APC	70.38 ± 1.58	69.73 ± 2.18

Table 1: Five-seed results using identical prespecified seeds and validation-selected checkpoints.

5.2 Cross-Domain Alignment Visualization

Figure 2 visualizes the same training samples before and after CCSG under identical t-SNE settings. Separate t-SNE embeddings are not coordinate-comparable, so the plots are used only for qualitative inspection.

The visualization suggests reduced same-polarity cross-domain discrepancy with preserved inter-class separation; quantitative conclusions rely on the controlled experiments below.

5.3 Cross-Dataset Reference Reuse

Table 1 contrasts dataset-specific PG-APC with the complete shared-reference pipeline under the same five fixed seeds and validation-selected checkpoints; Tables 2 and 3 provide portability and contextual evidence.

Dataset	Anchor pair used	Acc. (%)	Macro-F1 (%)
SMG	iMiGUE-specific	61.33 ± 1.39	59.78 ± 2.16
iMiGUE	SMG-specific	59.81 ± 0.43	56.60 ± 1.45
SMG + CCSG	CCSG shared	68.00 ± 4.31	67.65 ± 4.48
iMiGUE + CCSG	CCSG shared	70.38 ± 1.58	69.73 ± 2.18

Table 2: Anchor portability under the same five fixed seeds as Table 1; the first two rows directly swap dataset-specific anchors.

Anchor portability. Table 2 shows that direct swapping reduces SMG from $69.00/68.69$ to $61.33/59.78$ ($-7.67/-8.91$ points) and reaches $59.81/56.60$ on iMiGUE, versus $68.65/68.09$ for

Group	Method	SMG		iMiGUE	
		Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
RGB	TC-CLIP	55.00	54.89	63.46	61.41
	Video-FocalNet	61.67	60.33	64.42	59.27
	ViTTA	63.33	63.29	63.46	61.03
Skeleton	GCN-NAS	63.33	61.81	56.73	46.76
	BlockGCN	65.00	64.99	58.65	49.13
	PoseC3D	60.00	59.60	55.77	41.03
	Hyperformer	66.67	66.63	58.65	51.05
	H2OFormer	70.00	68.27	65.38	63.08
Adapter	Training-Only Adapter	61.67	61.66	58.65	46.86
	Add. Adapter + Same Losses	65.00	64.76	63.46	61.75
Metric	AdaSP	66.67	66.63	67.31	64.34
	DADA	65.00	64.99	65.38	58.64
Anchor-guided	PG-APC (w/o CCSG)	**73.33**	**73.30**	71.15	70.88
	PG-APC + Naive Pooled Anchors	58.33	56.88	59.62	55.95
	CCSG-PG-APC	**73.33**	**73.30**	**72.12**	**72.05**

Table 3: Contextual comparison; reporting protocols are specified in Sec. 5.3.

PG-APC and $70.38/69.73$ for CCSG-PG-APC. Thus prototype reuse degrades without a compatible class-conditional reference geometry.

Contextual comparison. Table 3 is contextual: SMG and non-PG-APC/naive-pooled iMiGUE entries use the prespecified diagnostic seed, while the two iMiGUE PG-APC rows report paired metrics from the seed attaining both maxima in the fixed five-seed pool and are not used for selection.

5.4 Multiclass Proof-of-Concept

Table 4 evaluates the four-class, four-domain HEROES extension, with each class reference constructed from the four training domains. CCSG-PG-APC reaches 72.92%/71.61% Accuracy/Macro-F1, versus 70.83%/69.41% for PG-APC and 68.75%/69.05% for Add. Adapter + Same Losses. Because HEROES uses conventional actions and a different feature configuration, it tests the K-class extension only, not the SMG/iMiGUE cross-dataset claim.

Method	Acc. (%)	Macro-F1 (%)
DADA	60.42	59.40
NC-Init + PI Park et al. (2026)	62.50	62.79
AdaSP	66.67	66.83
Add. Adapter + Same Losses	68.75	69.05
PG-APC	70.83	69.41
CCSG-PG-APC	**72.92**	**71.61**

Table 4: HEROES four-class proof-of-concept.

6 Conclusion

We study cross-dataset anchor generality: whether heterogeneous datasets can reuse one fixed class-reference system rather than maintain dataset-specific prototypes. Direct swapping and naive pooling both degrade recognition. On SMG, using the shared CCSG-reference pipeline instead of dataset-specific references changes five-seed mean Accuracy by -1.00 point under validation-selected checkpoints. On iMiGUE, five-seed mean Accuracy is $70.38 \pm 1.58\%$ for CCSG-PG-APC versus $68.65 \pm 1.46\%$ for PG-APC ($+1.73$ points) under the same prespecified seeds. Cross-dataset evidence is limited to the SMG–iMiGUE pair and one primary frozen TrackMAE representation; HEROES serves only as a multiclass proof-of-concept. Taken together, the controlled swapping and pooling results indicate that anchor reuse is not achieved by exchanging prototype vectors alone: the results support compatible reference geometry and source-to-reference alignment before local prototype calibration. These findings therefore characterize reference portability within the evaluated datasets rather than universal cross-domain transfer.

AI USE STATEMENT

Generative AI tools were used to assist with converting this manuscript from an AAAI LaTeX source format to the ICLR 2027 LaTeX template and with language editing. The authors reviewed the resulting text and take responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work uses previously released research datasets (SMG, iMiGUE, and HEROES) and does not introduce new data collection or participant recruitment. Our experiments study representation alignment and classification under the dataset protocols described in Sec. 5.1. Because affect-related behavioral recognition can raise privacy, bias, and misuse concerns, the benchmark polarity labels should not be interpreted as direct measurements of a person's internal emotional state. The reported results should be interpreted within these dataset protocols and not as evidence for high-stakes deployment.

REPRODUCIBILITY STATEMENT

Reproducibility is supported by the experimental protocol in Sec. 5.1, which specifies the dataset splits, label mapping, evaluation metrics, five fixed random seeds, validation-based checkpoint selection, frozen TrackMAE features, and the train-only construction of CCSG, mapper statistics, and anchors. The main text gives the mathematical definitions of CCSG, the soft-gated source-to-reference mapper, and PG-APC. Supplementary Secs. A–F provide additional architecture and optimization settings, diagnostic and ablation results, efficiency and sensitivity analyses, and multi-class derivations and evaluation protocols.

REFERENCES

Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023.

Meng Chen, Hongwei Jia, Zechen Li, Wenzhen Jia, Kai Zhao, Hongjun Dai, and Weiming Huang. Cross-city latent space alignment for consistency region embedding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8261–8274. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25ab.html.

Darryl Ho and Samuel Madden. Dejavid: Encoder-agnostic learned temporal matching for video classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24023–24032, 2025.

Mohammad Jalali, Bahar Dibaei Nia, and Farzan Farnia. Towards an explainable comparison and alignment of feature embeddings. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26757–26796. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/jalali25a.html.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pp. 237–255. Springer, 2024.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10631–10642, 2021.

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 143–152, 2020.

Ilaria Mannocchi, Kamal Lamichhane, Marco Carli, and Federica Battisti. Heroes: A video-based human emotion recognition database. In *2022 10th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6. IEEE, 2022.

Khushboo Mishra, Varun Trivedi, and Tanima Dutta. Towards uncertainty-aware unsupervised domain adaptation for videos and time-series with causal optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29421–29430, June 2026.

Jinhee Park, Hee Bin Yoo, Minjun Kim, Byoung-Tak Zhang, and Junseok Kwon. Neural collapse-informed initialization with perturbation injection in classification-based metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8287–8295, 2026.

Li Ren, Chen Chen, Liqiang Wang, and Kien Hua. Towards improved proxy-based deep metric learning via data-augmented domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14811–14819, 2024.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Renaud Vandeghen, Fida Mohammad Thoker, Marc Van Droogenbroeck, and Bernard Ghanem. Trackmae: Video representation learning via track mask and predict. *arXiv preprint arXiv:2603.27268*, 2026.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

Zhaoqiang Xia, Hexiang Huang, Haoyu Chen, Xiaoyi Feng, and Guoying Zhao. Hybrid-supervised hypergraph-enhanced transformer for micro-gesture based emotion recognition. *IEEE Transactions on Affective Computing*, 2025.

Chao Xu, Xijia Tang, and Chenping Hou. Heterogeneous label shift: Theory and algorithm. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69705–69724. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/xu25ab.html`.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Hao Zheng, Jingjun Yi, Qi Bi, Huimin Huang, Haolan Zhan, Yawen Huang, Yuexiang Li, Xian Wu, and Yefeng Zheng. Learning a cross-modal schrödinger bridge for visual domain generalization. *Advances in Neural Information Processing Systems*, 38:167613–167641, 2026.

Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, and Lin Ma. Adaptive sparse pairwise loss for object re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19691–19701, 2023.

Appendix to "Cross-Domain Micro-Action Learning via Class-Conditional Shared Geometry, Reference Alignment, and Prototype Calibration"

In this appendix, we provide the following materials:

A Architecture, optimization, and experimental settings (referring to Sec. 5.1 in the main paper);
B Additional anchor-relative diagnostic results (referring to Sec. 5.3 in the main paper);
C Component ablations for CCSG, the source-to-reference mapper, and PG-APC (referring to Sec. 5.3 in the main paper);
D Post-feature inference overhead and deployment characteristics (referring to Sec. 5.3 in the main paper);
E Full multiclass derivation and HEROES protocol details (referring to Sec. 4.4 and Sec. 5.4 in the main paper);
F Complete hyperparameter sensitivity analyses for CCSG, the mapper, and PG-APC (referring to Secs. 5.1 and 5.3 in the main paper).

A Architecture and Experimental Settings

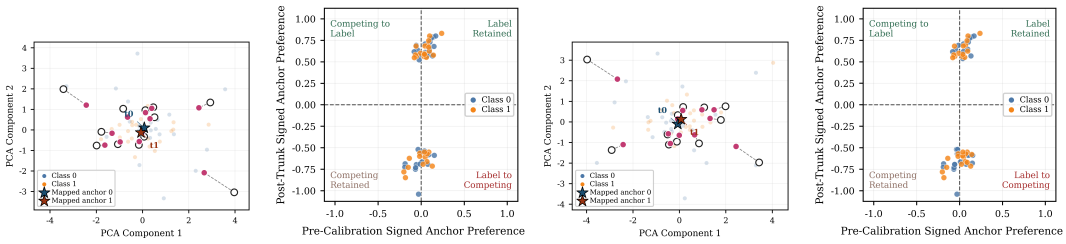
Features and Architecture. We instantiate the generic frozen encoder with TrackMAE Vandeghen et al. (2026) and extract one $D = 768$ representation per input instance (segmented clip for SMG and video for iMiGUE). The encoder remains frozen throughout CCSG construction, mapper fitting, PG-APC training, validation, and testing. The soft-gated source-to-reference mapper operates directly on these cached 768-D features: the class-conditional transforms T_0 and T_1 are estimated from first- and second-order training statistics pooled over the two observed domains, while the logistic routing gate is optimized by Eq. (14) of the main paper on standardized training features, then frozen; it produces (p_0, p_1) without reading a validation/test label. The mapped feature f_i is then passed to PG-APC. Each dataset branch uses the same trunk architecture, LayerNorm–Linear$(768, 128)$–GELU–Dropout, followed by Linear$(128, 2)$ main and auxiliary heads; predictor parameters are not shared across datasets. Consistent with the two scalar gates in Fig. 1 and the parameter count in Table 8, PG-APC uses separate attraction and separation branches, each with LayerNorm–Linear$(768, 128)$–GELU–Dropout–Linear$(128, 1)$ followed by Softplus to produce a_i or r_i. Dropout is 0.24 on both SMG and iMiGUE.

Optimization and Hyperparameter Configuration. PG-APC and adapter controls use AdamW with batch size 16, weight decay 2×10^{-4}, learning rate 3×10^{-4}, and 36 epochs on both SMG and iMiGUE. For CCSG construction, we use $R = 50$ bridge integration steps, RBF bandwidth $h = 40.0$, and a bounded-separation cap $\delta_{\max} = 4$. The two explicit CCSG geometry-loss weights are $\lambda_{\mathrm{CCS}} = 10.0$ and $\lambda_{\mathrm{sep}} = 1.0$. Reported training additionally uses the normalized within-class dispersion $\mathcal{C}_{\mathrm{intra}}$ from Eq. (7) of the main paper as an auxiliary compactness stabilizer with $\lambda_{\mathrm{comp}} = 6.0$; equivalently, $\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{CCSG}} + \lambda_{\mathrm{comp}}\mathcal{C}_{\mathrm{intra}}$. The bridge-only warm-up occupies the first $\eta_{\mathrm{warm}} = 0.20$ fraction of CCSG optimization, after which the core geometry terms and the auxiliary stabilizer are enabled; global geometry statistics are re-estimated every $K_{\mathrm{geo}} = 10$ optimization steps.

For the source-to-reference mapper, the selection marked in Fig. 5 is $\alpha = 1.0$, $\rho_{X,k} = 0.025$, $\rho_{C,k} = 0.06$ (for both classes), and $\tau_g = 0.25$. For PG-APC, we set $\beta_{\mathrm{step}} = 0.24$, $\lambda_{\mathrm{AP}} = 0.035$, ranking margin $m = 0.40$, anchor-affinity temperature $\tau_p = 0.35$, and separation radius $\mu_{\mathrm{sep}} = 0.50$. Auxiliary supervision lasts five epochs. All of these values are fixed before test evaluation and, unless explicitly stated otherwise, the same selected hyperparameter configuration is used for both SMG and iMiGUE. AdaSP and DADA retain their native objectives.

B Additional Anchor-Relative Diagnostics

Anchor-Relative Behavior. For the prespecified SMG run, we conduct a post-hoc, label-conditioned diagnostic of anchor-relative preference. Before calibration, define the signed



(a) PG-APC Post-Trunk Space

(b) PG-APC Preference-Sign Comparison

(c) CCSG-PG-APC Post-Trunk Space

(d) CCSG-PG-APC Preference-Sign Comparison

Figure 3: Post-hoc SMG anchor diagnostics for the prespecified run. Panels (a,c) show post-trunk PCA, while panels (b,d) compare raw/post-trunk preference signs by sign and quadrant only. Panels (a,b) correspond to PG-APC, while panels (c,d) correspond to CCSG-PG-APC.

unsquared-distance difference as

$$s_i^{\text{pre}} = \|f_i - t_{\bar{y}_i}\|_2 - \|f_i - t_{y_i}\|_2, \qquad \bar{y}_i = 1 - y_i. \tag{29}$$

A positive value indicates preference for the label anchor. The post-trunk diagnostic applies the same comparison to $H_\omega(\tilde{f}_i)$ using $H_\omega(t_k)$ as mapped references. These quantities are used only for post-hoc analysis. Figure 3 compares PG-APC and CCSG-PG-APC through within-space structure and pre/post-trunk sign/quadrant changes.

C Component Ablations

The ablations follow the same three consistency roles used in the main paper. CCSG is tested as the reference-consistency stage, the source-to-reference mapper as the source-to-reference consistency stage, and PG-APC as the prototype-relative consistency stage. This organization is important because the modules share one fixed reference system: the ablations test which requirement fails when a component is removed or simplified, rather than treating the components as independent accuracy-oriented add-ons. The single diagnostic seed is fixed a priori and independently of test performance. On iMiGUE, that same prespecified seed yields $72.12/72.05$ Accuracy/Macro-F1 for the complete configuration, which therefore coincides with the paired post-hoc maximum reported for CCSG-PG-APC in Table 3.

CCSG Objective and Stabilizer Ablation. The main paper presents the core CCSG objective in Eq. (9) as the bridge term plus two explicit geometric losses: same-class cross-domain alignment $\mathcal{L}_{\text{CCS}}$ and bounded class separation $\mathcal{L}_{\text{sep}}$. Reported training additionally applies $\lambda_{\text{comp}}\mathcal{C}_{\text{intra}}$ as an auxiliary compactness stabilizer. Accordingly, *w/o CCS* sets $\lambda_{\text{CCS}} = 0$, *w/o* $\mathcal{L}_{\text{sep}}$ sets $\lambda_{\text{sep}} = 0$, and *w/o Compactness Stabilizer* sets only the auxiliary weight $\lambda_{\text{comp}} = 0$. *Bridge Only* disables both explicit geometry losses and the auxiliary stabilizer, leaving only $\mathcal{L}_{\text{bridge}}$. Each variant is retrained from scratch; its anchors are recomputed, the source-to-reference mapper is refit, and PG-APC is retrained using the same downstream protocol. This prevents an ablated CCSG from being evaluated with anchors or mapper statistics estimated from the complete configuration.

In this fixed-pipeline diagnostic run, the complete CCSG training configuration attains the highest Accuracy and Macro-F1 values on both SMG and iMiGUE. On SMG, removing either explicit geometry loss or the auxiliary compactness stabilizer lowers both metrics; removing separation gives the largest Macro-F1 drop and ties the largest Accuracy drop. On iMiGUE, the effects are less uniform, so these single-run differences are interpreted as mechanism diagnostics rather than estimates of statistical superiority.

Source-to-Reference Mapper Ablation. To isolate source-to-reference consistency, we keep the learned CCSG reference geometry, its fixed anchors, and PG-APC protocol unchanged, and modify only the source-to-reference mapping mechanism. The diagnostic *Gate Only* control directly reuses the already trained routing gate from Eq. (14) of the main paper and predicts $\hat{y}_i = \arg\max_k p_k(x_i)$.

	SMG		iMiGUE	
CCSG Objective	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
Bridge Only	66.67	66.52	64.42	62.24
w/o $\mathcal{L}_{\mathrm{CCS}}$	61.67	60.78	66.35	66.19
w/o Compactness Stabilizer	60.00	58.86	65.38	61.76
w/o $\mathcal{L}_{\mathrm{sep}}$	60.00	52.38	62.50	61.90
Full CCSG	**73.33**	**73.30**	**72.12**	**72.05**

Table 5: CCSG loss-term ablation under the stated data splits and evaluation protocol. Each ablated CCSG objective is retrained from scratch, followed by mapper refitting, anchor recomputation, and PG-APC retraining.

	SMG		iMiGUE	
Mapper Variant	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
Gate Only	61.67	58.32	60.58	53.32
Global Affine Map	65.00	64.91	64.42	64.39
Uniform Routing	63.33	61.81	65.38	64.32
Hard Routing	66.67	66.52	61.54	60.35
Full Soft-Gated Mapper	**73.33**	**73.30**	**72.12**	**72.05**

Table 6: Shared-mapper ablation under the stated data splits and evaluation protocol. Gate Only directly evaluates the frozen routing gate and therefore does not invoke PG-APC; for all remaining mapper variants, the CCSG reference and PG-APC hyperparameters are held fixed, while PG-APC is retrained under the same protocol for each mapper variant. Macro-F1 is reported in percent.

It introduces no additional classifier and no extra gate retraining, and bypasses T_k, soft feature placement, and PG-APC. This control therefore tests whether the supervised routing logits alone explain the full mapper result, as opposed to the intended role of routing between class-conditional source-to-reference transforms. *Global Affine Map* replaces the two class-conditional maps T_0, T_1 with one class-agnostic affine map estimated from pooled source and CCSG training statistics. *Uniform Routing* keeps T_0, T_1 but fixes $p_0(x) = p_1(x) = 0.5$. *Hard Routing* replaces the soft mixture with the single map having the largest gate probability. For every variant, PG-APC is retrained under the same protocol so that only the mapper construction differs.

In this fixed-pipeline diagnostic run, the full soft-gated mapper attains the highest Accuracy and Macro-F1 values on both datasets. The Gate Only control separates direct prediction by the supervised router from its intended role of mixing T_0 and T_1. Global affine mapping, uniform routing, and hard routing all yield lower values than the full variant in this run, which is consistent with using both class-conditional transforms and sample-dependent soft fusion for source-to-reference placement before prototype calibration.

PG-APC Component Ablation. Table 7 isolates prototype-relative consistency under the complete pipeline by changing PG-APC while keeping the upstream shared reference fixed. For every component variant, the CCSG construction, shared anchors, and soft-gated source-to-reference mapper are held fixed; only the indicated PG-APC component is changed, and the downstream PG-APC model is retrained under the same protocol. Uniform Affinity fixes $\pi_{i,k} = 0.5$, while Global Scalar Gates uses two shared non-negative values. PG-APC (w/o CCSG) is included only as a dataset-specific-anchor reference and is not the upstream configuration used by the component variants.

The component effects are dataset-dependent in this diagnostic run: separation and affinity produce larger changes on iMiGUE, whereas replacing sample-adaptive gates with global scalars produces a larger change on SMG. Because all component variants share the same upstream CCSG and mapper, the table is used as a component-wise diagnostic of PG-APC rather than as a repeated-run ranking of variants. The PG-APC (w/o CCSG) row serves only to contextualize the difference between the dataset-specific-reference baseline and the complete shared-reference pipeline.

Setting	SMG		iMiGUE	
	Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
Uniform affinity	70.00	69.46	57.69	53.85
w/o Attraction	68.33	67.90	63.46	59.64
w/o Opp.-anchor sep.	65.00	64.99	57.69	50.35
Global scalar gates	63.33	62.96	68.27	67.17
w/o Ranking	68.33	68.25	60.58	54.86
w/o Step control	66.67	66.33	63.46	60.15
w/o Early Aux. Sup.	70.00	69.87	61.54	58.97
PG-APC (w/o CCSG)	73.33	73.30	71.15	70.88
CCSG-PG-APC	**73.33**	**73.30**	**72.12**	**72.05**

Table 7: PG-APC component analysis under fixed CCSG shared anchors and the fixed soft-gated mapper. PG-APC (w/o CCSG) is an external dataset-specific-anchor reference rather than a component variant. Accuracy and Macro-F1 are reported in percent.

D Post-Feature Inference Overhead

Post-Feature Inference Overhead. Table 8 reports inference cost for the downstream path under the same cached 768-D TrackMAE representation. The frozen encoder is excluded because it is shared by all methods; CCSG construction and mapper-statistic estimation are also offline procedures and therefore do not enter the measured inference path.

Method	Params (M)	FLOPs (M)	Latency (ms)
MLP Head	0.1002	0.2020	0.2027 ± 0.0571
Add. Adapter + Same Losses	0.2993	0.6008	0.5070 ± 0.0196
PG-APC (w/o CCSG)	0.3004	0.6277	0.7320 ± 0.2556
CCSG-PG-APC	0.3020	0.6400	0.8757 ± 0.1064

Table 8: Post-feature inference overhead under the same cached 768-D TrackMAE representation. The frozen encoder, offline CCSG construction, and mapper-statistic estimation are excluded; parameter counts include only trainable modules active at inference, excluding the early auxiliary head. Latency is measured on an NVIDIA RTX 4090 in FP32 with 100 warm-up and 1,000 measured runs at batch size one.

Relative to PG-APC without CCSG, the complete framework adds $1{,}538$ parameters (0.51%) and 0.0123M FLOPs per sample (1.96%), while mean latency increases from 0.7320 ms to 0.8757 ms, i.e., by 0.1437 ms (19.6%) at batch size one. The complete method is also slower than the additive-adapter control (0.8757 vs. 0.5070 ms), showing that the online mapping and prototype-calibration operations introduce measurable wall-clock overhead despite the small increase in parameter count and arithmetic cost. These measurements characterize only the post-feature path and should not be interpreted as end-to-end video inference latency.

E Full Multiclass Derivation and HEROES Protocol

The main paper presents the three multiclass definitions needed to understand the extension directly: the K-way soft-gated mapper, the nearest competing anchor, and the multiclass ranking loss (Sec. 4.4), together with the HEROES proof-of-concept results (Sec. 5.4). This section supplies the omitted class-aggregation terms, the complete PG-APC residual construction, and the HEROES evaluation protocol.

E.1 Complete K-Class Formulation

Let $K \geq 2$, $\mathcal{K} = \{0, \ldots, K-1\}$ be the class set, and $\mathcal{D}$ the observed-domain set. The binary SMG/iMiGUE setting uses $|\mathcal{D}| = 2$, whereas the HEROES protocol uses four domains.

Multiclass CCSG. The class-conditional bridge in Eqs. (4)–(5) is instantiated for every $(d, k) \in \mathcal{D} \times \mathcal{K}$. For more than two domains, the class-k CS term averages Eq. (6) over unordered domain pairs. The explicit alignment loss becomes

$$\mathcal{L}_{\mathrm{CCS}}^{(K)} = \frac{1}{K} \sum_{k \in \mathcal{K}} \mathcal{D}_{\mathrm{CS}}^{(k)}, \tag{30}$$

where $\mathcal{D}_{\mathrm{CS}}^{(k)}$ is the class-k Cauchy–Schwarz discrepancy across the observed domains (Eq. (6) in the two-domain case).

The multiclass auxiliary compactness statistic is

$$\mathcal{C}_{\mathrm{intra}}^{(K)} = \frac{1}{K} \sum_{k \in \mathcal{K}} \frac{|Z_k^C|^{-1} \sum_{z \in Z_k^C} \|z - \mu_k^C\|_2^2}{|Z_k^C|^{-1} \sum_{z \in Z_k^C} \|z\|_2^2 + \varepsilon}. \tag{31}$$

As in the binary case, this normalized dispersion is an auxiliary training stabilizer rather than a core explicit geometry loss.

For separation, define the explicit multiclass loss as the negative average bounded center separation over all unordered class pairs:

$$\mathcal{L}_{\mathrm{sep}}^{(K)} = -\frac{2}{K(K-1)} \sum_{\substack{k,j \in \mathcal{K} \\ k<j}} \mathrm{clip}\left(\frac{\|\mu_k^C - \mu_j^C\|_2^2}{\frac{1}{2}(\|\mu_k^C\|_2^2 + \|\mu_j^C\|_2^2) + \varepsilon}, 0, \delta_{\max} \right). \tag{32}$$

Minimizing $\mathcal{L}_{\mathrm{sep}}^{(K)}$ increases bounded inter-class center separation.

The core K-class CCSG objective is therefore

$$\mathcal{L}_{\mathrm{CCSG}}^{(K)} = \mathcal{L}_{\mathrm{bridge}}^{(K)} + \lambda_{\mathrm{CCS}} \mathcal{L}_{\mathrm{CCS}}^{(K)} + \lambda_{\mathrm{sep}} \mathcal{L}_{\mathrm{sep}}^{(K)}, \tag{33}$$

where $\mathcal{L}_{\mathrm{bridge}}^{(K)}$ retains Eq. (5)'s $\frac{1}{2}$ group-sum reduction over $d \in \mathcal{D}$ and $k \in \mathcal{K}$; the domain weights ω_d are used only in the barycentric root construction of Eq. (3) and do not additionally weight $\mathcal{L}_{\mathrm{bridge}}^{(K)}$. Reported training additionally uses $\lambda_{\mathrm{comp}} \mathcal{C}_{\mathrm{intra}}^{(K)}$ as the same auxiliary compactness stabilizer used in the binary formulation.

Multiclass soft-gated mapping. For every $k \in \mathcal{K}$, the class-conditional transform T_k keeps the form of Eq. (11) in the main paper. The routing gate produces K logits and the mapper uses the K-way mixture defined in Sec. 4.4 of the main paper. Applying Eq. (16) of the main paper to each class yields the shared anchor set $\mathcal{T}_K = \{t_k \mid k \in \mathcal{K}\}$.

Multiclass PG-APC. For mapped feature $f_i^{(K)}$, define $d_{i,k}^{(K)} = \|f_i^{(K)} - t_k\|_2$. The K-anchor affinity is

$$\pi_{i,k}^{(K)} = \frac{\exp(-d_{i,k}^{(K)}/\tau_p)}{\sum_{\ell \in \mathcal{K}} \exp(-d_{i,\ell}^{(K)}/\tau_p)}. \tag{34}$$

The nearest competing anchor index $c_i(k)$ is defined in Sec. 4.4 of the main paper. Using this competitor, the separation displacement becomes

$$V_{i,k}^{\mathrm{sep},(K)} = r_i [\mu_{\mathrm{sep}} - d_{i,c_i(k)}^{(K)}]_+ \frac{f_i^{(K)} - t_{c_i(k)}}{\|f_i^{(K)} - t_{c_i(k)}\|_2 + \varepsilon}. \tag{35}$$

The attraction term keeps the binary form, $V_{i,k}^{\mathrm{att}} = a_i(t_k - f_i^{(K)})$, and the complete anchor-weighted residual is

$$V_{\mathrm{ap}}^{(K)}(f_i^{(K)}; \mathcal{T}_K) = \sum_{k \in \mathcal{K}} \pi_{i,k}^{(K)} \left(V_{i,k}^{\mathrm{att}} + V_{i,k}^{\mathrm{sep},(K)} \right). \tag{36}$$

The calibrated feature is $\tilde{f}_i^{(K)} = f_i^{(K)} + V_{\mathrm{ap}}^{(K)}(f_i^{(K)}; \mathcal{T}_K)$. The multiclass ranking term is given in Sec. 4.4 of the main paper. The residual-step regularizer keeps the form of Eq. (24), while $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{L}_{\mathrm{aux}}$ become K-class classification cross-entropies with K-logit classifiers. At validation and test time, routing probabilities and anchor affinities are computed from the input feature using training-derived statistics and anchors; no validation/test label is used to choose a class-conditional transform.

E.2 HEROES Protocol Details

The HEROES experiment uses four domains and four classes. CCSG forms each class-conditional reference from the four training-domain distributions used in this protocol. We use an actor-disjoint split over the 16 actors: 10 actors for training, 3 for validation, and 3 for testing. This yields 160/48/48 videos, corresponding to 40/12/12 videos per class. All compared methods use the same split and the same separate feature configuration. Checkpoints are selected on the validation split and the test split is evaluated once after model selection.

PG-APC is the calibration-only multiclass reference, whereas CCSG-PG-APC uses the complete extension above. Because this feature configuration differs from the 768-D TrackMAE instantiation used for the main SMG/iMiGUE experiments, the HEROES experiment is used only to check the K-class extension of the anchor-based formulation; it is not presented as additional cross-dataset or backbone-matched evidence. The Accuracy and Macro-F1 results are reported only in Table 4 of the main paper to avoid duplicating the primary comparison in the supplementary material.

F Hyperparameter Sensitivity

CCSG Hyperparameter Sensitivity. We analyze four CCSG hyperparameters governing the reference geometry and its training stability: the RBF bandwidth h, the two explicit geometry-loss weights λ_{CCS} and λ_{sep}, and the auxiliary compactness-stabilizer weight λ_{comp}. Each parameter is varied independently on the SMG validation split while the remaining configuration is fixed; SMG validation Accuracy determines the selected value, and the test split is not used for hyperparameter selection. The validation-selected CCSG setting is then shared with iMiGUE and applied without dataset-specific retuning. Numerical-integration and optimization-schedule settings are kept fixed.

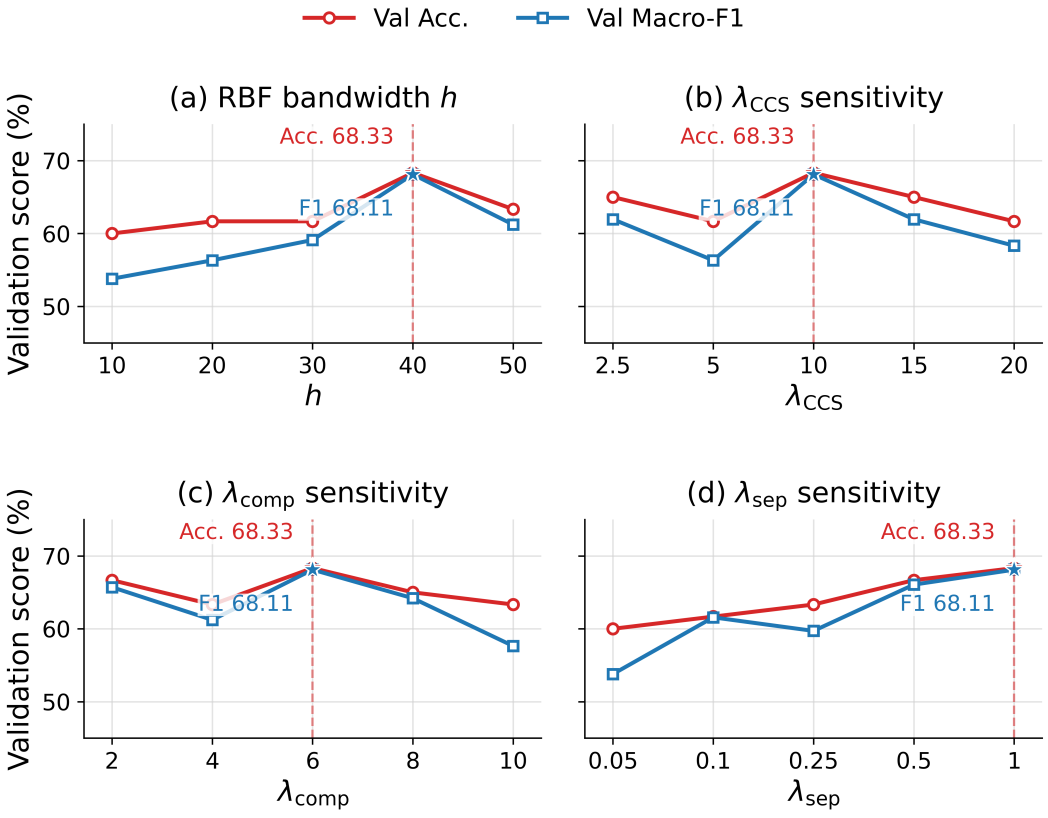


Figure 4: CCSG hyperparameter sensitivity on the SMG validation split. Each hyperparameter is varied individually while the remaining configuration is fixed; stars mark the selected values, which are shared with iMiGUE without retuning.

Shared-Mapper Hyperparameter Sensitivity. We next analyze the source-to-reference mapper by varying the covariance shrinkage coefficient α, the source and CCSG ridge coefficients $\rho_{X,k}$ and $\rho_{C,k}$, and the routing temperature τ_g on the SMG validation split while keeping CCSG and PG-APC fixed. SMG validation Accuracy determines each selected value. The selected mapper configuration is then shared with iMiGUE without dataset-specific retuning. This isolates the sensitivity of source-to-reference placement from downstream calibration. The four sweeps probe complementary aspects of the same mapping mechanism: α controls how strongly off-diagonal covariance structure is retained, the two ridge terms regulate the conditioning of the source and reference covariance estimates, and τ_g controls how sharply samples are routed between the class-conditional transforms. We therefore use these individually selected values together as one validation-derived mapper configuration rather than interpreting any single curve as an independent source of recognition gain. At the selected operating point, $\alpha = 1.0$ uses the diagonal-shrinkage endpoint of Eq. (10), while the distinct source/reference ridge terms retain separate conditioning for the two covariance estimates. The selected $\tau_g = 0.25$ is used in the same soft-routing rule of Eq. (13), rather than replacing it with the hard-routing control in Table 6. None of these sweeps changes the CCSG anchors or PG-APC settings, so the curves isolate mapper conditioning and routing sharpness rather than upstream or downstream retuning.

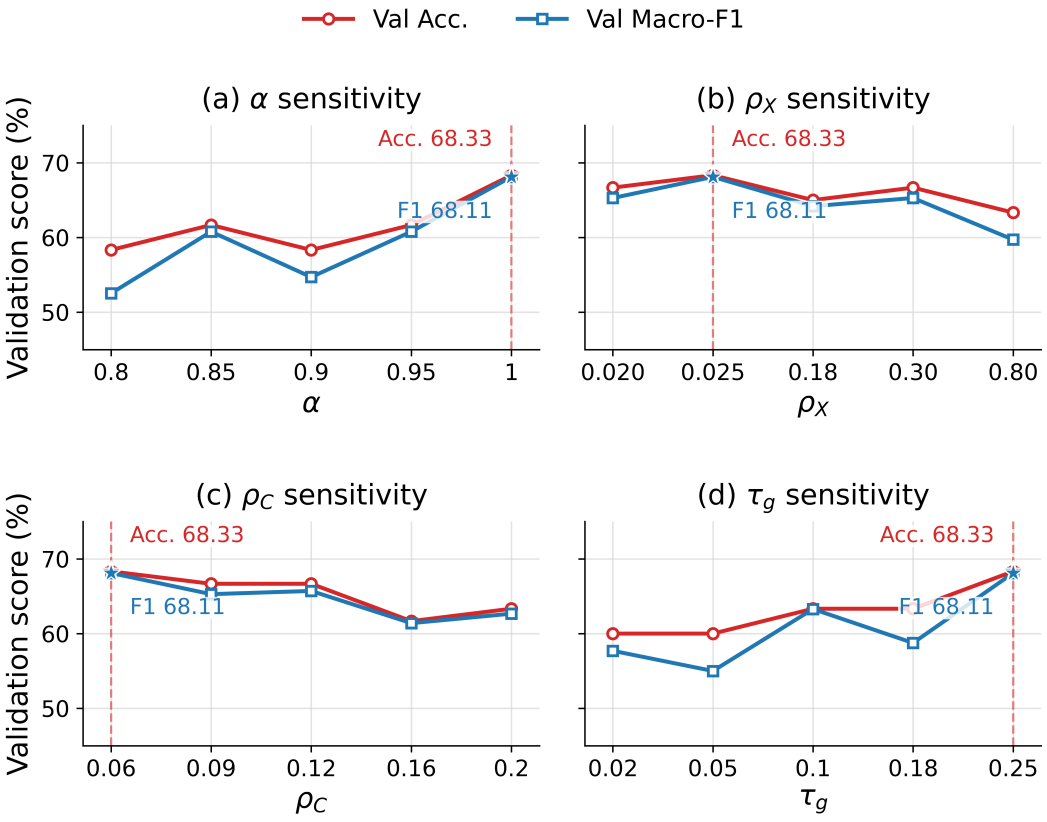


Figure 5: Source-to-reference mapper hyperparameter sensitivity on the SMG validation split. Each hyperparameter is varied individually while the remaining configuration is fixed; stars mark the selected values, which are shared with iMiGUE without retuning.

PG-APC Hyperparameter Sensitivity. After fixing the selected CCSG and source-to-reference mapper configurations, we analyze PG-APC on the SMG validation split by varying one calibration hyperparameter at a time while keeping the remaining settings unchanged. Specifically, we vary the anchor-affinity temperature τ_p, the opposite-anchor separation radius μ_{sep}, the anchor-ranking weight λ_{AP}, and the ranking margin m. For each setting, PG-APC is retrained from scratch with the upstream CCSG and mapper fixed, and validation Accuracy is used for selection. Consistent with the shared-hyperparameter protocol above, the selected PG-APC configuration is fixed and

applied unchanged to iMiGUE. These curves are local sensitivity diagnostics rather than a joint hyperparameter search.

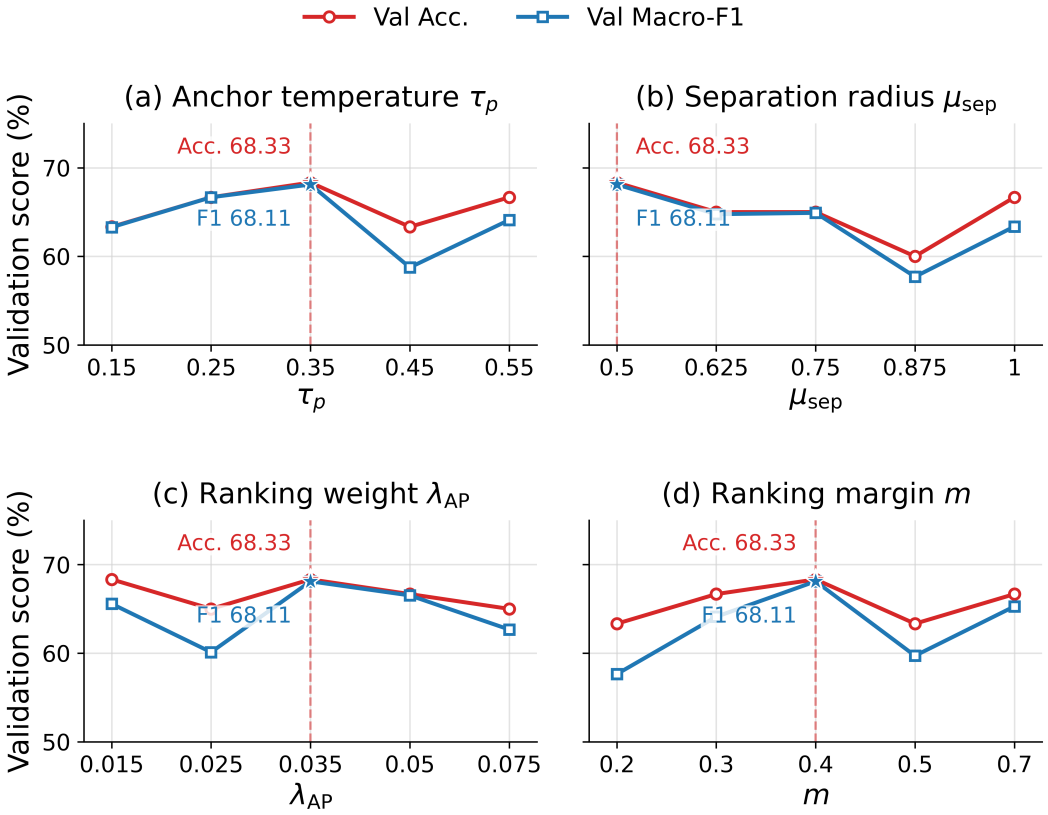


Figure 6: PG-APC hyperparameter sensitivity on the SMG validation split. Each hyperparameter is varied individually with CCSG and the source-to-reference mapper fixed; stars mark the selected values, which are shared with iMiGUE without retuning.

Sensitivity Summary. Across the three stages, hyperparameters are selected only from the SMG validation split and are then transferred unchanged to iMiGUE. The sweeps are deliberately one-dimensional: they test whether the reported operating point remains reasonable under local perturbations while preserving the rest of the pipeline, rather than performing a joint search over CCSG, mapper, and PG-APC settings. Accordingly, the curves support the stability of the reported configuration around the selected operating point, but they are not used to claim that the same numerical values are universally optimal for other feature encoders or datasets.