CONSENSUS-ANCHORED RESIDUAL LEARNING FOR INCOMPLETE MULTI-VIEW MULTI-LABEL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete multi-view multi-label classification seeks to leverage heterogeneous information from partially observed views and labels. Existing approaches often couple consensus learning with the extraction of view-specific information, allowing complementary updates to perturb the shared predictor. Under sparse supervision, such coupling may compromise reliable consensus and diminish the predictive gains from view-specific information. To address this issue, we propose Consensus-Anchored Residual Learning (CARE), a new two-stage framework that preserves shared knowledge while learning complementary corrections. Specifically, the first stage establishes a consensus predictor through cross-view fusion guided by label semantics. The second stage anchors this predictor and learns residual refinements around it, while retaining the original consensus as an incumbent during model selection. This design provides a stable reference for exploiting view-specific information and evaluating its predictive contribution. Furthermore, our analysis characterizes conditions for beneficial residual updates and relates the gains retained after selection to conditional complementary information, approximation error, and selection uncertainty. Extensive experiments on six benchmark datasets demonstrate the superiority of CARE over state-of-the-art methods, together with its effectiveness in chest radiograph screening.

1 INTRODUCTION

Multi-view multi-label recognition draws on measurements that support different aspects of the same instance. Color and texture descriptors can identify different image attributes, while frontal and lateral chest radiographs expose different anatomical structures. Acquiring every view and annotating every relevant label is often impractical. Portable radiography may provide a single projection, and clinical reports may record selected findings (Irvin et al., 2019). Missing inputs and annotations therefore restrict both the evidence available for prediction and the examples that supervise its use. Consensus learning combines evidence across observed sources, while complementary learning extracts label information that this combination leaves unresolved (Li et al., 2019; Baltrušaitis et al., 2019). A shared predictor pools supervision across views, while each view-specific correction depends on the joint observation of its view and label. Sparse supervision therefore makes the value of a correction relative to the shared predictor central to complementary learning.

Existing methods address this problem by improving representation, recovering missing evidence, and controlling fusion. Structured approaches connect cross-view representations with label dependencies, including iMVWL and NAIM3L (Tan et al., 2018; Li & Chen, 2022). Deep models extend this coupling through cross-view agreement, label-guided interactions, and embedding imputation, as in DICNet, LMVCAT, and AIMNet (Liu et al., 2023b;c; 2024a). Shared–specific models make complementarity explicit through separate channels, while semantic compression and reliability estimates refine the information assigned to the classifier (Liu et al., 2023a; Wen et al., 2026b; Liu et al., 2024b; 2025; Wen et al., 2026a). Joint optimization nevertheless allows updates from these channels to change the consensus function. A view-specific cue can improve a local prediction while its coupled update displaces useful shared knowledge. Related analyses identify uneven modality learning and gradient interference in joint training (Wang et al., 2020; Peng et al., 2022; Yu et al., 2020). With missing views and labels, these updates receive uneven supervision across branches.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

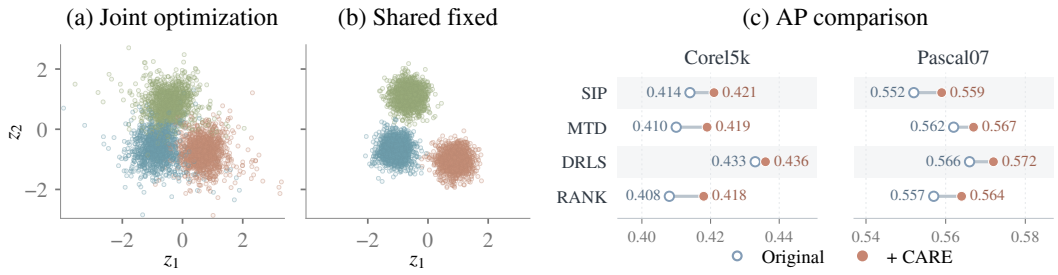


Figure 1: Preserving shared structure during complementary refinement. (a)–(b) compare joint refinement and a fixed shared branch in a controlled task-extension experiment, using common samples and coordinates. (c) compares AP before and after adapting four methods to CARE on Corel5k and Pascal07.

The objective can then improve by redistributing predictive information between shared and specific representations, even when the shared representation becomes less discriminative. Figure 1(a)–(b) illustrates this interaction in a controlled synthetic experiment that introduces view-specific tasks after shared learning. Joint refinement weakens the separation of previously learned semantic groups, while fixing the shared branch preserves their structure. A stable prediction reference separates the contribution of specific features from changes in the shared predictor.

We propose CARE, a new two-stage framework that learns complementary information as a correction to an established consensus predictor. Consensus estimation integrates observed views and labels into a shared predictor. Complementary learning expands its prediction family through gated residuals while preserving the full consensus state, including normalization statistics and label prototypes. Each residual can exploit a view for one label while leaving another prediction unchanged. The archived predictor remains available for every observation mask, giving refinement and selection a common reference for assessing the contribution of view-specific information. CARE's transferable two-stage framework can be applied to established multi-view multi-label methods that disentangle shared and view-specific representations, improving prediction through complementary corrections around a preserved consensus. Adapting SIP, MTD, DRLS (Yan et al., 2025), and RANK to CARE improves AP on Corel5k and Pascal07, as shown in Figure 1(c). These gains support the value of preserving consensus during complementary refinement.

Under binary cross-entropy, the gain of an augmented predictor equals the conditional label information available beyond consensus, plus the consensus approximation error, minus the augmented predictor's remaining approximation error. This identity specifies the information worth learning and the accuracy needed to recover its value. The residual objective admits improving directions around the fixed reference, while anchoring removes the shared-logit drift term from the same correction. Incumbent-inclusive selection recovers the strongest available candidate gain up to an estimation allowance. Paired stage comparisons on six benchmarks and mechanism ablations examine the resulting gains, while chest radiograph screening examines missing projections and labels.

- We propose CARE, a two-stage framework that separates consensus learning from complementary refinement. Gated, label-specific residuals and incumbent-inclusive selection preserve a common reference for learning corrections and retaining their predictive gains.
- CARE admits a unified analysis connecting conditional label information, residual approximation, and selection uncertainty. The analysis identifies improving update directions and characterizes when complementary information yields gains over the preserved predictor.
- CARE achieves leading performance on six benchmarks and supports chest radiograph screening under missing projections and annotations. Paired-stage comparisons and component ablations connect these results to refinement around a stable consensus.

2 Problem Formulation

Each of n instances has V views, $\mathbf{x}_i^{(v)} \in \mathbb{R}^{d_v}$, and a binary label vector $\mathbf{y}_i \in \{0,1\}^c$. Masks m_{iv} and u_{ij} indicate an observed view and an observed label. Every instance retains a nonempty view set $\mathcal{V}_i$. The predictor accepts $(\mathbf{x}_i^{\text{obs}}, \mathbf{m}_i)$ and returns scores for all labels. Training uses $\mathcal{O} = \{(i,j) : u_{ij} =$

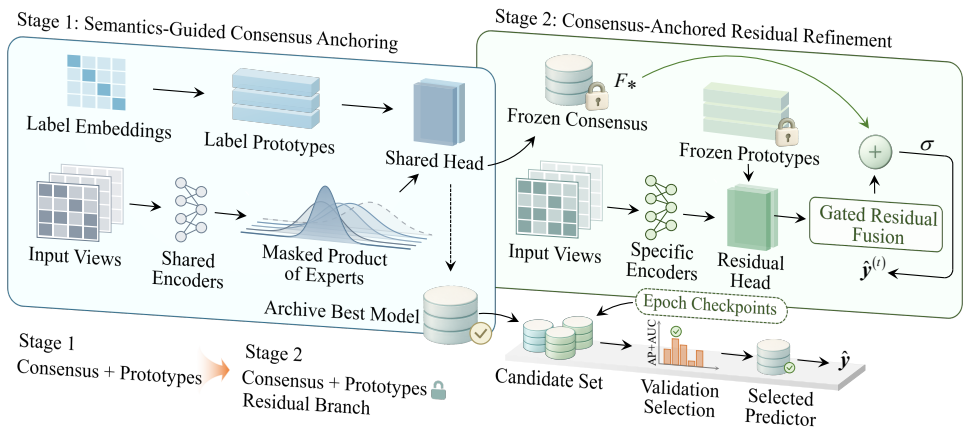


Figure 2: Semantics-guided consensus learning and anchored residual refinement in CARE. Stage 1 combines masked expert aggregation with label prototypes to learn and archive a shared predictor. Stage 2 freezes the consensus and prototypes while learning gated residuals.

1}, with per-view supervision additionally requiring $m_{iv} = 1$. For fixed positive class weights, the masked classification loss is

$$\mathcal{L}_{\text{cls}}(\ell) = \frac{1}{|\mathcal{O}|} \sum_{(i,j)\in\mathcal{O}} q_{ij}\big[\text{softplus}(\ell_{ij}) - y_{ij}\ell_{ij}\big], \quad q_{ij} = w_j^+ y_{ij} + w_j^-(1 - y_{ij}). \tag{1}$$

An empty supervised set contributes zero loss. Unit weights recover ordinary binary cross-entropy. Let F_* denote the archived consensus logit map and D_ϕ a correction. CARE learns $F_\phi = F_* + D_\phi$ and retains F_* in its candidate set. The central comparison is the improvement of the selected predictor over this same archived function, measured by a score S and its empirical estimate $\widehat{S}$.

3 Consensus-Anchored Residual Learning

CARE is a two-stage framework whose prediction family expands around a fixed consensus function, as shown in Figure 2. The first stage learns a shared predictor through masked expert aggregation guided by label semantics. The second preserves the consensus and prototypes, learns gated residual corrections, and selects among the archived predictor and residual checkpoints.

3.1 Availability-Aware Consensus Anchoring and Control of Prediction Drift

Each observed view defines a diagonal Gaussian expert $q_v(h) = \mathcal{N}(h; \mu_v, \Sigma_v)$. A standard normal prior and the available experts form $q_s(h) \propto p_0(h) \prod_v q_v(h)^{m_v}$ (Hinton, 2002; Wu & Goodman, 2018). Completing its quadratic log density gives

$$\Lambda_s = I + \sum_v m_v \Sigma_v^{-1}, \qquad q_s = \mathcal{N}\left(\Lambda_s^{-1} \sum_v m_v \Sigma_v^{-1} \mu_v, \Lambda_s^{-1}\right). \tag{2}$$

An absent view contributes zero precision. The prior keeps the aggregate precision positive definite for every mask. For nested observed-view sets $A \subseteq B$, adding an expert gives $\Lambda_B \succeq \Lambda_A$ and $\Sigma_B \preceq \Sigma_A$. Aggregation therefore adjusts its latent uncertainty with the evidence actually available.

Learnable Gaussian label prototypes define a common output space. A deterministic scorer maps interaction moments between the consensus Gaussian and each prototype to logits. For independent $H \sim \mathcal{N}(\mu, \Sigma)$ and label prototype $L_j \sim \mathcal{N}(\nu_j, T_j)$, useful moments include

$$\mathbb{E}[H^T L_j] = \mu^T \nu_j, \qquad \text{Var}(H^T L_j) = \mu^T T_j \mu + \nu_j^T \Sigma \nu_j + \text{tr}(\Sigma T_j). \tag{3}$$

For moment features Φ_j, the per-label affine scorer gives $F_j = w_j^T \Phi_j + b_j$. Its training objective combines observed-label classification, auxiliary per-view classification, masked reconstruction, prototype alignment, and view-to-fusion coherence,

$$\mathcal{L}_1 = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{aux}} + \alpha\mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{lab}} + \eta\mathcal{L}_{\text{coh}}. \tag{4}$$

The best consensus checkpoint under $\widehat{S}$ defines state $s_* = (\theta_*, b_*)$, including network parameters, prototypes, and persistent normalization buffers (Ioffe & Szegedy, 2015). All residual candidates use the deterministic logit map $F_*(x, m) = F(x, m; s_*)$. Fixing this complete state removes both gradient-driven parameter changes and changes caused by running statistics.

For a fixed logit correction d, changing the consensus by e adds a distinct contribution to the classification loss. Taylor's integral formula and positive semidefinite logistic curvature give

$$\begin{aligned}\mathcal{L}_{\mathrm{cls}}(F_* + d + e) - \mathcal{L}_{\mathrm{cls}}(F_* + d) &= \langle \nabla \mathcal{L}_{\mathrm{cls}}(F_* + d), e\rangle + \mathcal{C}(e),\\ \mathcal{C}(e) &= \int_0^1 (1-t) e^T \nabla^2 \mathcal{L}_{\mathrm{cls}}(F_* + d + te) e \, dt \geq 0.\end{aligned} \tag{5}$$

A positive inner product certifies an additional cost for the same correction. Fixing the complete consensus state sets $e = 0$ and removes this term. Complementary optimization then changes predictions through the residual alone, so its contribution has a stable reference.

3.2 Zero-Initialized Gated Corrections with Constructive Descent Directions

Each specific encoder produces moment features h^{p}_{iv} against the frozen prototypes. A residual output head and a scalar view gate define

$$\delta_{iv} = W_{\mathrm{r}} h^{\mathrm{p}}_{iv} + b_{\mathrm{r}}, \quad g_v = \sigma(a_v), \quad F_{\phi,i} = F_*(x_i, m_i) + \sum_v m_{iv} g_v \delta_{iv}. \tag{6}$$

We initialize $W^{(0)}_{\mathrm{r}} = b^{(0)}_{\mathrm{r}} = 0$. At $a^{(0)}_v = 0$, gates start at one half and have maximal sigmoid derivative. The zero head makes the correction vanish regardless of the gate values or observation pattern. Thus $F_{\phi^{(0)}} = F_*$, and the family $\mathcal{F}_* = \{F_* + D_\phi\}$ contains the archived predictor exactly. Each residual predicts a separate correction for every label, and its scalar gate controls the overall exposure of that view.

Residual learning updates the specific encoders, output head, and gates using

$$\begin{aligned}\mathcal{L}_2 = \mathcal{L}_{\mathrm{cls}}(F_* + D_\phi) + \sum_v \mathcal{L}_{\mathrm{cls}}(F_* + \delta_v; m_v \mathbf{u})\\ + \alpha \mathcal{L}_{\mathrm{rec}} + \gamma \mathcal{L}_{\mathrm{dis}} + \frac{\lambda_{\mathrm{gate}}}{V} \sum_v g_v.\end{aligned} \tag{7}$$

The auxiliary loss trains a correction wherever its view and label are observed. Reconstruction supports the private representation, and disentanglement penalizes duplication of the detached consensus representation. Gates have a separate optimizer group with zero weight decay. Writing $r_{ij} = u_{ij} q_{ij} [\sigma(F_{\phi,ij}) - y_{ij}]/|\mathcal{O}|$, their derivative is

$$\frac{\partial \mathcal{L}_2}{\partial a_v} = g_v(1 - g_v) \left[\sum_i m_{iv} \langle r_i, \delta_{iv} \rangle + \frac{\lambda_{\mathrm{gate}}}{V} \right]. \tag{8}$$

The gradient favors opening a gate when its correction aligns with the negative classification gradient strongly enough to offset its exposure penalty. The head can receive a classification gradient at zero initialization. As it moves away from zero, classification gradients also reach the specific encoder; reconstruction can train that encoder throughout.

For $|\mathcal{O}| > 0$, the correction has a controlled effect on classification. Let $Q = \mathrm{diag}(u_{ij} q_{ij}/|\mathcal{O}|)$ and evaluate $r = \nabla \mathcal{L}_{\mathrm{cls}}(F_*)$. Logistic curvature satisfies $0 \preceq \nabla^2 \mathcal{L}_{\mathrm{cls}} \preceq Q/4$. Taylor expansion along $d = D_\phi$, with $A = -\langle r, d \rangle$ and $B = d^T Q d$, gives

$$\mathcal{L}_{\mathrm{cls}}(F_* + td) - \mathcal{L}_{\mathrm{cls}}(F_*) \leq -tA + \frac{t^2 B}{8}, \qquad 0 < t < \frac{8A}{B} \quad \text{if } A, B > 0. \tag{9}$$

Scaling the residual output head realizes td with gates fixed. Every aligned direction therefore admits a strictly improving classification step. The scale $t = 4A/B$ gives a decrease of at least $2A^2/B$ in this upper bound.

The zero head supplies such a direction whenever the main classification gradient has a nonzero projection onto the residual features. With encoders and gates fixed, stack the head parameters into

ϑ and write $D = \Psi\vartheta$. The design matrix Ψ includes the masked, gated features and bias coordinates. Set $b = \Psi^T r$ at $\vartheta = 0$. The realizable head update $\vartheta(t) = -tb$ gives

$$\mathcal{L}_{\mathrm{cls}}(F_* - t\Psi b) - \mathcal{L}_{\mathrm{cls}}(F_*) \le -t\|b\|^2 + \frac{t^2}{8}\|Q^{1/2}\Psi b\|^2. \tag{10}$$

If $b \neq 0$, then $B_\Psi = \|Q^{1/2}\Psi b\|^2 > 0$, since $r^T\Psi b = \|b\|^2$ and r vanishes outside the supervised coordinates. Choosing $t = 4\|b\|^2/B_\Psi$ guarantees a classification decrease of at least $2\|b\|^4/B_\Psi$. This feature alignment identifies available corrections at initialization. Auxiliary and representation losses shape the features used in subsequent updates.

3.3 Incumbent-Inclusive Selection with Guarantees for Complementary Gain

Let $F_0 = F_*$ and $F_1, \ldots, F_K$ be complete residual candidates. All candidates use the same evaluation path and metric. The score $S(F)$ includes any subset prediction and late fusion used at evaluation. Selection starts from the archived score and updates upon strict improvement,

$$\widehat{k} \in \operatorname*{arg\,max}_{0 \le k \le K} \widehat{S}(F_k), \qquad \widehat{\Delta} = \widehat{S}(F_{\widehat{k}}) - \widehat{S}(F_*) \ge 0. \tag{11}$$

Ties with the incumbent retain the archived model. The rule needs only the incumbent and the strongest residual checkpoint in memory. Its empirical comparison is exact for every score sequence. Population improvement also depends on the information recovered by the candidate and the uncertainty of its estimated score.

Condition on the fitted functions. Let Z contain the frozen consensus representation and mask, and U the observed views used by every residual candidate. Both information sets are fixed across candidates. Write $p_j = \Pr(Y_j = 1 \mid Z)$ and $p_j^+ = \Pr(Y_j = 1 \mid Z, U)$. For unweighted population binary cross-entropy R, define $E_* = c^{-1}\sum_j \mathbb{E}D_{\mathrm{B}}(p_j\|\sigma(F_{*,j}))$ and $E_k = c^{-1}\sum_j \mathbb{E}D_{\mathrm{B}}(p_j^+\|\sigma(F_{k,j}))$, with Bernoulli divergence D_{B}.

Proposition 3.1 (Conditional value of a correction). *For finite risks and $p_j, p_j^+ \in (0,1)$,*

$$R(F_*) - R(F_k) = \underbrace{\frac{1}{c}\sum_j I(Y_j; U \mid Z)}_{\mathcal{I}_{\mathrm{comp}}} + E_* - E_k, \qquad \inf_{F \in \mathcal{F}_*} R(F) \le R(F_*). \tag{12}$$

To derive the identity, condition the cross-entropy on each information set. Its Bayes term is conditional entropy and its excess is Bernoulli divergence (Gneiting & Raftery, 2007),

$$R(F_*) = \frac{1}{c}\sum_j H(Y_j \mid Z) + E_*, \qquad R(F_k) = \frac{1}{c}\sum_j H(Y_j \mid Z, U) + E_k. \tag{13}$$

Subtracting gives the conditional-information term. The zero correction establishes inclusion of F_* and proves the second claim. Thus informative private features have value when the residual family recovers them with error below $\mathcal{I}_{\mathrm{comp}} + E_*$. Anchoring makes this target common across candidates.

Residual approximation makes the criterion constructive. Define the ideal correction $d_j^* = \mathrm{logit}(p_j^+) - F_{*,j}$ and $A(a) = \log(1 + e^a)$. For $a_j^* = \mathrm{logit}(p_j^+)$, the divergence is the Bregman remainder of A. Since $A'' \le 1/4$,

$$D_{\mathrm{B}}(p_j^+\|\sigma(a)) = A(a) - A(a_j^*) - A'(a_j^*)(a - a_j^*) \le \tfrac{1}{8}(a - a_j^*)^2. \tag{14}$$

Consequently $E_k \le (8c)^{-1}\sum_j \mathbb{E}(D_{k,j} - d_j^*)^2$ whenever this squared error is finite. The actual residual architecture enters through its ability to approximate the conditional log-odds correction.

Theorem 3.2 (Recovery of the available complementary gain). *If $\max_{0 \le k \le K} |S(F_k) - \widehat{S}(F_k)| \le \varepsilon$, then incumbent-inclusive selection satisfies*

$$S(F_{\widehat{k}}) - S(F_*) \ge \max\{\widehat{\Delta} - 2\varepsilon, \Gamma_K - 2\varepsilon\}, \qquad \Gamma_K = \max_k [S(F_k) - S(F_*)] \ge 0. \tag{15}$$

Indeed, empirical maximization gives $S(F_{\widehat{k}}) \geq \widehat{S}(F_{\widehat{k}}) - \varepsilon \geq \widehat{S}(F_k) - \varepsilon \geq S(F_k) - 2\varepsilon$ for every candidate. Applying the same deviation control to the observed incumbent comparison gives the other term. For candidates constructed independently of n_{sel} selection instances and instance scores in $[0,1]$, a union of Hoeffding bounds yields $\varepsilon = \sqrt{\log(2(K+1)/\delta)/(2n_{\mathrm{sel}})}$ with probability at least $1-\delta$ (Hoeffding, 1963). AUC uses class-conditional concentration; adaptive reuse requires uniform control over the resulting candidate family (Agarwal et al., 2005; Dwork et al., 2015).

Remark 3.3 (From complementary information to retained improvement)**.** For direct log-loss selection with risk-estimation error bound ε_R, define $\mathcal{A}_K = \min_{0\leq k\leq K} c^{-1}\sum_j \mathbb{E}(D_{k,j} - d_j^*)^2$. Then

$$R(F_*) - R(F_{\widehat{k}}) \geq \mathcal{I}_{\mathrm{comp}} + E_* - \min_k E_k - 2\varepsilon_R \geq \mathcal{I}_{\mathrm{comp}} + E_* - \frac{\mathcal{A}_K}{8} - 2\varepsilon_R. \tag{16}$$

The selected predictor improves when the available conditional information and consensus error exceed residual approximation and selection costs. Anchoring fixes this comparison, zero inclusion preserves the reference, and selection retains the strongest supported correction. AP and AUC use the general score bound in Theorem 3.2.

Table 1: Benchmark comparison at PER = LMR = 50%, reported as $\mathrm{mean}_{\mathrm{std}}$. **Bold** and underline denote the best and second-best unrounded means. *AVE* is the average rank over six metrics, with ties averaged.

Data	Metric	LMVCAT	AIMNet	DICNet	LVSL	MTD	SIP	DCSI	RANK	URDF	**CARE**
	1-HL	$0.986_{0.000}$	$0.988_{0.000}$	$0.987_{0.000}$	$0.987_{0.000}$	$\underline{0.988}_{0.000}$	$0.988_{0.000}$	$\mathbf{0.988}_{0.000}$	$0.988_{0.000}$	$\underline{0.988}_{0.000}$	$\mathbf{0.988}_{0.000}$
	1-OE	$0.448_{0.011}$	$0.478_{0.011}$	$0.460_{0.012}$	$0.353_{0.017}$	$0.492_{0.011}$	$0.492_{0.014}$	$\underline{0.520}_{0.017}$	$0.487_{0.016}$	$0.505_{0.015}$	$\mathbf{0.522}_{0.008}$
	1-Cov	$0.720_{0.006}$	$0.766_{0.004}$	$0.726_{0.007}$	$0.719_{0.005}$	$0.754_{0.005}$	$\underline{0.780}_{0.004}$	$0.774_{0.006}$	$0.766_{0.006}$	$0.756_{0.006}$	$\mathbf{0.810}_{0.004}$
COR	1-RL	$0.876_{0.004}$	$0.900_{0.002}$	$0.881_{0.004}$	$0.879_{0.003}$	$0.893_{0.004}$	$\underline{0.908}_{0.003}$	$0.903_{0.003}$	$0.902_{0.004}$	$0.898_{0.003}$	$\mathbf{0.921}_{0.002}$
	AP	$0.378_{0.006}$	$0.404_{0.005}$	$0.381_{0.006}$	$0.311_{0.005}$	$0.410_{0.007}$	$0.414_{0.006}$	$\underline{0.430}_{0.007}$	$0.408_{0.006}$	$0.424_{0.007}$	$\mathbf{0.450}_{0.003}$
	AUC	$0.879_{0.003}$	$0.903_{0.002}$	$0.883_{0.004}$	$0.882_{0.002}$	$0.896_{0.003}$	$\underline{0.910}_{0.002}$	$0.905_{0.003}$	$0.904_{0.003}$	$0.900_{0.003}$	$\mathbf{0.923}_{0.001}$
	AVE	9.5	5.9	8.0	9.5	5.7	3.2	2.4	5.1	4.6	**1.1**
	1-HL	$0.982_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$\underline{0.983}_{0.000}$	$0.983_{0.000}$	$\mathbf{0.983}_{0.000}$
	1-OE	$0.432_{0.006}$	$0.442_{0.006}$	$0.440_{0.009}$	$0.365_{0.006}$	$0.452_{0.007}$	$0.450_{0.006}$	$0.454_{0.007}$	$\underline{0.460}_{0.008}$	$0.438_{0.006}$	$\mathbf{0.496}_{0.004}$
	1-Cov	$0.587_{0.003}$	$0.621_{0.003}$	$0.601_{0.003}$	$0.578_{0.002}$	$0.617_{0.004}$	$0.622_{0.004}$	$0.613_{0.007}$	$\underline{0.627}_{0.003}$	$0.583_{0.007}$	$\mathbf{0.661}_{0.002}$
ESP	1-RL	$0.827_{0.002}$	$0.845_{0.002}$	$0.836_{0.002}$	$0.829_{0.001}$	$0.843_{0.002}$	$\underline{0.847}_{0.002}$	$0.836_{0.003}$	$0.846_{0.002}$	$0.822_{0.003}$	$\mathbf{0.866}_{0.001}$
	AP	$0.293_{0.003}$	$0.306_{0.003}$	$0.300_{0.003}$	$0.266_{0.003}$	$0.309_{0.003}$	$0.309_{0.004}$	$\underline{0.316}_{0.003}$	$0.313_{0.005}$	$0.307_{0.003}$	$\mathbf{0.344}_{0.003}$
	AUC	$0.832_{0.001}$	$0.850_{0.001}$	$0.841_{0.002}$	$0.834_{0.001}$	$0.847_{0.002}$	$0.851_{0.002}$	$0.840_{0.003}$	$\underline{0.851}_{0.001}$	$0.826_{0.003}$	$\mathbf{0.869}_{0.001}$
	AVE	9.0	5.5	6.5	8.7	5.2	3.7	5.3	2.3	7.8	**1.0**
	1-HL	$0.980_{0.000}$	$0.981_{0.000}$	$0.981_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.981_{0.000}$	$\underline{0.981}_{0.000}$	$0.981_{0.000}$	$\underline{0.981}_{0.000}$	$\mathbf{0.981}_{0.000}$
	1-OE	$0.433_{0.009}$	$0.456_{0.008}$	$0.464_{0.008}$	$0.377_{0.007}$	$0.479_{0.007}$	$0.458_{0.005}$	$\underline{0.489}_{0.006}$	$0.474_{0.007}$	$0.483_{0.010}$	$\mathbf{0.508}_{0.007}$
	1-Cov	$0.646_{0.004}$	$0.675_{0.004}$	$0.649_{0.005}$	$0.605_{0.004}$	$0.670_{0.004}$	$\underline{0.678}_{0.003}$	$0.678_{0.008}$	$0.675_{0.006}$	$0.644_{0.004}$	$\mathbf{0.712}_{0.005}$
IAP	1-RL	$0.868_{0.002}$	$0.884_{0.001}$	$0.874_{0.002}$	$0.857_{0.002}$	$0.882_{0.002}$	$\underline{0.886}_{0.001}$	$0.883_{0.004}$	$0.885_{0.002}$	$0.870_{0.002}$	$\mathbf{0.903}_{0.002}$
	AP	$0.313_{0.004}$	$0.329_{0.003}$	$0.326_{0.003}$	$0.262_{0.002}$	$0.340_{0.002}$	$0.331_{0.002}$	$\underline{0.354}_{0.004}$	$0.336_{0.004}$	$0.349_{0.005}$	$\mathbf{0.371}_{0.003}$
	AUC	$0.870_{0.002}$	$0.885_{0.001}$	$0.876_{0.002}$	$0.859_{0.001}$	$0.883_{0.002}$	$\underline{0.887}_{0.001}$	$0.883_{0.004}$	$0.886_{0.002}$	$0.870_{0.002}$	$\mathbf{0.902}_{0.002}$
	AVE	9.0	5.8	6.8	9.8	5.1	4.3	3.2	4.2	5.6	**1.0**
	1-HL	$0.880_{0.004}$	$0.890_{0.001}$	$0.890_{0.001}$	$0.877_{0.001}$	$0.893_{0.000}$	$0.890_{0.001}$	$\underline{0.894}_{0.002}$	$0.890_{0.001}$	$0.891_{0.001}$	$\mathbf{0.898}_{0.001}$
	1-OE	$0.639_{0.009}$	$0.646_{0.009}$	$0.647_{0.010}$	$0.609_{0.007}$	$0.667_{0.006}$	$0.654_{0.007}$	$\underline{0.670}_{0.007}$	$0.650_{0.007}$	$0.662_{0.006}$	$\mathbf{0.699}_{0.005}$
	1-Cov	$0.665_{0.002}$	$0.673_{0.003}$	$0.662_{0.004}$	$0.624_{0.002}$	$0.681_{0.002}$	$0.668_{0.006}$	$\underline{0.685}_{0.004}$	$0.645_{0.015}$	$0.660_{0.009}$	$\mathbf{0.698}_{0.001}$
MIR	1-RL	$0.862_{0.003}$	$0.874_{0.002}$	$0.869_{0.003}$	$0.847_{0.001}$	$0.878_{0.001}$	$0.873_{0.002}$	$\underline{0.879}_{0.003}$	$0.856_{0.010}$	$0.863_{0.006}$	$\mathbf{0.892}_{0.001}$
	AP	$0.589_{0.004}$	$0.599_{0.003}$	$0.595_{0.007}$	$0.548_{0.003}$	$0.614_{0.004}$	$0.603_{0.005}$	$\underline{0.618}_{0.006}$	$0.592_{0.007}$	$0.603_{0.006}$	$\mathbf{0.645}_{0.005}$
	AUC	$0.852_{0.003}$	$0.861_{0.001}$	$0.855_{0.002}$	$0.839_{0.001}$	$0.864_{0.001}$	$0.858_{0.002}$	$\underline{0.864}_{0.003}$	$0.842_{0.010}$	$0.848_{0.006}$	$\mathbf{0.876}_{0.001}$
	AVE	8.0	5.7	6.6	10.0	3.0	5.2	2.0	7.7	5.8	**1.0**
	1-HL	$0.940_{0.002}$	$0.948_{0.001}$	$0.948_{0.001}$	$0.935_{0.001}$	$0.949_{0.001}$	$0.948_{0.001}$	$\underline{0.949}_{0.001}$	$0.947_{0.001}$	$0.946_{0.001}$	$\mathbf{0.951}_{0.001}$
	1-OE	$0.604_{0.016}$	$0.619_{0.015}$	$0.601_{0.011}$	$0.450_{0.008}$	$\underline{0.627}_{0.011}$	$0.626_{0.009}$	$0.621_{0.014}$	$0.609_{0.011}$	$0.603_{0.015}$	$\mathbf{0.664}_{0.008}$
	1-Cov	$0.796_{0.008}$	$0.806_{0.006}$	$0.794_{0.006}$	$0.759_{0.006}$	$0.812_{0.006}$	$0.809_{0.006}$	$\underline{0.814}_{0.006}$	$0.807_{0.008}$	$0.794_{0.007}$	$\mathbf{0.828}_{0.002}$
OBJ	1-RL	$0.878_{0.006}$	$0.888_{0.005}$	$0.876_{0.004}$	$0.850_{0.004}$	$0.890_{0.005}$	$0.889_{0.004}$	$\underline{0.892}_{0.004}$	$0.885_{0.006}$	$0.878_{0.005}$	$\mathbf{0.904}_{0.001}$
	AP	$0.630_{0.012}$	$0.639_{0.010}$	$0.627_{0.009}$	$0.537_{0.008}$	$\underline{0.649}_{0.009}$	$0.649_{0.008}$	$\underline{0.649}_{0.011}$	$0.636_{0.011}$	$0.629_{0.009}$	$\mathbf{0.679}_{0.004}$
	AUC	$0.888_{0.006}$	$0.897_{0.004}$	$0.886_{0.004}$	$0.864_{0.004}$	$0.900_{0.005}$	$0.898_{0.004}$	$\underline{0.902}_{0.004}$	$0.895_{0.006}$	$0.888_{0.004}$	$\mathbf{0.912}_{0.001}$
	AVE	7.3	5.2	8.2	10.0	2.8	3.8	2.4	6.0	8.2	**1.0**
	1-HL	$0.915_{0.005}$	$0.932_{0.001}$	$0.931_{0.000}$	$0.928_{0.001}$	$\underline{0.933}_{0.001}$	$0.932_{0.001}$	$\underline{0.933}_{0.001}$	$0.932_{0.001}$	$0.930_{0.002}$	$\mathbf{0.936}_{0.000}$
	1-OE	$0.433_{0.016}$	$0.462_{0.010}$	$0.443_{0.007}$	$0.418_{0.008}$	$0.474_{0.008}$	$0.468_{0.008}$	$\underline{0.476}_{0.014}$	$0.467_{0.009}$	$0.454_{0.017}$	$\mathbf{0.508}_{0.001}$
	1-Cov	$0.759_{0.006}$	$0.781_{0.007}$	$0.749_{0.003}$	$0.738_{0.003}$	$0.790_{0.006}$	$0.778_{0.004}$	$\underline{0.791}_{0.006}$	$0.785_{0.008}$	$0.752_{0.006}$	$\mathbf{0.812}_{0.006}$
PAS	1-RL	$0.808_{0.006}$	$0.830_{0.006}$	$0.804_{0.002}$	$0.797_{0.002}$	$0.836_{0.006}$	$0.828_{0.004}$	$\underline{0.838}_{0.005}$	$0.834_{0.007}$	$0.804_{0.006}$	$\mathbf{0.859}_{0.005}$
	AP	$0.524_{0.009}$	$0.548_{0.007}$	$0.517_{0.004}$	$0.486_{0.005}$	$0.562_{0.005}$	$0.552_{0.006}$	$\underline{0.568}_{0.008}$	$0.557_{0.006}$	$0.537_{0.010}$	$\mathbf{0.599}_{0.004}$
	AUC	$0.830_{0.006}$	$0.851_{0.005}$	$0.827_{0.002}$	$0.823_{0.002}$	$\underline{0.855}_{0.006}$	$0.848_{0.005}$	$0.854_{0.005}$	$0.854_{0.006}$	$0.817_{0.006}$	$\mathbf{0.875}_{0.005}$
	AVE	8.0	5.5	8.3	9.7	2.8	5.3	2.2	4.2	8.0	**1.0**

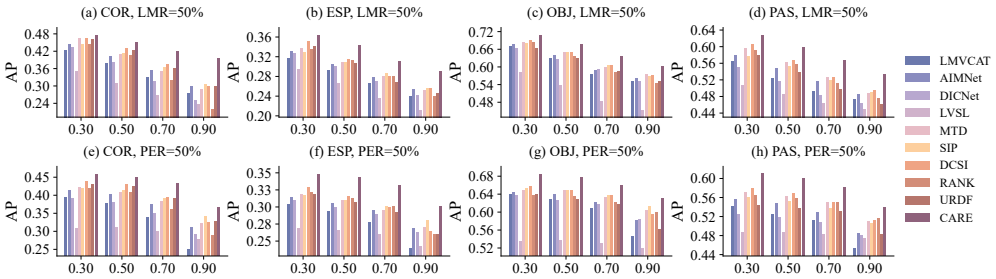


Figure 3: AP as PER varies in (a)–(d) and LMR varies in (e)–(h).

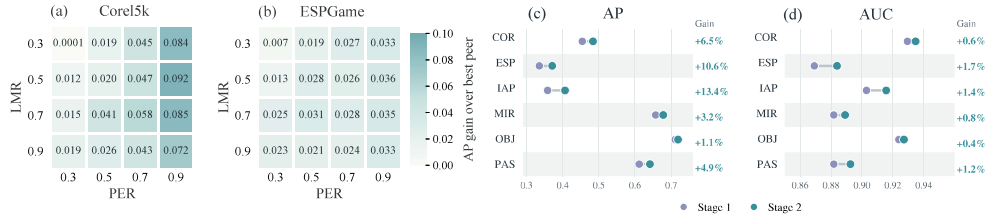


Figure 4: Complementary gains across missingness settings and learning stages. (a)–(b) show absolute AP gains over the strongest competitor on Corel5k and ESPGame. (c)–(d) compare stage-wise AP and AUC on six datasets at PER = LMR = 30%; annotations report relative gains.

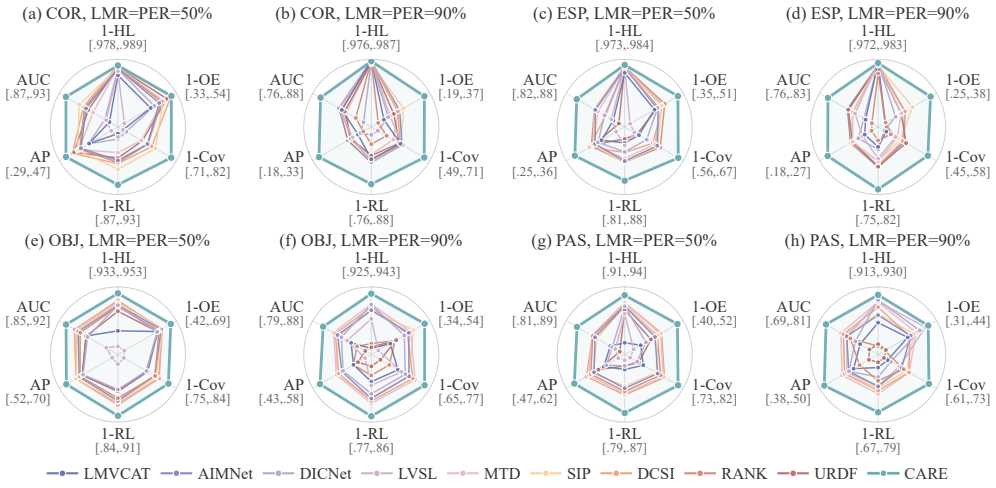


Figure 5: Six-metric profiles at PER = LMR = 50% and 90%. Axis-specific ranges are shown beside the metric labels.

4 Experiments

4.1 Evaluation Protocol and Benchmark Results

Datasets. We evaluate Corel5k (Duygulu et al., 2002), ESPGame (von Ahn & Dabbish, 2004), IAPRTC12 (Grubinger et al., 2006), Mirflickr (Huiskes & Lew, 2008), Pascal07 (Everingham et al., 2010), and OBJECT (Hao et al., 2024). The benchmarks contain 4,999 to 25,000 instances and 20 to 291 labels. Five use six descriptors covering color, local appearance, and scene structure. OBJECT uses five handcrafted views. Table 4 gives the feature dimensions, sample counts, and label spaces.

Comparison methods. The nine baselines cover complementary approaches to incomplete evidence. DICNet and LMVCAT learn cross-view agreement and label-guided interaction (Liu et al., 2023b;c). AIMNet imputes missing embeddings, while LVSL learns view-specific labels (Liu et al., 2024a; Zhao et al., 2023). MTD and DCSI separate shared and specific information, SIP learns semantic invariance, and RANK and URDF estimate reliable view contributions (Liu et al., 2023a;

Wen et al., 2026b; Liu et al., 2024b; 2025; Wen et al., 2026a). Methods use matched data partitions and observation masks. LVSL receives training-mean imputation for missing features and zero-filled missing label entries, to satisfy its complete-matrix input format.

Implementation details. PER removes complete view vectors and LMR hides positive and negative label coordinates. Every sample retains at least one view. The benchmark comparison uses PER $=$ LMR $= 50\%$, and robustness studies vary one rate over $\{30\%, 50\%, 70\%, 90\%\}$ while fixing the other at 50%. We use a 7:1:2 training, validation, and test split, Adam (Kingma & Ba, 2015) with learning rate 10^{-4}, and batches of 256. The main comparison uses five masking and splitting repetitions. Across all experiments, consensus learning runs for 50 epochs and residual refinement for 25 epochs. The consensus checkpoint and residual candidates are selected by the mean of validation AP and AUC. Test labels remain fixed across stages.

Evaluation metrics. AP measures relevant-label ranking within each instance, and AUC measures positive–negative discrimination. We also report 1-HL, 1-OE, 1-Cov, and 1-RL for coordinate accuracy, the first-ranked label, coverage, and pairwise ordering. All metrics favor higher scores.

CARE achieves the highest AP and AUC on all six datasets in Table 1, with AP margins of 1.7 to 3.1 percentage points over the strongest competitor. Its leading average ranks also reflect gains in coverage, pairwise ordering, and top-label accuracy.

Figure 3 compares AP as view and label missingness vary separately. Figure 4(a)–(b) maps AP gains across their joint variation on Corel5k and ESPGame. CARE exceeds the strongest competitor in all 32 displayed settings. Corel5k shows larger margins at high PER, reaching 0.092 at PER $= 90\%$ and LMR $= 50\%$. ESPGame exhibits smaller, more even gains. Figure 5 complements this comparison with six-metric profiles at matched missingness rates.

Figure 4(c)–(d) isolates residual refinement at PER $=$ LMR $= 30\%$ over three folds. Relative AP gains range from 1.08% to 13.41%, and AUC gains from 0.36% to 1.69%. IAPRTC12 and ESPGame gain most in AP; OBJECT retains a smaller correction to its stronger consensus. All seven metrics improve across the six datasets in Table 5, which reports paired-fold variation. Absolute AP gains range from 0.00767 to 0.04804.

4.2 Ablation Analysis and Optimization Behavior

Table 2 ablates consensus anchoring A, learned gating G, and incumbent inclusion I at PER $=$ LMR $= 50\%$ using matched splits and masks. The variants unfreeze consensus, fix gates, or exclude the incumbent. CARE leads in AUC and AP across three datasets, tying in Pascal07 AP. Anchoring contributes most, while gating and incumbent inclusion provide further gains.

Table 2: Component ablation.

Design			Pascal07		OBJECT		ESPGame	
A	G	I	AP	AUC	AP	AUC	AP	AUC
–	✓	✓	.587	.863	.653	.861	.339	.847
✓	–	✓	.591	.865	.669	.879	.341	.862
✓	✓	–	**.599**	.870	.671	.882	.343	.863
✓	✓	✓	**.599**	**.875**	**.679**	**.912**	**.344**	**.869**

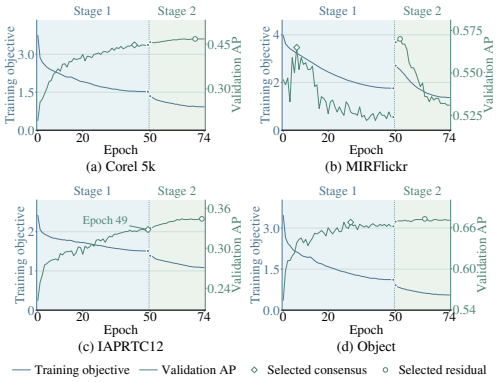


Figure 6: Training objectives and validation AP across both stages.

Figure 6 shows the optimization behavior of CARE across four datasets. Training objectives decrease and gradually flatten within each stage. Consensus learning establishes the shared predictor, after which optimization is confined to residual corrections around the fixed reference. The selected residual checkpoints achieve higher validation AP than the archived consensus in the displayed runs. The stage boundary therefore preserves the established prediction function while allowing complementary learning to improve its output. Selection retains a complete predictor from this refinement trajectory. The archived consensus remains available throughout, so refinement is assessed by the predictive value added to an established solution.

Figure 7 examines the semantic alignment weight β and residual gating weight λ_{gate} on two datasets. Semantic alignment shapes the consensus representation, while the gating penalty controls the contribution of residual corrections. AP remains broadly stable over the selected ranges, with extended regions of comparable performance. Neighboring combinations give similar scores in both datasets. The surfaces indicate that useful refinements are available across a range of regularization strengths. Since the consensus is fixed during the second stage, the gating penalty adjusts complementary corrections against the same preserved predictor.

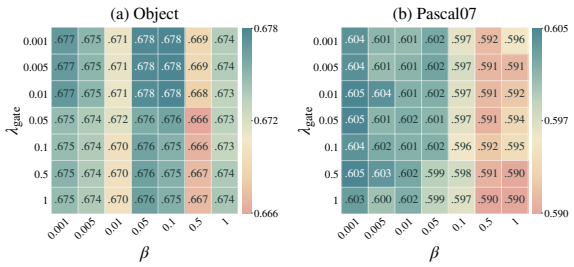


Figure 7: AP sensitivity to semantic alignment weight β and residual gating weight λ_{gate}.

4.3 Chest Radiograph Disease Screening

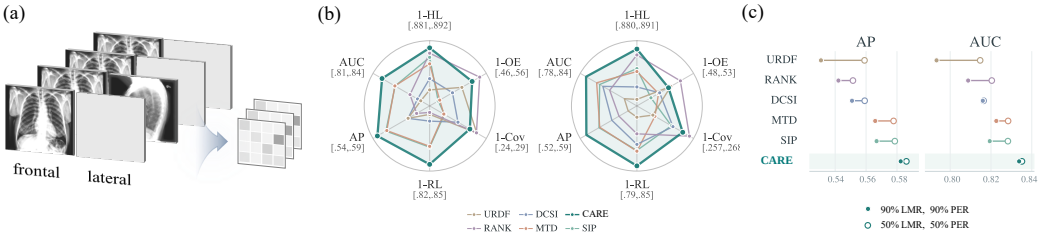


Figure 8: Chest radiograph screening. (a) shows paired projections, (b) six-metric profiles, and (c) AP/AUC comparisons. Open and filled markers denote nominal PER = LMR of 50% and 90%.

Frontal and lateral chest radiographs expose complementary anatomy. Portable examinations may provide one projection, and reports may omit pathology labels. CheXpert-MV retains 31,413 frontal–lateral pairs from CheXpert (Irvin et al., 2019). Resized images pass through an ImageNet-pretrained DenseNet-121 to obtain 1,024-dimensional projection embeddings (Huang et al., 2017). The 14 observations use U-Zeros to map uncertain labels to zero. Separate encoders process the projection embeddings. PER removes whole embeddings, and LMR hides supervised coordinates while preserving evaluation labels. Both stages use availability-aware aggregation, leave-one-view-out prediction, and confidence-weighted late fusion.

Each projection supplies label-specific corrections, allowing its contribution to differ between pleural effusion and cardiomegaly. Figure 8 compares CARE with URDF, RANK, DCSI, MTD, and SIP at nominal 50% and 90% missingness. CARE occupies the highest AP and AUC positions in both displayed comparisons, supporting its applicability to paired radiographs with incomplete observations. The advantage appears in both relevant-label ranking and positive–negative discrimination. The six-metric profiles place these changes alongside label coverage and top-label accuracy. Preserving consensus gives each available projection a common prediction reference, while the residual branch can exploit anatomical information that contributes beyond the shared prediction.

5 Conclusion

CARE introduces a two-stage framework that establishes consensus before learning complementary corrections around the preserved predictor. The stage boundary makes additional information comparable against a stable prediction function throughout optimization and selection. Conditional information identifies what remains useful beyond consensus, and the approximation and selection bounds specify when this evidence becomes a retained gain. Incumbent-inclusive selection preserves the option to deploy the consensus. Paired benchmark evaluations quantify the contribution of the corrections, while chest radiograph screening demonstrates the framework in a clinical imaging setting. Complementarity is measured as predictive improvement over the preserved consensus.

AI USE STATEMENT

Generative AI assisted language editing, typesetting and visual review.

REFERENCES

Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, and Dan Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2019.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain Separation Networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018.

Pinar Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, volume 2353 of *Lecture Notes in Computer Science*, pp. 97–112, 2002.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving Statistical Validity in Adaptive Data Analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pp. 117–126, 2015.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *Proceedings of the International Workshop OntoImage: Language Resources for Content-Based Image Retrieval*, pp. 13–23, 2006.

Pingting Hao, Kunpeng Liu, and Wanfu Gao. Anchor-guided global view reconstruction for multi-view multi-label feature selection. *Information Sciences*, 679:121124, 2024.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

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

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, July 2017.

Mark J. Huiskes and Michael S. Lew. The MIR Flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, 2008.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 2015.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Xiang Li and Songcan Chen. A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5918–5932, 2022.

Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2019.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task Learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18878–18890, 2021.

Chengliang Liu, Jie Wen, Yabo Liu, Chao Huang, Zhihao Wu, Xiaoling Luo, and Yong Xu. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 32387–32400, 2023a.

Chengliang Liu, Jie Wen, Xiaoling Luo, Chao Huang, Zhihao Wu, and Yong Xu. DICNet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8807–8815, 2023b.

Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view- and category-aware transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8816–8824, 2023c.

Chengliang Liu, Jinlong Jia, Jie Wen, Yabo Liu, Xiaoling Luo, Chao Huang, and Yong Xu. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13864–13872, 2024a.

Chengliang Liu, Gehui Xu, Jie Wen, Yabo Liu, Chao Huang, and Yong Xu. Partial multi-view multi-label classification via semantic invariance learning and prototype modeling. In *Proceedings of the International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32253–32267, 2024b.

Chengliang Liu, Jie Wen, Yong Xu, Bob Zhang, Liqiang Nie, and Min Zhang. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4940–4956, 2025.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, pp. 148–188. Cambridge University Press, 1989.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247, June 2022.

Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Advances in Neural Information Processing Systems*, volume 32, pp. 15692–15703, 2019.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized Multimodal ELBO. In *International Conference on Learning Representations*, 2021.

Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2703–2709, 2018.

Xin Tan, Ce Zhao, Chengliang Liu, Jie Wen, and Zhanyan Tang. A two-stage information extraction network for incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15249–15257, 2024.

Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, 2004.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, June 2020.

Jie Wen, Jiang Long, Xiaohuan Lu, Chengliang Liu, Xiaozhao Fang, and Yong Xu. Partial multiview incomplete multilabel learning via uncertainty-driven reliable dynamic fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):236–250, 2026a.

Jie Wen, Lian Zhao, Xiaohuan Lu, Chengliang Liu, Li Shen, Chao Huang, and Yong Xu. Disentangling consistent and specific information for double incomplete multi-view multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):7307–7320, 2026b.

Mike Wu and Noah Goodman. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xu Yan, Jun Yin, and Jie Wen. Incomplete multi-view multi-label learning via disentangled representation and label semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30722–30731, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Non-aligned multi-view multi-label classification via learning view-specific labels. *IEEE Transactions on Multimedia*, 25:7235–7247, 2023.

A Mathematical Analysis of the Two-Stage Framework

The derivation follows the gain from conditional label information to residual approximation and model selection. Let $z = (\mathbf{x}, \mathbf{m}, \mathbf{y}, \mathbf{u})$ denote an observation, and write the archived consensus state as $s_* = (\theta_*, b_*)$, where b_* contains normalization statistics and other persistent evaluation buffers. The consensus logits and augmented logits are

$$F_*(z) = F(\mathbf{x}, \mathbf{m}; s_*), \qquad F_\phi(z) = F_*(z) + D_\phi(z), \qquad D_\phi(z) = \sum_{v=1}^{V} m_v g_v \delta_v(z). \tag{A.1}$$

Here $g_v = \sigma(a_v)$ and ϕ collects specific-encoder, residual-head, and gate parameters. All equalities between functions below refer to a common deterministic evaluation procedure. Optimization statements concern the stated objective and gradient-descent update. Statistical statements concern a specified population score and its estimator.

A.1 Conditional Information and the Population Value of a Residual

All expectations in this subsection condition on the fitted consensus state. The variable Z contains the frozen consensus representation and observation mask. The variable U contains the observed input views and is common to every candidate. Each candidate's learned specific features are measurable functions of these inputs. Let $p_j = \Pr(Y_j = 1 \mid Z)$, $p_j^+ = \Pr(Y_j = 1 \mid Z, U)$, $q_{*,j} = \sigma(F_{*,j})$, and $q_{\phi,j} = \sigma(F_{\phi,j})$. For $p, q \in (0, 1)$ define

$$\begin{aligned} h(p) &= -p \log p - (1-p) \log(1-p), \\ D_{\mathrm{B}}(p \| q) &= p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \\ \ell(q, y) &= -y \log q - (1-y) \log(1-q). \end{aligned} \tag{A.2}$$

The conditional expected loss decomposes exactly as

$$\mathbb{E}[\ell(q, Y) \mid \mathcal{A}] = -p_{\mathcal{A}} \log q - (1 - p_{\mathcal{A}}) \log(1-q) = h(p_{\mathcal{A}}) + D_{\mathrm{B}}(p_{\mathcal{A}} \| q), \tag{A.3}$$

where q is measurable with respect to $\mathcal{A}$ and $p_{\mathcal{A}} = \Pr(Y = 1 \mid \mathcal{A})$. This proper-scoring identity identifies both the Bayes risk and the approximation cost (Gneiting & Raftery, 2007). Define

$$\begin{aligned} R(F) &= \frac{1}{c} \sum_{j=1}^{c} \mathbb{E}\, \ell(\sigma(F_j), Y_j), \\ E_* &= \frac{1}{c} \sum_j \mathbb{E} D_{\mathrm{B}}(p_j \| q_{*,j}), \qquad E_\phi = \frac{1}{c} \sum_j \mathbb{E} D_{\mathrm{B}}(p_j^+ \| q_{\phi,j}). \end{aligned} \tag{A.4}$$

Applying Equation A.3 to the two information sets gives

$$R(F_*) = \frac{1}{c} \sum_j H(Y_j \mid Z) + E_*, \qquad R(F_\phi) = \frac{1}{c} \sum_j H(Y_j \mid Z, U) + E_\phi. \tag{A.5}$$

Consequently,

$$\begin{aligned} R(F_*) - R(F_\phi) &= \frac{1}{c} \sum_j \{H(Y_j \mid Z) - H(Y_j \mid Z, U)\} + E_* - E_\phi \\ &= \mathcal{I}_{\mathrm{comp}} + E_* - E_\phi, \qquad \mathcal{I}_{\mathrm{comp}} = \frac{1}{c} \sum_j I(Y_j; U \mid Z). \end{aligned} \tag{A.6}$$

This proves the identity in Proposition 3.1. The information term also admits a divergence representation. Since $\mathbb{E}[p_j^+ \mid Z] = p_j$,

$$\begin{aligned} \mathbb{E} D_{\mathrm{B}}(p_j^+ \| p_j) &= \mathbb{E}\left[p_j^+ \log p_j^+ + (1 - p_j^+) \log(1 - p_j^+)\right] \\ &\quad - \mathbb{E}\left[p_j \log p_j + (1 - p_j) \log(1 - p_j)\right] \\ &= H(Y_j \mid Z) - H(Y_j \mid Z, U) = I(Y_j; U \mid Z). \end{aligned} \tag{A.7}$$

The term vanishes exactly when the specific information leaves the conditional label probability unchanged almost surely. A view can therefore contain information about the label while adding no information beyond the consensus. For chest radiographs, the relevant quantity concerns a finding resolved by the retained projection after conditioning on the consensus prediction features.

Let $\mathcal{D}$ denote the actual gated additive residual family, with $0 \in \mathcal{D}$. For its approximation floor $E_{\mathcal{D}} = \inf_{d\in\mathcal{D}} c^{-1} \sum_j \mathbb{E} D_{\mathrm{B}}(p_j^+ \| \sigma(F_{*,j} + d_j))$, the best attainable gain is

$$\Gamma_{\mathcal{D}} := R(F_*) - \inf_{d\in\mathcal{D}} R(F_* + d) = \mathcal{I}_{\mathrm{comp}} + E_* - E_{\mathcal{D}} \geq 0. \tag{A.8}$$

The inequality follows from $0 \in \mathcal{D}$. The zero correction satisfies

$$E_0 = \frac{1}{c} \sum_j \mathbb{E} D_{\mathrm{B}}(p_j^+ \| q_{*,j}) = \mathcal{I}_{\mathrm{comp}} + E_*. \tag{A.9}$$

For unrestricted measurable corrections, the Bayes correction is

$$d_j^*(Z,U) = \log \frac{p_j^+(Z,U)}{1 - p_j^+(Z,U)} - F_{*,j}(Z), \qquad R(F_* + d^*) = \frac{1}{c} \sum_j H(Y_j \mid Z, U). \tag{A.10}$$

The actual architecture enters through $E_{\mathcal{D}}$. When d^* belongs to that family, its approximation floor is zero; otherwise Equation A.8 retains the corresponding positive approximation cost. Attainable gain depends on conditional information and residual expressiveness.

The approximation cost has a useful logit-scale upper bound. Let $a_j^* = \mathrm{logit}(p_j^+)$ and $A(a) = \log(1 + e^a)$. Since $A'(a_j^*) = p_j^+$,

$$D_{\mathrm{B}}(p_j^+ \| \sigma(a)) = A(a) - A(a_j^*) - A'(a_j^*)(a - a_j^*) \leq \frac{1}{8}(a - a_j^*)^2. \tag{A.11}$$

The last inequality uses $A''(a) \leq 1/4$. For any realizable d with finite squared approximation error,

$$R(F_*) - R(F_* + d) \geq \mathcal{I}_{\mathrm{comp}} + E_* - \frac{1}{8c} \sum_j \mathbb{E}\big[d_j - d_j^*\big]^2. \tag{A.12}$$

A residual that approximates the conditional log-odds correction sufficiently closely therefore improves population log loss. This criterion specifies a representational target for the private pathway. It also explains why separation alone is insufficient to establish a gain: the extracted features must support an accurate correction relative to the fixed predictor.

Observation patterns refine the same identity. Write $Z = (H_*, M)$ and let $\omega_m = \Pr(M = m)$. Conditioning on the finitely many masks gives

$$\mathcal{I}_{\mathrm{comp}} = \frac{1}{c} \sum_j \sum_m \omega_m I(Y_j; U \mid H_*, M = m). \tag{A.13}$$

The residual can have different conditional value under different acquisition patterns. The availability mask determines which terms can be evaluated, while the population gain averages over the distribution of those patterns.

A.2 Gaussian Aggregation and Deterministic Moment Identities

For a latent vector $h \in \mathbb{R}^d$, let $p_0(h) = \mathcal{N}(h; 0, I)$ and $q_v(h) = \mathcal{N}(h; \mu_v, \Sigma_v)$, with positive diagonal Σ_v. The masked product of experts is

$$q_{\mathrm{s}}(h) = Z_{\mathbf{m}}^{-1} p_0(h) \prod_{v=1}^{V} q_v(h)^{m_v}, \qquad Z_{\mathbf{m}} = \int p_0(h) \prod_v q_v(h)^{m_v}\, dh. \tag{A.14}$$

Expanding the log density gives

$$\log q_{\mathrm{s}}(h) = -\tfrac{1}{2} h^T h - \tfrac{1}{2} \sum_v m_v (h - \mu_v)^T \Sigma_v^{-1} (h - \mu_v) + C. \tag{A.15}$$

Collect the quadratic and linear terms using

$$\Lambda_{\mathrm{s}} = I + \sum_v m_v \Sigma_v^{-1}, \qquad b_{\mathrm{s}} = \sum_v m_v \Sigma_v^{-1} \mu_v. \tag{A.16}$$

Completing the square yields

$$-\tfrac{1}{2} h^T \Lambda_{\mathrm{s}} h + h^T b_{\mathrm{s}} = -\tfrac{1}{2}(h - \Lambda_{\mathrm{s}}^{-1} b_{\mathrm{s}})^T \Lambda_{\mathrm{s}} (h - \Lambda_{\mathrm{s}}^{-1} b_{\mathrm{s}}) + \tfrac{1}{2} b_{\mathrm{s}}^T \Lambda_{\mathrm{s}}^{-1} b_{\mathrm{s}}. \tag{A.17}$$

The completed quadratic identifies the mean and covariance of the fused Gaussian,

$$\Sigma_{\mathrm{s}} = \Lambda_{\mathrm{s}}^{-1}, \qquad \mu_{\mathrm{s}} = \Lambda_{\mathrm{s}}^{-1} b_{\mathrm{s}}, \qquad q_{\mathrm{s}} = \mathcal{N}(\mu_{\mathrm{s}}, \Sigma_{\mathrm{s}}). \tag{A.18}$$

The precision is positive definite because $\Lambda_{\mathrm{s}} \succeq I$. In coordinate k,

$$\sigma_{\mathrm{s},k}^2 = \left(1 + \sum_v m_v \sigma_{v,k}^{-2}\right)^{-1}, \qquad \mu_{\mathrm{s},k} = \sigma_{\mathrm{s},k}^2 \sum_v m_v \sigma_{v,k}^{-2} \mu_{v,k}. \tag{A.19}$$

If $m_v = 0$, the corresponding precision and linear terms vanish. For observed-view sets $A \subseteq B$, the covariance matrices satisfy the following order relation

$$\Lambda_B = \Lambda_A + \sum_{v \in B \setminus A} \Sigma_v^{-1} \succeq \Lambda_A, \qquad \Sigma_B \preceq \Sigma_A. \tag{A.20}$$

Adding an observed expert increases precision and contracts the covariance of the Gaussian product.

Closed-form interaction moments provide deterministic inputs to a scorer. For independent Gaussian vectors $H \sim \mathcal{N}(\mu, \Sigma)$ and $L \sim \mathcal{N}(\nu, T)$,

$$\mathbb{E}[HL^T] = \mu\nu^T, \qquad \mathbb{E}[\|H - L\|^2] = \|\mu - \nu\|^2 + \mathrm{tr}(\Sigma + T). \tag{A.21}$$

Writing $H = \mu + \epsilon$ and $L = \nu + \zeta$, with independent centered ϵ and ζ, gives

$$H^T L = \mu^T \nu + \mu^T \zeta + \nu^T \epsilon + \epsilon^T \zeta. \tag{A.22}$$

The three centered terms have zero pairwise covariance, so

$$\mathrm{Var}(H^T L) = \mu^T T \mu + \nu^T \Sigma \nu + \mathrm{tr}(\Sigma T). \tag{A.23}$$

The moment identities give deterministic features for independent latent and prototype distributions. A fixed scorer built from these features satisfies the deterministic evaluation condition.

A.3 Exact Anchoring and Stability of the Prediction Function

Assumption A.1. The consensus evaluation state s_* is fixed during residual learning. The evaluation path is deterministic. All residual output weights and biases start at zero.

A linear last layer makes the initialization condition explicit. For moment features $h_v(z; \psi_v)$, write

$$\delta_v(z) = W h_v(z; \psi_v) + b, \qquad W^{(0)} = 0, \quad b^{(0)} = 0. \tag{A.24}$$

The same argument applies to separate view-specific output heads.

Proposition A.2. *Under Assumption A.1, $F_{\phi^{(0)}}(z) = F_*(z)$ for every input and mask. At every later residual-training step, the consensus component remains F_*.*

Proof. For each view, $\delta_v^{(0)} = 0 h_v + b^{(0)} = 0$. Substitution into Equation A.1 gives

$$D_{\phi^{(0)}}(z) = \sum_v m_v \sigma(a_v^{(0)}) 0 = 0, \qquad \sigma(F_{\phi^{(0)}}(z)) = \sigma(F_*(z)). \tag{A.25}$$

The identity holds for every gate value. For later steps, $s_t = s_*$ implies $F(\mathbf{x}, \mathbf{m}; s_t) = F(\mathbf{x}, \mathbf{m}; s_*)$. Only $D_{\phi^{(t)}}$ changes. □

The state condition includes normalization buffers. With coefficient κ, its running mean evolves as

$$b_{t+1} = (1-\kappa)b_t + \kappa\bar{h}_t, \qquad b_T - b_0 = \kappa \sum_{t=0}^{T-1} (1-\kappa)^{T-1-t}(\bar{h}_t - b_0). \tag{A.26}$$

Freezing parameter gradients alone leaves this recurrence active. Fixing evaluation buffers sets $b_T - b_0 = 0$. If the evaluation map is locally Lipschitz in its state, a perturbation has the bound

$$\|F(z;\theta_t,b_t) - F(z;\theta_*,b_*)\|_\infty \le L_\theta\|\theta_t - \theta_*\| + L_b\|b_t - b_*\|. \tag{A.27}$$

CARE fixes both state components. The bound $0 < \sigma'(x) \le 1/4$ controls the probability change,

$$\|\sigma(F_*(z) + D_\phi(z)) - \sigma(F_*(z))\|_\infty \le \tfrac{1}{4}\|D_\phi(z)\|_\infty \le \tfrac{1}{4}\sum_v m_v g_v \|\delta_v(z)\|_\infty. \tag{A.28}$$

The bound relates each gate's exposure to its possible effect on predicted probabilities. Functional equality uses finite intermediate values and the same deterministic evaluation computation.

For a fixed correction d, shared-logit drift e changes the same prediction by

$$\begin{aligned}\mathcal{L}_{\mathrm{cls}}(F_* + d + e) - \mathcal{L}_{\mathrm{cls}}(F_* + d) &= \langle \nabla\mathcal{L}_{\mathrm{cls}}(F_* + d), e\rangle \\ &\quad + \int_0^1 (1-t)e^T\nabla^2\mathcal{L}_{\mathrm{cls}}(F_* + d + te)e\,dt.\end{aligned} \tag{A.29}$$

Positive semidefinite curvature proves Equation 5. For fixed positive class weights,

$$\begin{aligned}\langle \nabla\mathcal{L}_{\mathrm{cls}}(F_* + d), e\rangle &\le \mathcal{L}_{\mathrm{cls}}(F_* + d + e) - \mathcal{L}_{\mathrm{cls}}(F_* + d) \\ &\le \langle \nabla\mathcal{L}_{\mathrm{cls}}(F_* + d), e\rangle + \frac{1}{8|\mathcal{O}|}\sum_{ij} u_{ij}q_{ij}e_{ij}^2.\end{aligned} \tag{A.30}$$

The lower bound certifies a cost when drift opposes the local descent direction. Anchoring sets both terms to zero while retaining the residual direction. This controlled comparison isolates the effect of moving the reference for the same correction.

Let $C(\theta)$ be the consensus objective and $X(\theta)$ the terms that couple it to a specific pathway. At a fixed parameter point write $g_c = \nabla C(\theta)$ and $g_x = \nabla X(\theta)$. Assume C has an L-Lipschitz gradient on the update segment. For $d = g_c + g_x$ and step size $\eta > 0$,

$$C(\theta - \eta d) - C(\theta) = -\eta\langle g_c, d\rangle + r, \qquad r = -\eta\int_0^1 \langle \nabla C(\theta - t\eta d) - g_c, d\rangle\,dt. \tag{A.31}$$

Cauchy–Schwarz and gradient Lipschitzness yield

$$|r| \le \eta\int_0^1 Lt\eta\|d\|^2\,dt = \tfrac{1}{2}L\eta^2\|d\|^2. \tag{A.32}$$

Put $A = \|g_c\|^2 + \langle g_c, g_x\rangle$. Then

$$-\eta A - \tfrac{1}{2}L\eta^2\|d\|^2 \le C(\theta - \eta d) - C(\theta) \le -\eta A + \tfrac{1}{2}L\eta^2\|d\|^2. \tag{A.33}$$

If $A < 0$, every $0 < \eta < -2A/(L\|d\|^2)$ gives a strict increase in consensus loss when $L\|d\|^2 > 0$. If $A > 0$, every $0 < \eta < 2A/(L\|d\|^2)$ gives a strict decrease. At $A = 0$, the remainder determines the sign. With $L = 0$, the linear term determines the change exactly.

Under frozen-consensus augmentation the update to θ is zero, so

$$\theta_{t+1} = \theta_t = \theta_*, \qquad C(\theta_{t+1}) - C(\theta_t) = 0. \tag{A.34}$$

Freezing removes the shared-parameter update from residual optimization. Let F_J be a joint predictor and F_J^{s} its internal shared pathway. For any scalar score S,

$$S(F_J) - S(F_*) = \big[S(F_J^{\mathrm{s}}) - S(F_*)\big] + \big[S(F_J) - S(F_J^{\mathrm{s}})\big] \tag{A.35}$$

The two terms isolate the shared-pathway change and the added contribution within the joint model. The gradient argument supplies a local explanation for the former.

A.4 Weighted Classification Gradients and Residual Descent

Let $\mathcal{O} = \{(i,j) : u_{ij} = 1\}$, $Z = |\mathcal{O}| > 0$, and $q_{ij} = w_j^+ y_{ij} + w_j^-(1 - y_{ij})$, with fixed positive class weights. For binary y_{ij}, the weighted loss can be written in logit form as

$$\mathcal{L}_{\mathrm{cls}}(\ell) = \frac{1}{Z} \sum_{(i,j)\in\mathcal{O}} q_{ij} \big[\log(1 + e^{\ell_{ij}}) - y_{ij}\ell_{ij}\big]. \tag{A.36}$$

Differentiating with respect to the logits gives the gradient and diagonal curvature,

$$r_{ij} = \frac{\partial \mathcal{L}_{\mathrm{cls}}}{\partial \ell_{ij}} = \frac{u_{ij}q_{ij}}{Z}(p_{ij} - y_{ij}), \qquad H_{ij,ij} = \frac{u_{ij}q_{ij}}{Z} p_{ij}(1 - p_{ij}). \tag{A.37}$$

Off-diagonal entries $H_{ij,kl}$ vanish when $(i,j) \neq (k,l)$. Thus $0 \preceq H \preceq \mathrm{diag}(u_{ij}q_{ij}/4Z)$. For a logit correction d, Taylor's formula with integral remainder gives

$$\mathcal{L}_{\mathrm{cls}}(\ell + d) - \mathcal{L}_{\mathrm{cls}}(\ell) = \langle r, d\rangle + \int_0^1 (1 - t) d^T H(\ell + td) d \, dt. \tag{A.38}$$

The Hessian bound implies

$$0 \leq \mathcal{L}_{\mathrm{cls}}(\ell + d) - \mathcal{L}_{\mathrm{cls}}(\ell) - \langle r, d\rangle \leq \frac{1}{8Z} \sum_{ij} u_{ij}q_{ij}d_{ij}^2. \tag{A.39}$$

A sufficient condition for a strict reduction of the classification loss is therefore

$$-\langle r, d\rangle > \frac{1}{8Z} \sum_{ij} u_{ij}q_{ij}d_{ij}^2. \tag{A.40}$$

The linear term rewards alignment with the negative classification gradient, while the quadratic term controls correction size. For $d = \sum_v m_v g_v \delta_v$, residuals and gates determine both terms.

For a single candidate direction d and scale $t \geq 0$, define $A = -\langle r, d\rangle$ and $B = Z^{-1} \sum_{ij} u_{ij}q_{ij}d_{ij}^2$. Then

$$\mathcal{L}_{\mathrm{cls}}(\ell + td) - \mathcal{L}_{\mathrm{cls}}(\ell) \leq -tA + \tfrac{1}{8}t^2 B. \tag{A.41}$$

When $A > 0$ and $B > 0$, every $0 < t < 8A/B$ decreases this upper bound below zero. Its minimizer is $t_* = 4A/B$, with value $-2A^2/B$. The minimizing scale quantifies shrinkage along an informative residual direction at the fixed training predictor.

The head parameterization makes a descent direction explicit. With features and gates fixed, define the augmented feature vector

$$\widetilde{h}_i = \sum_v m_{iv} g_v \begin{bmatrix} h_{iv}^{\mathrm{p}} \\ 1 \end{bmatrix}, \qquad D_i = [W_{\mathrm{r}} \; b_{\mathrm{r}}]\widetilde{h}_i. \tag{A.42}$$

Stack the rows of $[W_{\mathrm{r}} \; b_{\mathrm{r}}]$ into ϑ and define the corresponding linear map Ψ by $\mathrm{vec}(D) = \Psi\vartheta$. At the zero head, let $r = \nabla\mathcal{L}_{\mathrm{cls}}(F_*)$, $Q = \mathrm{diag}(u_{ij}q_{ij}/Z)$, and $b = \Psi^T r$. A main-classification gradient step on the head produces

$$\vartheta(t) = -tb, \qquad D(t) = -t\Psi\Psi^T r, \qquad -\langle r, D(t)\rangle = t\|\Psi^T r\|^2. \tag{A.43}$$

Substitution into the curvature bound gives

$$\mathcal{L}_{\mathrm{cls}}(F_* + D(t)) - \mathcal{L}_{\mathrm{cls}}(F_*) \leq -t\|b\|^2 + \tfrac{1}{8}t^2 B_\Psi, \qquad B_\Psi = \|Q^{1/2}\Psi b\|^2. \tag{A.44}$$

For $b \neq 0$, suppose $B_\Psi = 0$. Positive class weights imply $(\Psi b)_{ij} = 0$ on every observed coordinate. Since $r_{ij} = 0$ elsewhere, this would give $\|b\|^2 = r^T\Psi b = 0$, a contradiction. Hence $B_\Psi > 0$, and minimizing the quadratic yields

$$t_* = \frac{4\|b\|^2}{B_\Psi}, \qquad \mathcal{L}_{\mathrm{cls}}(F_*) - \mathcal{L}_{\mathrm{cls}}(F_* + D(t_*)) \geq \frac{2\|b\|^4}{B_\Psi} > 0. \tag{A.45}$$

The nonzero projected gradient is a sufficient condition for a realizable improving head update. The zero-initialized residual family therefore contains both the archived predictor and a strictly better

training prediction whenever its fixed features expose such a direction. This statement concerns the main classification term with features and gates fixed; the full residual objective also updates those features through its auxiliary and representation terms.

With residual encoders and heads fixed, the classification loss and gate penalty define

$$J(g) = \mathcal{L}_{\text{cls}}\left(\ell_* + \sum_v m_v g_v \delta_v\right) + c_g \sum_v g_v, \qquad c_g = \lambda_{\text{gate}}/V. \tag{A.46}$$

The other terms in the residual objective are independent of g. The partial derivative in gate space is

$$\frac{\partial J}{\partial g_v} = B_v + c_g, \qquad B_v = \sum_i m_{iv}\langle r_i, \delta_{iv}\rangle. \tag{A.47}$$

Applying $g_v = \sigma(a_v)$ yields

$$\frac{\partial J}{\partial a_v} = g_v(1-g_v)(B_v + c_g). \tag{A.48}$$

For finite a_v, the multiplier is strictly positive. A gradient-descent update opens the gate exactly when $B_v + c_g < 0$. At a finite stationary point,

$$-\sum_i m_{iv}\langle r_i, \delta_{iv}\rangle = c_g. \tag{A.49}$$

The equality balances training-gradient alignment with the cost of exposing that view.

The curvature in gate space is

$$\frac{\partial^2 J}{\partial g_v \partial g_w} = \frac{1}{Z}\sum_{ij} u_{ij} q_{ij} p_{ij}(1-p_{ij}) m_{iv} m_{iw} \delta_{ivj}\delta_{iwj}. \tag{A.50}$$

For any $t \in \mathbb{R}^V$,

$$t^T \nabla_g^2 J\, t = \frac{1}{Z}\sum_{ij} u_{ij} q_{ij} p_{ij}(1-p_{ij})\left(\sum_v t_v m_{iv}\delta_{ivj}\right)^2 \geq 0. \tag{A.51}$$

Thus J is convex in g when the residual functions are fixed. A minimizer over the closed cube $[0,1]^V$ satisfies the coordinate optimality conditions

$$\begin{cases} B_v + c_g \geq 0, & g_v = 0, \\ B_v + c_g = 0, & 0 < g_v < 1, \\ B_v + c_g \leq 0, & g_v = 1. \end{cases} \tag{A.52}$$

Finite sigmoid logits parameterize the open cube; its boundary is approached only as $a_v \to \pm\infty$. In logit coordinates the Hessian is

$$\nabla_a^2 J = D\nabla_g^2 J D + \operatorname{diag}\big(\sigma''(a_v)(B_v + c_g)\big), \qquad D = \operatorname{diag}(\sigma'(a_v)). \tag{A.53}$$

The second term can have either sign. The full optimization also changes the residual functions, so the convexity statement applies specifically to the fixed-residual gate subproblem.

At $W = b = 0$ and $a_v = 0$, all residuals vanish. Equation A.48 becomes

$$\left.\frac{\partial J}{\partial a_v}\right|_{t=0} = \frac{c_g}{4}, \qquad a_v^{(1)} = -\frac{\eta_g c_g}{4}, \qquad g_v^{(1)} = \frac{1}{2} - \frac{\eta_g c_g}{16} + O((\eta_g c_g)^3). \tag{A.54}$$

The expansion describes an ordinary gradient-descent step on the gate objective.

For a shared residual head $\delta_{iv} = Wh_{iv} + b$, the main classification term has gradients

$$\nabla_W \mathcal{L}_{\text{cls}} = \sum_{i,v} m_{iv} g_v r_i h_{iv}^T, \qquad \nabla_b \mathcal{L}_{\text{cls}} = \sum_{i,v} m_{iv} g_v r_i. \tag{A.55}$$

At initialization these expressions can be nonzero even though every residual output is zero. The gradient reaching encoder parameters ψ_v is

$$\nabla_{\psi_v} \mathcal{L}_{\text{cls}} = \sum_i m_{iv} g_v J_{h_{iv},\psi_v}^T W^T r_i = 0 \quad \text{when } W = 0. \tag{A.56}$$

Auxiliary classification terms likewise update the output head and initially contribute zero encoder gradient through that head. Reconstruction and representation penalties may contribute nonzero encoder gradients independently. After a head update $W^{(1)} = -\eta_W G_W$, the classification gradient through the encoder is generally of order η_W under bounded local Jacobians.

A.5 Incumbent Selection and Finite-Candidate Population Bounds

Let $\theta_0 = \theta^*$ and let $\theta_1, \dots, \theta_K$ denote residual checkpoints. Define the recursive selected index by

$$i_0 = 0, \qquad i_t = \begin{cases} t, & \widehat{S}(\theta_t) > \widehat{S}(\theta_{i_{t-1}}), \\ i_{t-1}, & \text{otherwise.} \end{cases} \tag{A.57}$$

Proposition A.3. *For every realized score sequence, $\widehat{S}(\theta_{i_K}) = \max_{0\le k\le K} \widehat{S}(\theta_k) \ge \widehat{S}(\theta_0)$. If no checkpoint strictly exceeds the incumbent, $i_K = 0$.*

Proof. The assertion is immediate at $t = 0$. Suppose $\widehat{S}(\theta_{i_{t-1}}) = \max_{k<t} \widehat{S}(\theta_k)$. The recursion selects the greater of this score and $\widehat{S}(\theta_t)$, preserving the maximum. Induction proves the first claim. If no new score exceeds the incumbent, the recursion retains $i_t = 0$ throughout. □

Theorem A.4. *On an event $\mathcal{E}_\varepsilon$ satisfying $\sup_{0\le k\le K} |S(\theta_k) - \widehat{S}(\theta_k)| \le \varepsilon$, the selected predictor $\widehat{\theta}$ obeys*

$$S(\widehat{\theta}) - S(\theta_0) \ge \widehat{\Delta} - 2\varepsilon, \qquad S(\widehat{\theta}) \ge \max_{0\le k\le K} S(\theta_k) - 2\varepsilon, \tag{A.58}$$

where $\widehat{\Delta} = \widehat{S}(\widehat{\theta}) - \widehat{S}(\theta_0) \ge 0$.

Proof. On $\mathcal{E}_\varepsilon$,

$$S(\widehat{\theta}) - S(\theta_0) \ge \widehat{S}(\widehat{\theta}) - \varepsilon - \widehat{S}(\theta_0) - \varepsilon = \widehat{\Delta} - 2\varepsilon. \tag{A.59}$$

For any candidate θ_k, empirical maximization gives

$$S(\widehat{\theta}) \ge \widehat{S}(\widehat{\theta}) - \varepsilon \ge \widehat{S}(\theta_k) - \varepsilon \ge S(\theta_k) - 2\varepsilon. \tag{A.60}$$

Taking the maximum over k proves the second inequality. □

Let $k^\star$ attain the best population score and let $\Gamma_K = S(F_{k^\star}) - S(F_*)$. Combining the two inequalities gives

$$S(F_{\widehat{k}}) - S(F_*) \ge \max\{\widehat{\Delta} - 2\varepsilon, \Gamma_K - 2\varepsilon\}, \tag{A.61}$$

which proves Theorem 3.2. For negative log loss, define E_k with respect to p_j^+ for every candidate, including the incumbent. Equation A.6 then gives

$$\Gamma_K = \mathcal{I}_{\mathrm{comp}} + E_* - \min_{0\le k\le K} E_k, \qquad R(F_*) - R(F_{\widehat{k}}) \ge \mathcal{I}_{\mathrm{comp}} + E_* - \min_k E_k - 2\varepsilon_R. \tag{A.62}$$

Here ε_R bounds risk-estimation errors in the units of log loss, and selection uses the direct log-loss score of the functions entering the information identity. Define $a_k = c^{-1} \sum_j \mathbb{E}(D_{k,j} - d_j^*)^2$ and $\mathcal{A}_K = \min_{0\le k\le K} a_k$. For k_a attaining this finite minimum, the logit approximation inequality implies

$$\min_k E_k \le E_{k_a} \le \frac{a_{k_a}}{8} = \frac{\mathcal{A}_K}{8}, \qquad R(F_*) - R(F_{\widehat{k}}) \ge \mathcal{I}_{\mathrm{comp}} + E_* - \frac{\mathcal{A}_K}{8} - 2\varepsilon_R. \tag{A.63}$$

The approximation-minimizing and divergence-minimizing candidates may differ. Evaluating the divergence at k_a proves the required bound without identifying their minimizers. An evaluation rule with late fusion defines its own score S for the general selection theorem. For an ideal residual family $\mathcal{D}$ containing all candidates, set $\xi_K = \min_k E_k - E_{\mathcal{D}} \ge 0$. The gain decomposes as

$$\Gamma_K = \Gamma_{\mathcal{D}} - \xi_K, \qquad R(F_*) - R(F_{\widehat{k}}) \ge \Gamma_{\mathcal{D}} - \xi_K - 2\varepsilon_R. \tag{A.64}$$

The family gain is reduced by candidate approximation and selection uncertainty. A sufficient condition for a positive selected gain is $\Gamma_{\mathcal{D}} > \xi_K + 2\varepsilon_R$. If the incumbent is selected, the realized gain is exactly zero. The theorem distinguishes improvement available in the residual family, improvement reached by optimization, and improvement supported by selection data.

To establish such an event, let $\mathcal{G}$ contain all data and randomness used to construct the incumbent and residual candidates. Conditional on $\mathcal{G}$, assume the selection observations $Z_1, \dots, Z_n$ are independent and identically distributed and independent of that construction. Let

$$\widehat{S}(\theta_k) = \frac{1}{n} \sum_{i=1}^{n} q(\theta_k; Z_i), \qquad S(\theta_k) = \mathbb{E}[q(\theta_k; Z) \mid \mathcal{G}], \qquad 0 \le q \le 1. \tag{A.65}$$

For fixed k, Hoeffding's lemma and independence give (Hoeffding, 1963)

$$\mathbb{E}\left[\exp\left(t\sum_i (q(\theta_k; Z_i) - S(\theta_k))\right)\middle|\mathcal{G}\right] \le \exp(nt^2/8). \tag{A.66}$$

Markov's inequality therefore implies, for $t > 0$,

$$\Pr(\widehat{S}(\theta_k) - S(\theta_k) \ge \varepsilon \mid \mathcal{G}) \le \exp(-tn\varepsilon + nt^2/8). \tag{A.67}$$

Minimizing at $t = 4\varepsilon$ yields the one-sided bound $e^{-2n\varepsilon^2}$. Applying it to the negative deviation and taking a union over candidates gives

$$\Pr(\mathcal{E}_\varepsilon^c \mid \mathcal{G}) \le 2(K+1)e^{-2n\varepsilon^2}. \tag{A.68}$$

Set $\varepsilon = \sqrt{\log(2(K+1)/\delta)/(2n)}$. Integration over $\mathcal{G}$ establishes the same probability bound unconditionally. Theorem A.4 then yields

$$S(\widehat{\theta}) - S(\theta_0) \ge \widehat{\Delta} - 2\sqrt{\frac{\log(2(K+1)/\delta)}{2n}}. \tag{A.69}$$

An independent selection sample yields the finite-candidate bound above. Adaptive reuse of selection observations requires a uniform concentration event over the resulting candidate family (Cawley & Talbot, 2010; Dwork et al., 2015). Proposition A.3 holds for every realized sequence of scores.

A direct paired analysis uses $d_k(Z) = q(\theta_k; Z) - q(\theta_0; Z) \in [-1, 1]$. Let $\Delta_k = \mathbb{E}[d_k(Z) \mid \mathcal{G}]$ and $\widehat{\Delta}_k = n^{-1}\sum_i d_k(Z_i)$. A one-sided Hoeffding bound and union over $k = 1, \dots, K$ give

$$\Pr\Big(\exists k : \widehat{\Delta}_k - \Delta_k > \eta \mid \mathcal{G}\Big) \le K\exp(-n\eta^2/2). \tag{A.70}$$

For $K \ge 1$, with probability at least $1 - \delta$, every augmented candidate satisfies

$$\Delta_k \ge \widehat{\Delta}_k - \sqrt{\frac{2\log(K/\delta)}{n}}. \tag{A.71}$$

The incumbent has paired difference zero exactly. The paired bound compares each correction directly with the incumbent under the same independence conditions.

A.6 Ranking Criteria and Class-Conditional Concentration

AUC compares positive and negative examples. For class j, condition on $n_{j,+} > 0$ positive and $n_{j,-} > 0$ negative selection examples, drawn independently from their class-conditional distributions. For a fixed score function f, define

$$h_f(x^+, x^-) = \mathbf{1}\{f(x^+) > f(x^-)\} + \tfrac{1}{2}\mathbf{1}\{f(x^+) = f(x^-)\}, \tag{A.72}$$

$$\widehat{A}_j(f) = \frac{1}{n_{j,+}n_{j,-}}\sum_{a=1}^{n_{j,+}}\sum_{b=1}^{n_{j,-}} h_f(X_a^+, X_b^-), \qquad A_j(f) = \mathbb{E}[h_f(X^+, X^-)]. \tag{A.73}$$

Replacing one positive observation changes $\widehat{A}_j$ by at most $1/n_{j,+}$; replacing one negative changes it by at most $1/n_{j,-}$. The sum of squared bounded-difference constants is

$$H_j = n_{j,+}(1/n_{j,+})^2 + n_{j,-}(1/n_{j,-})^2 = 1/n_{j,+} + 1/n_{j,-}. \tag{A.74}$$

McDiarmid's inequality gives the conditional concentration bound (McDiarmid, 1989)

$$\Pr(|\widehat{A}_j(f) - A_j(f)| > \varepsilon \mid n_{j,+}, n_{j,-}, \mathcal{G}) \le 2\exp(-2\varepsilon^2/H_j). \tag{A.75}$$

For a fixed set of c evaluated classes and $K+1$ independently constructed candidates, allocate failure probability $\delta/[c(K+1)]$ to each pair. For random class counts, integrate each class-conditional bound over its own counts before taking the union. On the event that every class has both labels represented, simultaneous errors satisfy

$$\varepsilon_j = \sqrt{\frac{H_j}{2}\log\frac{2c(K+1)}{\delta}}, \qquad \left|\frac{1}{c}\sum_j \widehat{A}_j(f) - \frac{1}{c}\sum_j A_j(f)\right| \le \frac{1}{c}\sum_j \varepsilon_j. \tag{A.76}$$

The probability that any defined bound fails while all counts are positive is at most δ. For iid selection instances with class prevalence $\pi_j \in (0,1)$, let E_+ denote the event that every class has both signs represented. Then

$$\Pr(E_+^c) \le \sum_{j=1}^{c} \left[\pi_j^n + (1-\pi_j)^n\right] =: \varepsilon_{\text{empty}}. \tag{A.77}$$

Thus all class AUCs are defined and Equation A.76 holds with probability at least $1-\delta-\varepsilon_{\text{empty}}$. Classes with few positive or negative examples receive wider bounds. The union bound allows correlated labels and uses independence across the sampled selection instances.

For instance-averaged AP, an instance with positive-label set Y_i and ranking π_i has

$$\mathrm{AP}_i(f) = \frac{1}{|Y_i|} \sum_{r=1}^{c} \mathbf{1}\{\pi_i(r) \in Y_i\} \frac{\sum_{s=1}^{r} \mathbf{1}\{\pi_i(s) \in Y_i\}}{r} \in [0,1]. \tag{A.78}$$

A fixed convention handles instances with $|Y_i| = 0$. Averaging these bounded instance scores permits Equation A.69. Class-wise AP that ranks examples uses a different estimator and requires its own concentration analysis. Allocate $\delta_{\mathrm{AP}} + \delta_{\mathrm{AUC}} = \delta$ to the two criteria. For $S = (S_{\mathrm{AP}} + S_{\mathrm{AUC}})/2$, the probability that either defined bound fails is at most δ. On their joint success event,

$$|\widehat{S} - S| \le \tfrac{1}{2}(\varepsilon_{\mathrm{AP}} + \varepsilon_{\mathrm{AUC}}). \tag{A.79}$$

For iid selection, the joint success event with all estimators defined has probability at least $1-\delta-\varepsilon_{\text{empty}}$. Substitution into Theorem A.4 gives the corresponding population comparison.

A.7 Observation Masks and the Effective Sample Size

Let M_{iv} indicate an observed view and U_{ij} an observed label. For a view–label pair define

$$A_{ivj} = M_{iv} U_{ij}, \qquad N_{vj} = \sum_{i=1}^{N} A_{ivj}, \qquad q_{vj} = \Pr(A_{ivj} = 1). \tag{A.80}$$

Assume observations and masks are identically distributed and independent across instances, and the joint observation event is independent of the underlying features and labels. Conditional on $N_{vj} = n > 0$, the retained examples then have the target population distribution. If view and label observation indicators are also mutually independent within an instance, $q_{vj} = \rho_v \pi_j$. Otherwise q_{vj} is their joint probability.

Let $\mathcal{H}_{vj}$ be a fixed class of loss functions taking values in $[0,1]$, and let

$$R(h) = \mathbb{E}[h(Z)], \qquad \widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(Z_i), \qquad \mathfrak{R}_n(\mathcal{H}_{vj}) = \mathbb{E}_{Z,\epsilon} \sup_{h \in \mathcal{H}_{vj}} \frac{1}{n} \sum_i \epsilon_i h(Z_i). \tag{A.81}$$

The class is fixed independently of these retained examples. A data-dependent anchor permits an independent fitting split or a fixed class containing all admissible anchor–residual losses.

For an independent ghost sample $Z_1', \ldots, Z_n'$, Jensen's inequality and symmetrization give (Bartlett & Mendelson, 2002)

$$\begin{aligned} \mathbb{E} \sup_h (R(h) - \widehat{R}_n(h)) &\le \mathbb{E} \sup_h \frac{1}{n} \sum_i (h(Z_i') - h(Z_i)) \\ &= \mathbb{E} \sup_h \frac{1}{n} \sum_i \epsilon_i (h(Z_i') - h(Z_i)) \\ &\le 2\mathfrak{R}_n(\mathcal{H}_{vj}). \end{aligned} \tag{A.82}$$

The same bound holds for $\mathbb{E}\sup_h(\widehat{R}_n(h) - R(h))$. Replacing one example changes either supremum by at most $1/n$. Bounded differences and a union over the two signs yield, with probability at least $1-\delta$,

$$\sup_{h \in \mathcal{H}_{vj}} |R(h) - \widehat{R}_n(h)| \le 2\mathfrak{R}_n(\mathcal{H}_{vj}) + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{A.83}$$

Under independent Bernoulli observation across instances, $N_{vj} \sim \mathrm{Binomial}(N, q_{vj})$. The observation count satisfies the following lower-tail bound

$$\Pr(N_{vj} < Nq_{vj}/2) \le \exp(-Nq_{vj}/8). \tag{A.84}$$

If $\mathfrak{R}_n(\mathcal{H}_{vj}) \le C\sqrt{\kappa/n}$ and $Nq_{vj} \ge 2$, combining the events gives

$$\sup_h |R(h) - \widehat{R}_{N_{vj}}(h)| \le 2C\sqrt{\frac{2\kappa}{Nq_{vj}}} + \sqrt{\frac{\log(2/\delta)}{Nq_{vj}}} \tag{A.85}$$

with probability at least $1 - \delta - \exp(-Nq_{vj}/8)$. Simultaneous control over all Vc pairs follows by replacing δ by $\delta/(Vc)$ and adding the corresponding observation-tail probabilities. The effective supervised sample size is therefore governed by the joint availability of the view and label.

The loss-range assumption can be made explicit for logistic risk. If $|\ell| \le B$ and $q_{ij} \le w_{\max}$, the per-coordinate weighted logistic loss satisfies

$$0 \le q_{ij}[\log(1 + e^{\ell}) - y\ell] \le w_{\max} \log(1 + e^{B}) = L_B. \tag{A.86}$$

Applying Equation A.83 to the loss divided by L_B and multiplying back gives a bounded-logit risk guarantee. Unrestricted logits require a tail assumption; explicit clipping supplies bounded losses.

B Related Work

B.1 Learning from Incomplete Views and Labels

Multi-view multi-label classification connects heterogeneous observations with a shared label space. Missing views restrict the information available for individual instances, while missing labels reduce supervision for learning which information supports each category. Methods differ in how they use the observed entries, infer unavailable information, and transfer supervision across views. The broader multi-view and multimodal literature provides a common perspective on alignment, fusion, and representation learning (Li et al., 2019; Baltrušaitis et al., 2019).

Structured models use relationships among instances, views, and labels to constrain learning from incomplete observations. iMVWL learns a common representation for incomplete views and weak labels (Tan et al., 2018). NAIM3L extends structured learning to non-aligned views and missing labels (Li & Chen, 2022), while LVSL learns view-specific labels for non-aligned multi-view classification (Zhao et al., 2023). These formulations connect the observed portions of the data through shared structure or view-dependent supervision. Their treatment of correspondence and annotation availability also determines which information can be transferred between instances.

Deep methods learn nonlinear interactions among the available sources. DICNet uses instance-level contrastive learning to encourage cross-view agreement (Liu et al., 2023b). LMVCAT combines label-guided masking with view- and category-aware transformers (Liu et al., 2023c). AIMNet uses attention-induced embedding imputation to supply missing representations (Liu et al., 2024a). Agreement learning, semantic interaction, and imputation address different consequences of incomplete observation. They respectively strengthen common representations, organize information around the prediction task, and reconstruct unavailable inputs in the latent space.

B.2 Shared Semantics, View-Specific Information, and Reliable Fusion

Shared–specific representation learning assigns common and individual-source information to distinct pathways. Domain Separation Networks introduce shared and private representations for domain adaptation (Bousmalis et al., 2016), and MISA applies modality-invariant and modality-specific representations to multimodal sentiment analysis (Hazarika et al., 2020). Within incomplete multi-view multi-label learning, MTD separates shared and view-proprietary channels through masked representation learning, cross-channel contrastive objectives, and label-guided regularization (Liu et al., 2023a). DCSI develops a dual-channel formulation for consistent and specific information (Wen et al., 2026b). These methods make the allocation of information across representation channels an explicit part of model design.

Semantic compression provides a related route to task-relevant representations. SIP uses an information-bottleneck formulation to concentrate shared predictive information and models label prototypes in the latent space (Liu et al., 2024b). Reliability-aware approaches further control how each source contributes to prediction. RANK combines label-guided representation learning with view-quality estimation (Liu et al., 2025), while URDF uses uncertainty estimates for dynamic fusion (Wen et al., 2026a). Semantic relevance and view reliability are distinct considerations. A view may be reliable yet offer little information beyond the shared prediction, or contain a category-specific cue that is useful only for selected labels.

Probabilistic fusion represents uncertainty when combining observed modalities. Multimodal variational autoencoders use products of experts to aggregate available sources (Wu & Goodman, 2018). Mixture-of-experts inference and generalized multimodal variational objectives provide further constructions for learning across modality subsets (Shi et al., 2019; Sutter et al., 2021). CARE uses Gaussian expert aggregation and label semantics to establish its consensus predictor. Its residual stage then evaluates view-specific information relative to this fixed prediction function. Representation separation describes where information is encoded; the anchored formulation additionally specifies the prediction that complementary learning is intended to refine.

B.3 Staged Learning and Anchored Residual Adaptation

Staged training already appears in incomplete multi-view multi-label learning. TSIEN first trains view-specific classifiers to concentrate task-relevant information. It then processes their hidden states through alignment, autoencoder-based information extraction, and weighted fusion (Tan et al., 2024). This design separates task-oriented feature extraction from subsequent information integration. CARE assigns a different role to the stage boundary. The first stage produces a complete consensus predictor whose state is preserved during residual learning and whose predictions remain available during final model selection.

The interaction between learning pathways has also motivated methods for balancing joint optimization. Studies of multimodal classification identify uneven learning across modalities (Wang et al., 2020; Peng et al., 2022). GradNorm balances gradient magnitudes, while gradient surgery and conflict-averse descent adjust update directions across objectives (Chen et al., 2018; Yu et al., 2020; Liu et al., 2021). These methods regulate how jointly trained components change. In CARE, fixing the consensus state gives the residual objective a stable prediction reference, so the learned correction can be assessed separately from changes to the shared predictor.

Additive learning and frozen-model adaptation provide complementary foundations. Gradient boosting builds predictors through successive corrections (Friedman, 2001). Adapters and LoRA add trainable capacity around pretrained networks (Houlsby et al., 2019; Hu et al., 2022); ControlNet and Flamingo demonstrate zero-initialized or gated components for extending pretrained models (Zhang et al., 2023; Alayrac et al., 2022). CARE organizes residual adaptation around incomplete views and labels, combining a preserved consensus with label-specific corrections and incumbent-inclusive selection. Its analysis connects the conditional information available beyond consensus to residual approximation and selection uncertainty. This connection identifies when a complementary correction can yield a retained predictive gain. Table 3 summarizes the representative formulations discussed here.

Table 3: Representative methods grouped by their main modeling emphasis. The categories describe related work; the experimental comparison uses the nine baselines specified in Section 4.

Method	Publication	Main modeling choice
Incomplete observation and cross-view structure		
iMVWL	IJCAI 2018	Common representation with incomplete views and weak labels
NAIM3L	TPAMI 2022	Structured learning with non-aligned views and missing labels
LVSL	TMM 2023	View-specific label learning for non-aligned views
DICNet	AAAI 2023	Instance-level contrastive consensus learning
LMVCAT	AAAI 2023	Masked view interaction and category-aware label modeling
AIMNet	AAAI 2024	Attention-induced embedding imputation
Shared semantics, specific information, and reliability		
MTD	NeurIPS 2023	Masked learning with shared and view-proprietary channels
SIP	ICML 2024	Semantic invariance and label prototype modeling
DCSI	TPAMI 2026	Disentanglement of consistent and specific information
RANK	TPAMI 2025	Label-guided representation and view-quality estimation
URDF	TPAMI 2026	Uncertainty-driven reliable dynamic fusion
Staged information learning and predictive refinement		
TSIEN	AAAI 2024	View-specific task learning followed by information integration
CARE	This work	Preserved consensus, residual corrections, and incumbent selection

C Datasets and Experimental Protocol

Corel5k contains natural images with keyword annotations (Duygulu et al., 2002). The evaluated feature version retains 4,999 instances and 260 labels. Its relatively small sample count and broad vocabulary make supervision sparse for individual label coordinates. ESPGame obtains semantic tags through a human image-labeling game (von Ahn & Dabbish, 2004). Its 20,770 instances and 268 labels provide a larger annotation setting with varied object and scene content. IAPRTC12 originates from a benchmark for visual information retrieval (Grubinger et al., 2006). The feature version contains 19,627 instances and 291 labels, giving the largest label vocabulary in the evaluation.

Table 4: Benchmark instances, labels, and view dimensions. Counts refer to the feature versions used for multi-view evaluation. Dimensions are listed in the order of the descriptors described below.

Dataset	Instances	Labels	Views	Feature dimensions
Corel5k	4,999	260	6	100 / 1,000 / 512 / 4,096 / 4,096 / 4,096
ESPGame	20,770	268	6	100 / 1,000 / 512 / 4,096 / 4,096 / 4,096
IAPRTC12	19,627	291	6	100 / 1,000 / 512 / 4,096 / 4,096 / 4,096
Mirflickr	25,000	38	6	100 / 1,000 / 512 / 4,096 / 4,096 / 4,096
Pascal07	9,963	20	6	100 / 1,000 / 512 / 4,096 / 4,096 / 4,096
OBJECT	6,047	31	5	64 / 225 / 144 / 73 / 128

Mirflickr contains 25,000 Flickr images and 38 annotation labels in the evaluated representation (Huiskes & Lew, 2008). Pascal07 contains 9,963 images and 20 object categories (Everingham et al., 2010). These two datasets provide smaller label spaces with different image collections and annotation protocols. OBJECT contains 6,047 instances with 31 labels and five handcrafted views (Hao et al., 2024). Its feature dimension totals 634; each six-view representation totals 13,900.

The six-view representation contains DenseHue, DenseSift, GIST, HSV, RGB, and LAB descriptors with dimensions 100, 1,000, 512, 4,096, 4,096, and 4,096. They encode color, local appearance, global layout, and color-channel structure. OBJECT uses CH, CM, CORR, EDH, and WT with dimensions 64, 225, 144, 73, and 128. Table 4 describes these released feature configurations; its counts specify the evaluated versions rather than the size of every original image collection.

Masks act at different levels of the pipeline. A view mask excludes the corresponding encoder input, Gaussian expert, auxiliary loss, and residual contribution. A label mask excludes the supervised loss coordinate while retaining the reference label for evaluation. Each repeated realization is shared across comparison methods. The same test instances and reference vectors are used for the consensus and augmented predictors. The paired stage difference consequently compares two prediction functions on a common evaluation set.

Training and selection. All experiments use 50 epochs for consensus learning and 25 epochs for residual refinement. Consensus training archives the selected state, including label prototypes and normalization buffers. Residual training initializes all output weights and biases to zero and all gate logits to zero. The archived predictor and residual checkpoints are evaluated by the same selection score. Hyperparameter choices use validation data. The finite-candidate concentration result applies when this final selection sample is independent of all candidate construction; adaptive reuse is covered by the uniform-event formulation in Appendix A.

Screening subset. CheXpert-MV contains 31,413 pairs and 14 observation labels. Frontal and lateral DenseNet-121 embeddings each contain 1,024 features (Irvin et al., 2019; Huang et al., 2017).

Controlled motivation experiment. Figure 1(a)–(b) uses a linear shared–private model on a synthetic task extension. Two six-dimensional views mix two shared factors, two private factors per view, and noise. Four semantic groups define two shared labels, and each view supplies three private binary attributes. We generate 768 training and 2,048 held-out examples per seed. Shared and private labels are observed with probabilities 0.05 and 0.30. Each view is removed with probability 0.30, restoring one whenever both are missing. The common predictor learns the shared labels with cross-view agreement for 50 epochs. Both refinement arms then optimize all eight labels for 25 epochs using gated residuals and a shared–private covariance penalty. The primary refinement objective omits the pretraining agreement term. Paired runs share their initialization, examples, masks, batches, optimizer, and budget; only the trainability of the shared branch differs.

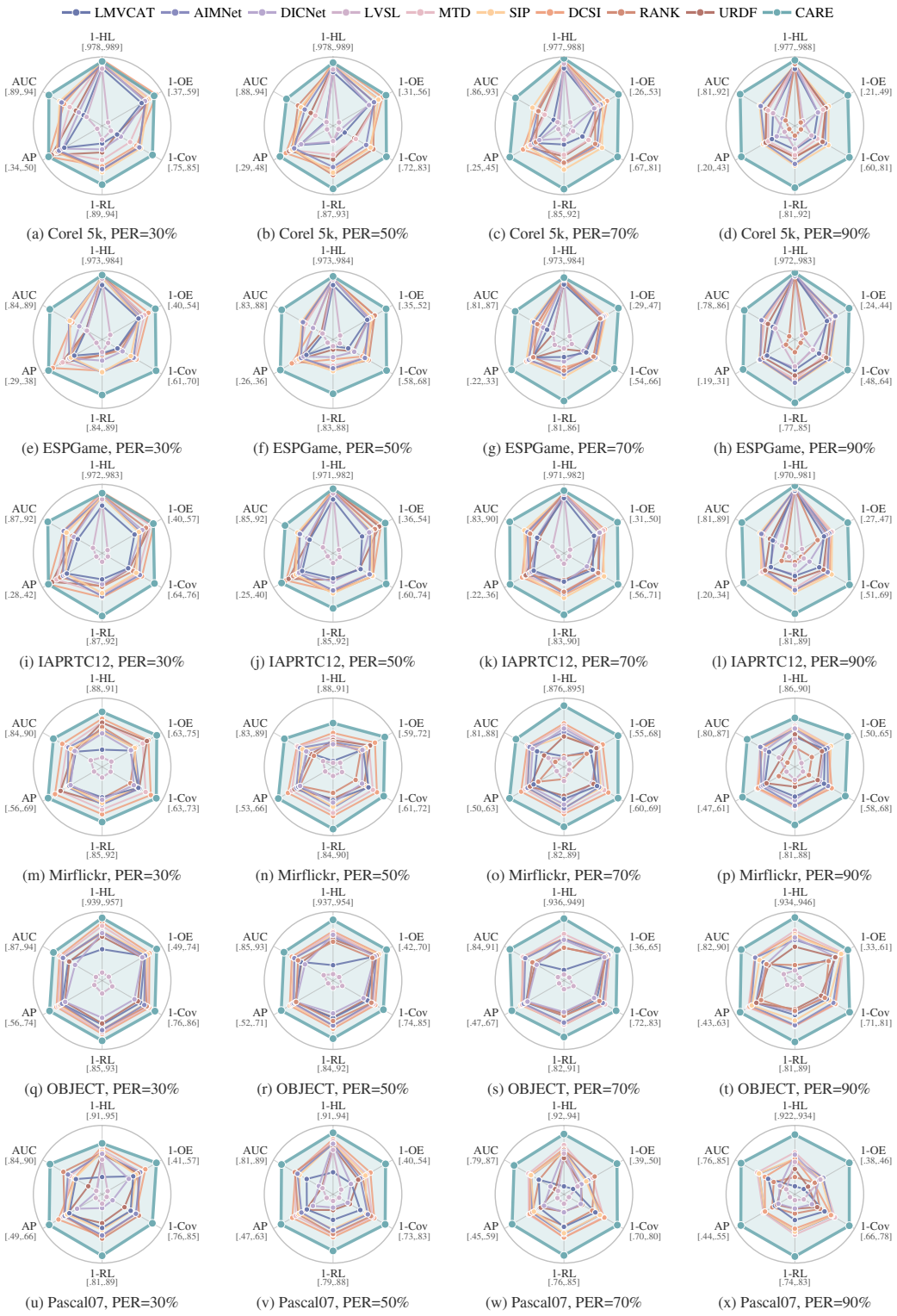
A linear coordinate map is fitted from the archived shared features to the generating shared factors on 1,024 held-out examples. The same map displays both refined representations on the remaining examples. The figure pools five seeds and displays three of the four groups with common axes. Across all groups, a separate held-out linear probe recovers the shared factors with mean R^2 of 0.807 after joint refinement and 0.992 with the shared branch fixed. Retaining cross-view agreement during joint refinement also yields $R^2 = 0.992$, identifying continued semantic constraints as another means of preserving shared structure. The fixed branch supplies that preservation directly while complementary tasks are learned. Figure 1(c) reports the separate framework-adaptation comparison on Corel5k and Pascal07.

D Detailed Results of Two-Stage Refinement

Table 5 compares the consensus and refined predictors at PER $=$ LMR $=30\%$ over three folds. Each absolute gain is computed within a fold before aggregation, so its standard deviation reflects variation in the paired improvement. Relative gains compare the two stage means. The stage-wise evaluation covers the six benchmark metrics and Macro-F_1 with calibrated thresholds. Positive mean gains across all seven metrics connect residual refinement to ranking, label coverage, and thresholded prediction.

Table 5: Stage-wise performance at PER $=30\%$ and LMR $=30\%$ over three folds. Scores and absolute gains are mean_{std}. Absolute gains use paired fold differences; relative gains are $100(\bar{x}_2/\bar{x}_1-1)$. All metrics favor larger values. Macro-F_1 uses calibrated thresholds.

Data	Metric	Stage 1	Stage 2	Absolute gain	Relative gain (%)
COR	1-HL	$0.98784_{0.00011}$	$\mathbf{0.98832}_{0.00012}$	$0.00048_{0.00011}$	**+0.05%**
	1-OE	$0.52500_{0.00455}$	$\mathbf{0.56200}_{0.00942}$	$0.03700_{0.00490}$	**+7.05%**
	1-Cov	$0.82413_{0.00246}$	$\mathbf{0.83461}_{0.00277}$	$0.01049_{0.00058}$	**+1.27%**
	1-RL	$0.92724_{0.00083}$	$\mathbf{0.93275}_{0.00095}$	$0.00550_{0.00044}$	**+0.59%**
	AP	$0.45407_{0.00416}$	$\mathbf{0.48375}_{0.00416}$	$0.02968_{0.00388}$	**+6.54%**
	AUC	$0.92941_{0.00076}$	$\mathbf{0.93469}_{0.00090}$	$0.00528_{0.00049}$	**+0.57%**
	Macro-F_1	$0.13915_{0.00572}$	$\mathbf{0.15459}_{0.00312}$	$0.01544_{0.00879}$	**+11.10%**
ESP	1-HL	$0.98288_{0.00010}$	$\mathbf{0.98328}_{0.00013}$	$0.00040_{0.00004}$	**+0.04%**
	1-OE	$0.49101_{0.00728}$	$\mathbf{0.52512}_{0.00748}$	$0.03410_{0.00030}$	**+6.95%**
	1-Cov	$0.65534_{0.00034}$	$\mathbf{0.69147}_{0.00142}$	$0.03613_{0.00122}$	**+5.51%**
	1-RL	$0.86487_{0.00077}$	$\mathbf{0.88026}_{0.00080}$	$0.01540_{0.00009}$	**+1.78%**
	AP	$0.33525_{0.00419}$	$\mathbf{0.37096}_{0.00305}$	$0.03570_{0.00132}$	**+10.65%**
	AUC	$0.86916_{0.00056}$	$\mathbf{0.88388}_{0.00111}$	$0.01471_{0.00066}$	**+1.69%**
	Macro-F_1	$0.10334_{0.00227}$	$\mathbf{0.14560}_{0.00240}$	$0.04226_{0.00462}$	**+40.90%**
IAP	1-HL	$0.98099_{0.00017}$	$\mathbf{0.98159}_{0.00009}$	$0.00060_{0.00009}$	**+0.06%**
	1-OE	$0.49236_{0.00675}$	$\mathbf{0.54619}_{0.00098}$	$0.05383_{0.00625}$	**+10.93%**
	1-Cov	$0.71293_{0.00121}$	$\mathbf{0.74544}_{0.00012}$	$0.03251_{0.00127}$	**+4.56%**
	1-RL	$0.90148_{0.00107}$	$\mathbf{0.91563}_{0.00062}$	$0.01415_{0.00057}$	**+1.57%**
	AP	$0.35824_{0.00086}$	$\mathbf{0.40629}_{0.00115}$	$0.04804_{0.00200}$	**+13.41%**
	AUC	$0.90291_{0.00064}$	$\mathbf{0.91564}_{0.00068}$	$0.01274_{0.00048}$	**+1.41%**
	Macro-F_1	$0.15966_{0.00174}$	$\mathbf{0.21135}_{0.00129}$	$0.05170_{0.00206}$	**+32.38%**
MIR	1-HL	$0.89999_{0.00072}$	$\mathbf{0.90402}_{0.00061}$	$0.00403_{0.00011}$	**+0.45%**
	1-OE	$0.71373_{0.00557}$	$\mathbf{0.74013}_{0.00458}$	$0.02640_{0.00343}$	**+3.70%**
	1-Cov	$0.70861_{0.00090}$	$\mathbf{0.72102}_{0.00161}$	$0.01242_{0.00071}$	**+1.75%**
	1-RL	$0.89814_{0.00071}$	$\mathbf{0.90607}_{0.00061}$	$0.00793_{0.00014}$	**+0.88%**
	AP	$0.65684_{0.00252}$	$\mathbf{0.67794}_{0.00211}$	$0.02109_{0.00203}$	**+3.21%**
	AUC	$0.88174_{0.00086}$	$\mathbf{0.88908}_{0.00083}$	$0.00734_{0.00021}$	**+0.83%**
	Macro-F_1	$0.38238_{0.00606}$	$\mathbf{0.41376}_{0.00686}$	$0.03138_{0.00370}$	**+8.21%**
OBJ	1-HL	$0.95369_{0.00048}$	$\mathbf{0.95542}_{0.00074}$	$0.00172_{0.00076}$	**+0.18%**
	1-OE	$0.71240_{0.00178}$	$\mathbf{0.71901}_{0.00527}$	$0.00661_{0.00600}$	**+0.93%**
	1-Cov	$0.84406_{0.00141}$	$\mathbf{0.84871}_{0.00098}$	$0.00465_{0.00182}$	**+0.55%**
	1-RL	$0.91665_{0.00003}$	$\mathbf{0.92005}_{0.00099}$	$0.00340_{0.00100}$	**+0.37%**
	AP	$0.71112_{0.00229}$	$\mathbf{0.71878}_{0.00429}$	$0.00767_{0.00337}$	**+1.08%**
	AUC	$0.92384_{0.00012}$	$\mathbf{0.92715}_{0.00085}$	$0.00331_{0.00088}$	**+0.36%**
	Macro-F_1	$0.31777_{0.01836}$	$\mathbf{0.33081}_{0.01442}$	$0.01305_{0.00497}$	**+4.11%**
PAS	1-HL	$0.93709_{0.00028}$	$\mathbf{0.93965}_{0.00074}$	$0.00256_{0.00050}$	**+0.27%**
	1-OE	$0.52668_{0.00492}$	$\mathbf{0.55595}_{0.00644}$	$0.02927_{0.00247}$	**+5.56%**
	1-Cov	$0.82067_{0.00392}$	$\mathbf{0.83587}_{0.00490}$	$0.01520_{0.00099}$	**+1.85%**
	1-RL	$0.86658_{0.00364}$	$\mathbf{0.88154}_{0.00473}$	$0.01496_{0.00126}$	**+1.73%**
	AP	$0.61138_{0.00224}$	$\mathbf{0.64110}_{0.00494}$	$0.02971_{0.00280}$	**+4.86%**
	AUC	$0.88176_{0.00413}$	$\mathbf{0.89252}_{0.00434}$	$0.01077_{0.00044}$	**+1.22%**
	Macro-F_1	$0.36920_{0.01075}$	$\mathbf{0.39431}_{0.00484}$	$0.02511_{0.01003}$	**+6.80%**



(a) Corel 5k, PER=30% (b) Corel 5k, PER=50% (c) Corel 5k, PER=70% (d) Corel 5k, PER=90%

(e) ESPGame, PER=30% (f) ESPGame, PER=50% (g) ESPGame, PER=70% (h) ESPGame, PER=90%

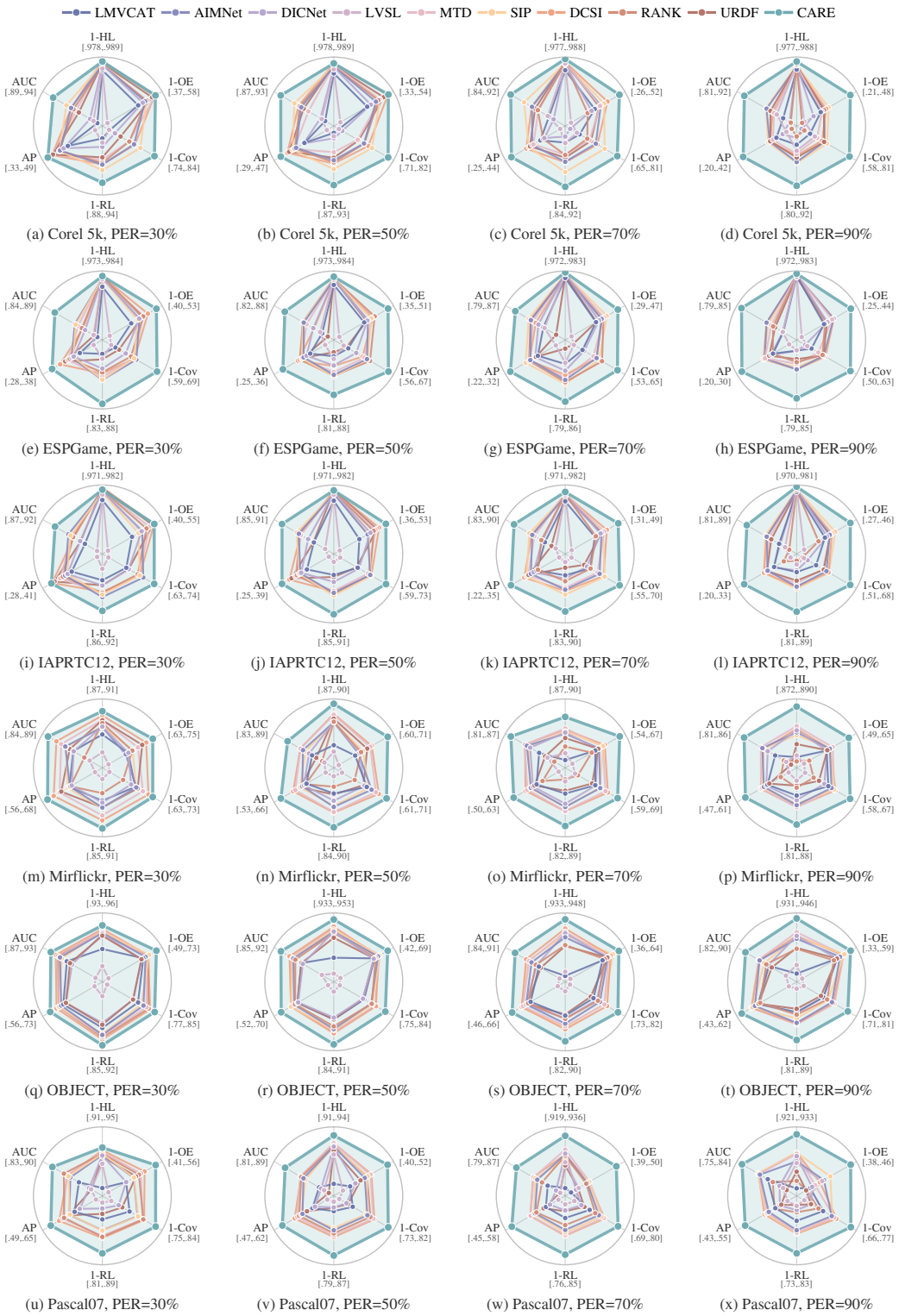
(i) IAPRTC12, PER=30% (j) IAPRTC12, PER=50% (k) IAPRTC12, PER=70% (l) IAPRTC12, PER=90%

(m) Mirflickr, PER=30% (n) Mirflickr, PER=50% (o) Mirflickr, PER=70% (p) Mirflickr, PER=90%

(q) OBJECT, PER=30% (r) OBJECT, PER=50% (s) OBJECT, PER=70% (t) OBJECT, PER=90%

(u) Pascal07, PER=30% (v) Pascal07, PER=50% (w) Pascal07, PER=70% (x) Pascal07, PER=90%

Figure 9: Six-metric comparisons at LMR $= 30\%$ and PER from 30% to 90%.



(a) Corel 5k, PER=30% (b) Corel 5k, PER=50% (c) Corel 5k, PER=70% (d) Corel 5k, PER=90%

(e) ESPGame, PER=30% (f) ESPGame, PER=50% (g) ESPGame, PER=70% (h) ESPGame, PER=90%

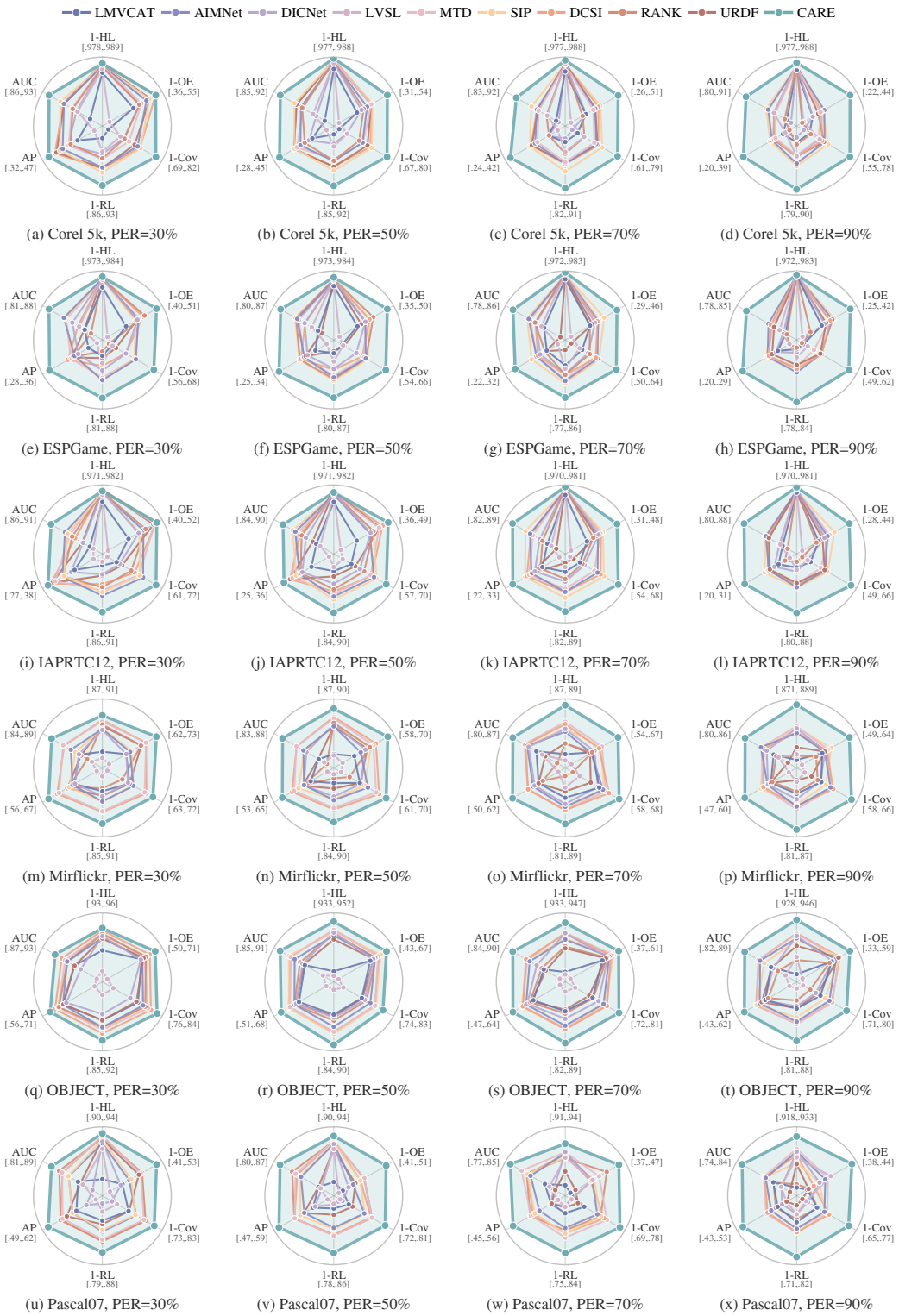
(i) IAPRTC12, PER=30% (j) IAPRTC12, PER=50% (k) IAPRTC12, PER=70% (l) IAPRTC12, PER=90%

(m) Mirflickr, PER=30% (n) Mirflickr, PER=50% (o) Mirflickr, PER=70% (p) Mirflickr, PER=90%

(q) OBJECT, PER=30% (r) OBJECT, PER=50% (s) OBJECT, PER=70% (t) OBJECT, PER=90%

(u) Pascal07, PER=30% (v) Pascal07, PER=50% (w) Pascal07, PER=70% (x) Pascal07, PER=90%

Figure 10: Six-metric comparisons at LMR $= 50\%$ and PER from 30% to 90%.



(a) Corel 5k, PER=30% (b) Corel 5k, PER=50% (c) Corel 5k, PER=70% (d) Corel 5k, PER=90%

(e) ESPGame, PER=30% (f) ESPGame, PER=50% (g) ESPGame, PER=70% (h) ESPGame, PER=90%

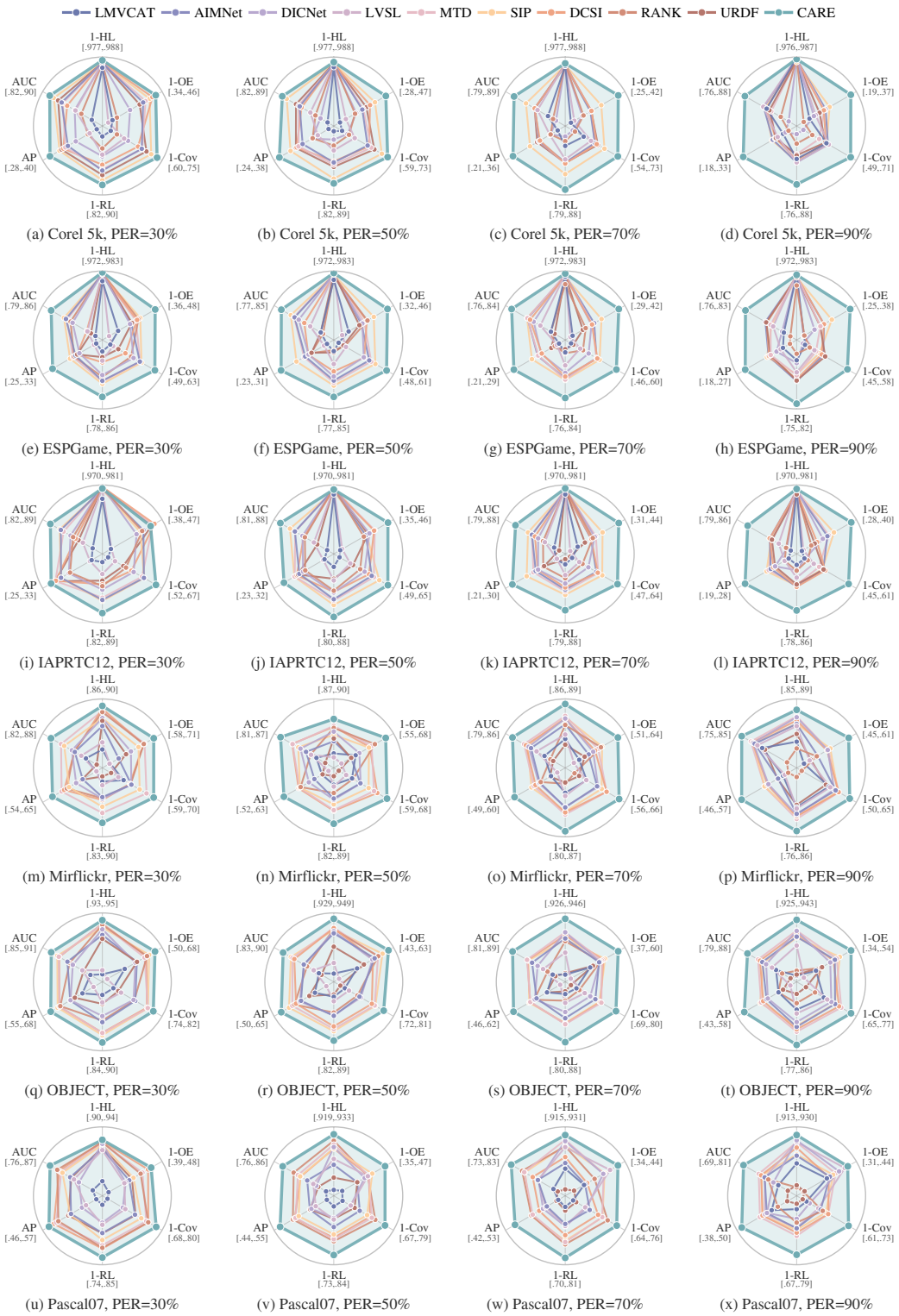
(i) IAPRTC12, PER=30% (j) IAPRTC12, PER=50% (k) IAPRTC12, PER=70% (l) IAPRTC12, PER=90%

(m) Mirflickr, PER=30% (n) Mirflickr, PER=50% (o) Mirflickr, PER=70% (p) Mirflickr, PER=90%

(q) OBJECT, PER=30% (r) OBJECT, PER=50% (s) OBJECT, PER=70% (t) OBJECT, PER=90%

(u) Pascal07, PER=30% (v) Pascal07, PER=50% (w) Pascal07, PER=70% (x) Pascal07, PER=90%

Figure 11: Six-metric comparisons at LMR $= 70\%$ and PER from 30% to 90%.



(a) Corel 5k, PER=30% (b) Corel 5k, PER=50% (c) Corel 5k, PER=70% (d) Corel 5k, PER=90%

(e) ESPGame, PER=30% (f) ESPGame, PER=50% (g) ESPGame, PER=70% (h) ESPGame, PER=90%

(i) IAPRTC12, PER=30% (j) IAPRTC12, PER=50% (k) IAPRTC12, PER=70% (l) IAPRTC12, PER=90%

(m) Mirflickr, PER=30% (n) Mirflickr, PER=50% (o) Mirflickr, PER=70% (p) Mirflickr, PER=90%

(q) OBJECT, PER=30% (r) OBJECT, PER=50% (s) OBJECT, PER=70% (t) OBJECT, PER=90%

(u) Pascal07, PER=30% (v) Pascal07, PER=50% (w) Pascal07, PER=70% (x) Pascal07, PER=90%

Figure 12: Six-metric comparisons at LMR $= 90\%$ and PER from 30% to 90%.