USIL: UNCERTAINTY-AWARE SELECTIVE INVARIANT LEARNING FOR LOW-LIGHT VISIBLE-INFRARED PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light visible–infrared person re-identification (VI-ReID) aims to retrieve the same pedestrian across heterogeneous cameras under severe illumination degradation. However, low-light corruption and cross-spectral discrepancies make identity evidence spatially non-uniform in reliability. When such cues are treated indiscriminately, unreliable local observations may be aligned, enhanced, and eventually propagated into erroneous retrieval rankings. To address this problem, we propose Uncertainty-Aware Selective Invariant Learning (USIL), a unified framework that couples local identity-cue reliability modeling, reliability-guided representation enhancement, and distributional retrieval-risk regularization. Specifically, USIL first employs Uncertainty-Aware Selective Invariant Modeling to selectively preserve stable identity cues and decode location-wise reliability from input-dependent selective-state representations together with its complementary uncertainty. The resulting reliability is then reused by Uncertainty-Guided Low-Light Enhancement to regulate feature compensation and descriptor aggregation. Finally, Distributional Cross-modal Retrieval-Risk Regularization constructs sample-wise cross-modal retrieval risks from ranking violations and regularizes the high-risk tail of the resulting mini-batch risk distribution using a Bayesian-quadrature-inspired Dirichlet reweighting scheme. Extensive experiments on three standard VI-ReID benchmarks demonstrate consistent improvements over the reproduced IRL baseline, including gains of 0.77 percentage points in Rank-1 and 1.10 percentage points in mAP on SYSU-MM01 All-Search.

1 INTRODUCTION

Deep person re-identification (ReID) models Ye et al. (2021c); Wu et al. (2017) have achieved substantial progress in retrieving identities across non-overlapping cameras. Visible cameras capture color, texture, and fine-grained appearance, whereas infrared sensors preserve complementary intensity, contour, and structural information when visible illumination is insufficient Nguyen et al. (2017); Ye et al. (2021b). This complementary sensing has made visible–infrared person ReID (VI-ReID) important for round-the-clock surveillance, and recent studies have advanced low-light benchmarks, modality unification, knowledge transfer, invariant representation learning, and semantic alignment Zhang & Wang (2023); Yu et al. (2023); Ren & Zhang (2024); Wang et al. (2025); Dong et al. (2025). However, low-light VI-ReID is not conventional VI-ReID with merely a larger modality gap: illumination degradation and spectral discrepancy jointly alter which local observations still provide identity-consistent evidence. To examine this effect directly, we gradually decrease the exposure of visible gallery samples and evaluate the reproduced IRL baseline Wang et al. (2025) and USIL on SYSU-MM01. As shown in Fig. 1, both Rank-1 and mAP consistently decrease as visible-light evidence is reduced, revealing the importance of preserving reliable identity information under severe illumination degradation.

Existing VI-ReID methods learn cross-modal correspondence through shared embeddings, modality translation, part-level regularization, or pairwise interaction Kim et al. (2023); Jiang et al. (2024); Xu & Niu (2026). IRL further extracts modality- and scale-invariant cues and enhances low-light representations Wang et al. (2025). Although these approaches reduce average cross-modal discrepancy, they commonly lack an explicit distinction between identity-reliable and unreliable local

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.

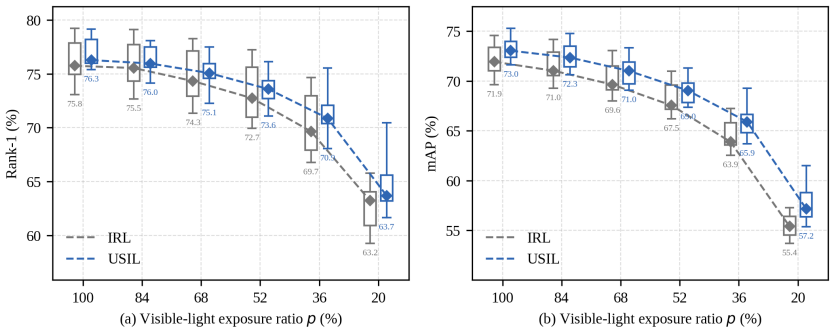


Figure 1: Effect of visible-light exposure on cross-modal retrieval. As exposure decreases on SYSU-MM01, both IRL Wang et al. (2025) and USIL degrade, while USIL consistently maintains higher Rank-1 and mAP over ten gallery trials.

evidence. Under low illumination, strong activation does not necessarily imply reliable identity information: color and texture may disappear, material boundaries may become ambiguous, and reflections, sensor noise, or background responses may remain highly activated. If such unreliable cues are treated indiscriminately during alignment, enhancement, and aggregation, they may contaminate the resulting identity descriptor and eventually induce cross-modal ranking violations. We refer to the uncertainty that a local observation represents identity-consistent evidence rather than illumination corruption, missing spectral information, or background interference as *cross-modal identity-cue uncertainty*. As illustrated in Fig. 2(a), this uncertainty is spatially non-uniform across modalities and samples. Fig. 2(b) further motivates two complementary modeling scales: locally, the model must determine which identity cues remain reliable; globally, unreliable evidence propagated into the identity descriptor may cause a hard negative to outrank the true cross-modal match. Such violations define sample-wise retrieval risk, whose high-risk tail can be regularized at the mini-batch distribution level. A held-out diagnostic in Appendix E.4 further shows that location-wise reliability decoded from Selective SSM states is strongly associated with local cross-modal identity consistency, supporting selective state propagation as the basis of local uncertainty modeling.

Motivated by this reliability-to-risk view, we propose **Uncertainty-Aware Selective Invariant Learning (USIL)**, a unified framework that connects location-wise identity-cue reliability modeling with distributional retrieval-risk regularization, as summarized in Fig. 2(c). First, **Uncertainty-Aware Selective Invariant Modeling (USIM)** organizes deep pedestrian features as spatial sequences and performs bidirectional, input-dependent state propagation following the content-selection principle of Mamba Gu & Dao (2023). Rather than treating raw selective-state parameters as uncertainty, USIM decodes location-wise identity-cue reliability from the resulting selective-state representations and defines its complementary uncertainty. Second, **Uncertainty-Guided Low-Light Enhancement (UGLE)** reuses the same reliability state to regulate low-light feature compensation and reliability-weighted descriptor aggregation, preventing unreliable local responses from being uniformly amplified. Finally, **Distributional Cross-modal Retrieval-Risk Regularization (DCRR)** constructs sample-wise retrieval risks from cross-modal ranking violations and regularizes the high-risk tail of the resulting mini-batch risk distribution through a Bayesian-quadrature-inspired Dirichlet reweighting scheme Snell & Griffiths (2025). The resulting retrieval-risk supervision provides optimization feedback to both reliability estimation and identity representation learning, thereby forming a continuous reliability-to-risk learning flow instead of a stack of independent modules.

Extensive experiments on three standard VI-ReID benchmarks, including SYSU-MM01, RegDB, and LLCM, consistently demonstrate the effectiveness of USIL over the reproduced IRL baseline. In particular, USIL improves Rank-1 and mAP by 0.77 and 1.10 percentage points, respectively, on the SYSU-MM01 All-Search setting. Moreover, as shown in Fig. 1, USIL consistently maintains higher Rank-1 and mAP than the reproduced IRL baseline under progressively reduced visible-light exposure, indicating improved robustness to illumination degradation.

Our contributions are summarized as follows:

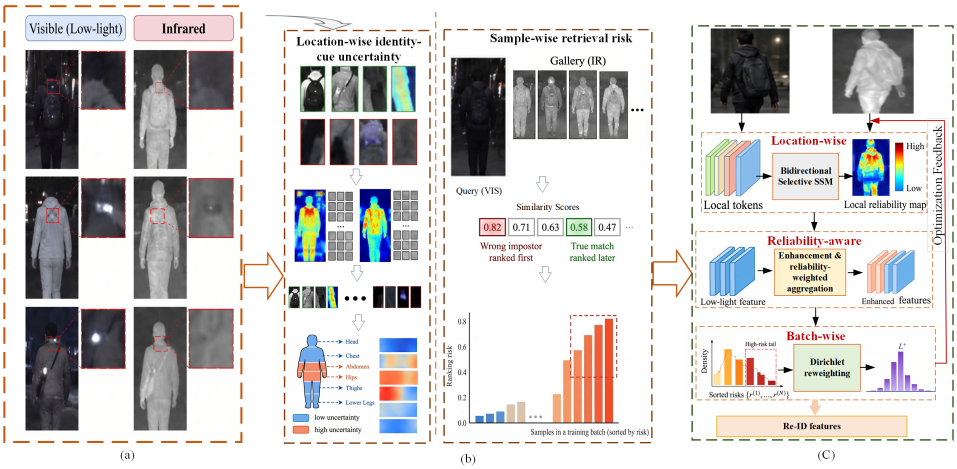


Figure 2: Motivation and reliability–risk structure of USIL. (a) Low-light cross-modal identity cues are spatially non-uniform. (b) USIL models location-wise reliability/uncertainty and sample-wise retrieval risk with Dirichlet-reweighted tail regularization. (c) Reliability guides enhancement and aggregation, while retrieval risk provides optimization feedback.

- We formulate *cross-modal identity-cue uncertainty* for low-light VI-ReID and reveal a reliability-to-risk propagation process in which spatially unreliable observations can enter representation learning and eventually induce erroneous cross-modal rankings.
- We propose a unified USIM–UGLE mechanism that uses content-adaptive bidirectional state propagation to decode location-wise identity-cue reliability and reuses the same reliability state to regulate low-light feature compensation and descriptor aggregation.
- We develop DCRR, a Bayesian-quadrature-inspired distributional cross-modal retrieval-risk regularizer that constructs sample-wise risks from ranking violations and emphasizes the high-risk tail of the mini-batch risk distribution through Dirichlet reweighting.

2 Related Work

Low-Light VI-ReID and Selective State-Space Modeling. Early RGB–infrared and visible–thermal benchmarks established heterogeneous cross-camera matching protocols Wu et al. (2017); Nguyen et al. (2017), followed by channel augmentation, embedding expansion, and modality unification Ye et al. (2021b); Zhang & Wang (2023); Yu et al. (2023). Recent studies further address low-light enhancement, privacy-preserving deployment, any-scenario multispectral retrieval, and adaptive frequency modeling Wang et al. (2025); Jiang et al. (2025); Ground; Li et al. (2026), while local and complementary representations are improved through part discovery, modality-specific knowledge transfer, domain shifting, foundation-model semantics, base–detail learning, diffusion synthesis, dual-level matching, pairwise interaction, and spatial–frequency modeling Kim et al. (2023); Ren & Zhang (2024); Jiang et al. (2024); Hu et al. (2024); Dong et al. (2025); Gong et al. (2025); Dai et al. (2025); Ye et al. (2025); Xu & Niu (2026); Yu et al. (2026). Selective SSMs provide input-dependent state propagation Gu & Dao (2023), with Vision Mamba, VMamba, and ReIDMamba extending this principle to visual and person representation Zhu et al. (2024); Liu et al. (2024); Gu et al. (2026); however, these methods do not maintain a location-wise uncertainty state shared by cue selection, enhancement, and aggregation, which motivates the uncertainty-aware bidirectional modeling in USIM.

Uncertainty Modeling and Distributional Retrieval Risk. Uncertainty-aware person retrieval models prediction reliability Dou et al. (2022), while calibration methods assess confidence consistency Silva Filho et al. (2023). Conformal risk control provides distribution-free control for bounded losses Angelopoulos et al. (2024) and has been extended to image retrieval and retrieve–then-rank systems Cai et al. (2025); Xu et al. (2024); conformal self-calibration further addresses low-quality multimodal learning Jiang et al. (2026), while Bayesian quadrature represents plausible risks through Dirichlet quantile spacings Snell & Griffiths (2025). These approaches do not directly couple location-wise identity-cue uncertainty with low-light enhancement and cross-modal ranking-risk

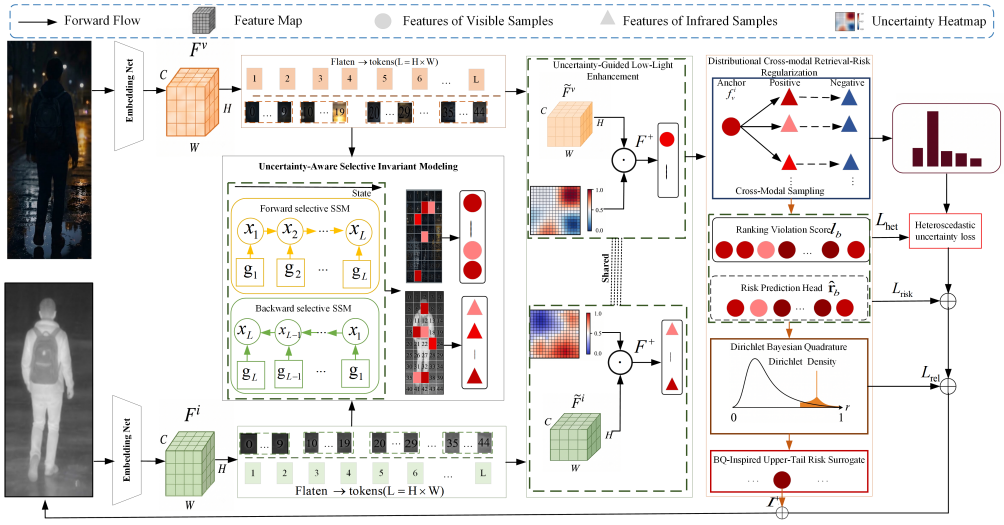


Figure 3: Overall framework of USIL. USIM models location-wise uncertainty/reliability via bidirectional selective state propagation; UGLE uses this reliability to guide low-light enhancement and aggregation; DCRR regularizes cross-modal ranking risk through Bayesian-quadrature-inspired Dirichlet reweighting, forming an uncertainty-to-risk feedback flow.

regularization; DCRR therefore employs a Bayesian-quadrature-inspired upper-tail risk surrogate together with USIM and UGLE, without claiming a deployment-time conformal guarantee.

3 Proposed Method

3.1 Preliminaries

Given a visible-infrared training set $\mathcal{D} = \{(x_b, y_b, m_b)\}_{b=1}^{N}$, $x_b \in \mathbb{R}^{H_0 \times W_0 \times 3}$ denotes a pedestrian image, $y_b \in \mathcal{Y} = \{1, \ldots, N_{\mathrm{id}}\}$ is its identity label, and $m_b \in \{v, i\}$ indicates the visible or infrared modality. A dual-stream invariant backbone $f_\theta(\cdot)$ Wang et al. (2025); Ye et al. (2021c) is used to extract deep features. The backbone maps x_b to the output of its third residual stage, immediately before the final stage; this spatial map is unfolded into a body-region token sequence,

$$F_b = f_\theta(x_b) \in \mathbb{R}^{C \times H \times W}, \quad X_b = \mathrm{flatten}(F_b) = \{x_b^1, \ldots, x_b^L\} \in \mathbb{R}^{L \times C}, \tag{1}$$

where $L = H \times W$ and each $x_b^q \in \mathbb{R}^C$ is a local body-region token that preserves fine-grained identity evidence. After that, USIM refines X_b into a selective invariant state and decodes a location-wise reliability $s_b \in (0,1)^{H \times W}$ together with its complementary identity-cue uncertainty $u_b = \mathbf{1} - s_b$ (Sec. 3.2). The remaining enhancement path and fourth residual stage then produce G_b, which reliability-weighted generalized-mean pooling Radenović et al. (2018) (Sec. 3.3) aggregates into a global identity embedding $f_b \in \mathbb{R}^d$. A shared classifier C receives its batch-normalized counterpart and produces identity logits,

$$\mathbf{z}_b = C(\mathrm{BN}(f_b)) \in \mathbb{R}^{N_{\mathrm{id}}}, \qquad p_{b,c} = \frac{\exp(z_{b,c})}{\sum_{c'} \exp(z_{b,c'})}. \tag{2}$$

The embedding f_b is used for identity classification and cross-modal retrieval, i.e., cross-modal same-identity pairs are expected to be closer than cross-identity pairs under a distance $d(\cdot, \cdot)$.

Unlike conventional VI-ReID that treats every activated token as equally trustworthy, illumination degradation and spectral discrepancy make cross-modal identity evidence spatially non-uniform, so a visible or infrared response at a given location need not convey the same identity. USIL therefore exploits the reliability s_b to decide which cross-spectral cues should be enhanced, aggregated, and trusted during retrieval, preventing low-light and modality-specific corruption from propagating into f_b.

3.2 Uncertainty-Aware Selective Invariant Modeling

Since the reliability s_b is not annotated, USIM estimates it while selectively refining identity semantics. We unfold the deep feature map F_b into a spatial sequence of $L{=}H{\times}W$ tokens and project it,

$$Q_b = \mathrm{Proj}_{\mathrm{in}}\big(\,\mathrm{LN}(\mathrm{flatten}(F_b))\big) \in \mathbb{R}^{L\times d}. \tag{3}$$

Following the selective state-space principle Gu & Dao (2023), each token updates an input-dependent hidden state and emits an output,

$$o_t = \bar{A}_t o_{t-1} + \bar{B}_t Q_{b,t}, \qquad \tilde{y}_t = C_t o_t, \tag{4}$$

where $(\bar{A}_t, \bar{B}_t, C_t)$ are generated from the input, allowing the model to propagate stable structures and forget modality-specific nuisance. To respect the non-causal nature of ReID, we scan bidirectionally Zhu et al. (2024); Liu et al. (2024) and fuse the two directions,

$$Y_b = \mathrm{LN}\big(\,\mathrm{SSM}_{\rightarrow}(Q_b) + \mathrm{flip}(\mathrm{SSM}_{\leftarrow}(\mathrm{flip}(Q_b)))\big). \tag{5}$$

The selective state is decoded into a residual refinement and a reliability map,

$$\tilde{F}_b = F_b + \gamma\ \mathrm{Proj}_{\mathrm{out}}(Y_b), \qquad s_b = \sigma\big(\,\mathrm{Head}(Y_b)\big), \quad u_b = \mathbf{1} - s_b, \tag{6}$$

where $\mathrm{Proj}_{\mathrm{out}}$ is zero-initialized and γ is a learnable scale, so that USIM starts as an identity mapping and departs from the baseline Wang et al. (2025); Gu et al. (2026) only in a non-destructive way. The reliability s_b is shared by UGLE and DCRR.

3.3 Uncertainty-Guided Low-Light Enhancement

UGLE reuses the reliability s_b from Eq. equation 6 to jointly govern low-light compensation and descriptor aggregation, so that unreliable regions are not blindly amplified. We first convert reliability into a compensation modulation centered at one,

$$\mathrm{mod}_b = 1 + \rho\,(s_b - \tfrac{1}{2}), \qquad \tilde{F}_b^{+} = \tilde{F}_b + \delta\,\mathrm{mod}_b \odot \mathrm{NAF}(\tilde{F}_b), \tag{7}$$

where $\mathrm{NAF}(\cdot)$ is the low-light compensation branch Chen et al. (2022); Wang et al. (2025), $\odot$ is the element-wise product, and ρ, δ are learnable scales. Equation equation 7 is compact notation for the two compensated enhancement stages in the implementation: the first uses the learnable channel-wise δ, while the second uses the fixed IRL coefficient 0.3; both reuse the same mod_b. Their output is propagated through the remaining enhancement and fourth residual stage to obtain G_b. Since $\mathrm{mod}_b \to 1$ when $\rho \to 0$, Eq. equation 7 recovers the original compensation as a special case. We then aggregate the enhanced feature with a reliability-weighted generalized-mean pooling Radenović et al. (2018),

$$w_b = \big(1 + \alpha(s_b/\bar{s}_b - 1)\big)_{\geq \eta_0}, \qquad f_b = \Big(\frac{\sum w_b \odot (G_b)^p}{\sum w_b}\Big)^{1/p}, \tag{8}$$

where s_b is bilinearly resized to the resolution of G_b before forming w_b, $\bar{s}_b$ is its spatial mean, p is the pooling exponent, and η_0 clamps the weight positive so that Eq. equation 8 reduces to uniform pooling when $\alpha \to 0$. We further summarize a per-sample uncertainty $\bar{u}_b = 1 - \bar{s}_b$ for the feedback in DCRR. UGLE thus lets a single reliability state control both enhancement and aggregation, rather than introducing an independent attention branch.

3.4 Distributional Cross-modal Retrieval-Risk Regularization

DCRR turns the identity embedding into a differentiable cross-modal retrieval risk and constrains its high-risk tail. Let $S_{ab} = \langle \hat{f}_a, \hat{f}_b \rangle$ be the cosine similarity between ℓ_2-normalized embeddings. For an anchor b, we define its cross-modal positives $\mathcal{P}_b = \{k : y_k = y_b, m_k \neq m_b\}$ and negatives $\mathcal{N}_b = \{k : y_k \neq y_b, m_k \neq m_b\}$, and measure the soft ranking violation Dou et al. (2022),

$$r_b = \frac{1}{|\mathcal{N}_b|} \sum_{k\in\mathcal{N}_b} \sigma\Big(\frac{S_{bk} - \min_{p\in\mathcal{P}_b} S_{bp}}{\tau}\Big) \in (0,1), \tag{9}$$

where τ is a temperature and $r_b \to 1$ indicates that negatives are ranked above the hardest cross-modal positive. In parallel, a risk head ψ predicts a smoothly bounded per-sample risk,

$$\ell_b = 4\tanh\big(\psi(f_b)/4\big), \qquad \hat{r}_b = \sigma(\ell_b), \tag{10}$$

so that $\ell_b \in (-4, 4)$ keeps $e^{-\ell_b}$ numerically safe with non-vanishing gradients. To constrain the tail rather than the mean, we follow the Bayesian-quadrature view of conformal risk Snell & Griffiths (2025); Angelopoulos et al. (2024); Cai et al. (2025)and draw M Dirichlet quantile spacings,

$$\mathbf{w}^{(j)} \sim \mathrm{Dirichlet}(\mathbf{1}_{N+1}), \quad L^+ = \mathrm{Quantile}_\eta\Big(\textstyle\sum_{b=1}^{N} w_b^{(j)}\, r_{(b)} + w_{N+1}^{(j)}\, r_{\max} \Big), \tag{11}$$

where $r_{(1)} \le \cdots \le r_{(N)}$ are ordered batch risks, $r_{\max}$ is a worst-case risk, and η is the upper-quantile level. L^+ is a BQ-inspired upper-tail surrogate of the cross-modal mismatch risk. Its gradient is retained so that ranking violations directly regularize the learned identity representation; we do not claim a deployment-time conformal guarantee.

3.5 Theoretical Properties

We summarize the algorithmic properties used to justify the three components; complete proofs are given in Appendix B. **Proposition 1 (uncertainty–reliability complementarity).** Since $s_{bq} = \sigma(a_{bq})$ and $u_{bq} = 1 - s_{bq}$, both scores lie in $(0, 1)$, sum to one, and vary smoothly with the selective state. **Proposition 2 (UGLE stability).** If $|\rho| < 2$, the NAF modulation is strictly positive and bounded by $1 \pm |\rho|/2$, while every pooling weight is at least η_0. **Proposition 3 (conditional state stability).** If the selective transition is uniformly contractive, $\|\bar{A}_t\|_{\mathrm{op}} \le a < 1$, with bounded input and injection operators, then every directional USIM state is uniformly bounded. **Theorem 1 (linear spatial scaling).** For fixed channel and state dimensions, bidirectional USIM has $\mathcal{O}(HW)$ time and activation-memory complexity with respect to its spatial sequence length.

Lemma 1 (soft ranking risk). For an anchor with valid cross-modal positives and negatives, $r_b \in (0, 1)$; away from ties, it increases with every negative similarity and decreases with the hardest-positive similarity. **Theorem 2 (Dirichlet surrogate).** Conditioned on a fixed batch, the implemented Dirichlet functional is a convex combination of the ordered risks and $r_{\max}$, hence lies in $[0, 1]$; its mean and variance are available in closed form, and the M-sample empirical quantile converges to the corresponding conditional quantile under standard quantile regularity. **Corollary 1 (feedback connectivity).** The stop-gradient placements retain three distinct optimization paths: observed risk supervises the risk head, risk and correctness supervise spatial reliability, and the non-detached L^+ branch regularizes the descriptor. These properties characterize the training surrogate and do not constitute a conformal or deployment-risk guarantee.

3.6 Training

The invariant backbone, USIM, UGLE, and DCRR are jointly optimized end-to-end. The standard identity and triplet objectives preserve identity discrimination, while the proposed terms couple uncertainty estimation, enhancement, and cross-modal retrieval risk. Given the predicted log-risk ℓ_b from the DCRR risk head, a heteroscedastic uncertainty loss Kendall & Gal (2017) relaxes high-risk samples,

$$\mathcal{L}_{\mathrm{het}} = \frac{1}{B} \sum_{b=1}^{B} \Big(\tfrac{1}{2} e^{-\ell_b}\, \mathrm{CE}(\mathbf{z}_b, y_b) + \tfrac{1}{2}\, \mathrm{softplus}(\ell_b) \Big), \tag{12}$$

while risk-prediction fitting and reliability feedback couple the predicted risk, observed ranking risk, and USIM reliability. To prevent a spatially constant reliability solution, we partition the layer-3 feature into six horizontal body stripes and form detached stripe descriptors h_{bk}. Without pixel-level registration (but assuming coarse-stripe semantic correspondence), their cross-modal identity-margin target compares the least-similar cross-modal positive with the most-similar cross-modal negative per stripe,

$$t_{bk}^{\mathrm{loc}} = \sigma\left(\frac{\min_{p \in \mathcal{P}_b} \langle \hat{h}_{bk}, \hat{h}_{pk} \rangle - \max_{n \in \mathcal{N}_b} \langle \hat{h}_{bk}, \hat{h}_{nk} \rangle}{\tau_{\mathrm{loc}}} \right), \tag{13}$$

where the target branch is stop-gradient and $\bar{s}_{bk}^{\mathrm{stripe}}$ is the mean USIM reliability in stripe k. The sample-level reliability feedback is

$$\mathcal{L}_{\mathrm{risk}} = \mathrm{MSE}\big(\hat{r}, \mathrm{sg}[r]\big), \qquad \mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}} = \mathrm{MSE}\big(\bar{s}, \mathrm{sg}[t]\big) + \mathrm{MSE}\big(\hat{r}, 1 - \bar{s}\big), \tag{14}$$

where $t_b = \mathbb{1}(\hat{y}_b = y_b)(1 - r_b)$ and $\mathrm{sg}[\cdot]$ denotes stop-gradient. The *local* reliability supervision is deliberately a *ranking* objective rather than an absolute-value regression: because the absolute reliability level is uninformative, an MSE against t^{loc} is minimized by predicting a spatially constant map, so the discriminative signal (the relative ordering of the six stripes) is discarded. We therefore mean-center the predicted reliability and the target across the six stripes and maximize their alignment,

$$\mathcal{L}_{\mathrm{rel}}^{\mathrm{loc}} = \frac{1}{|\mathcal{V}|} \sum_{b \in \mathcal{V}} \Big(1 - \cos\left(\tilde{s}_b^{\mathrm{stripe}}, \, \mathrm{sg}[\tilde{t}_b^{\mathrm{loc}}] \right) \Big), \tag{15}$$

where $\tilde{\cdot}$ denotes stripe-wise mean removal and $\mathcal{V}$ is the set of samples with a valid cross-modal positive/negative pair. This listwise alignment cannot be satisfied by a constant map and directly optimizes the reliability–margin correlation reported in Appendix E.4. We regard this local ordering constraint as an internal component of hierarchical reliability learning rather than an additional top-level objective:

$$\mathcal{L}_{\mathrm{rel}} = \mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}} + \kappa \mathcal{L}_{\mathrm{rel}}^{\mathrm{loc}}, \tag{16}$$

where κ controls the local component inside reliability learning. Finally, L^+ from Eq. equation 11 constrains the high-risk tail of observed ranking violations. The proposed objective is

$$\mathcal{L}_{\mathrm{USIL}} = \lambda_{\mathrm{het}} \mathcal{L}_{\mathrm{het}} + \lambda_{\mathrm{risk}} \mathcal{L}_{\mathrm{risk}} + \lambda_{\mathrm{rel}} \mathcal{L}_{\mathrm{rel}} + \lambda^+ L^+, \tag{17}$$

where the four coefficients balance uncertainty-aware identity weighting, risk-prediction fitting, hierarchical reliability learning, and upper-tail risk control, respectively. We integrate the task and USIL objectives as

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{USIL}}, \tag{18}$$

where $\mathcal{L}_{\mathrm{task}} = \mathcal{L}_{\mathrm{id}} + \mathcal{L}_{\mathrm{tri}}$ is the primary VI-ReID task loss. The implementation-specific warm-up and risk-ramp schedules are detailed in Appendix C. Thus, every innovation is tied to the same identity representation: USIM selects stable evidence, UGLE enhances and aggregates it according to reliability, and DCRR uses observed ordering violations to regularize both the descriptor and its uncertainty estimate.

4 Experiments

4.1 Datasets and Experimental Settings

Experimental setup. We evaluate USIL on three public VI-ReID benchmarks: SYSU-MM01 (Wu et al., 2017), LLCM (Zhang & Wang, 2023), and RegDB Nguyen et al. (2017). SYSU-MM01 uses 395/96 training/testing identities under all-search and indoor-search protocols, LLCM uses 713/351 identities, and RegDB uses 206/206 identities; LLCM and RegDB are evaluated in both visible-to-infrared and infrared-to-visible directions. Following standard VI-ReID practice Ye et al. (2021c); Wang et al. (2025), we report CMC and mAP, with SYSU-MM01 and LLCM averaged over 10 gallery trials and RegDB over its 10 official splits. We adopt the AGW backbone used by IRL Wang et al. (2025) and augment it with bidirectional Mamba-based USIM Gu & Dao (2023); images are resized to 384×192, training uses CAJ Ye et al. (2021b) and standard augmentations, and SGD is run for 150 epochs with an initial learning rate of 0.1. USIM uses state dimension/convolution width/expansion ratio $16/4/2$, UGLE uses $p = 3$, and DCRR uses $\tau = 0.1$, 64 Dirichlet draws, and quantile level 0.9 Snell & Griffiths (2025).

4.2 Comparison with State-of-the-Art Methods

Tables 1 and 2 compare USIL with state-of-the-art methods on SYSU-MM01, RegDB, and LLCM, with IRL reproduced under the same evaluation protocols for controlled comparison. On SYSU-MM01, USIL improves Rank-1/mAP from 76.23%/72.08% to 77.00%/73.18% under all-search and gains 2.07/1.47 percentage points under indoor-search, achieving the best Rank-1 among the listed methods in both settings. On RegDB, it improves Rank-1/mAP from 90.35%/81.90% to 90.98%/82.44% for visible-to-infrared retrieval and from 87.82%/79.29% to 88.54%/80.08% in the reverse direction; on LLCM, the corresponding gains are 0.47/0.43 and 0.57/0.52 points, with the best visible-to-infrared Rank-1 among the listed methods.

Table 1: Comparison with state-of-the-art methods on the SYSU-MM01 and RegDB datasets. The best and second-best results are highlighted in **bold** and underlined, respectively.

Methods	Venue	SYSU-MM01						RegDB			
		All-search			Indoor-search			VIS to IR		IR to VIS	
		Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP	Rank-1	mAP	Rank-1	mAP
Zero-Padding (Wu et al., 2017)	ICCV'17	14.80	–	15.95	20.58	–	26.92	17.80	18.90	16.70	17.90
AlignGAN (Wang et al., 2019)	ICCV'19	42.40	–	40.70	45.90	–	54.30	57.90	53.60	56.30	53.40
DDAG Ye et al. (2020)	ECCV'20	54.75	90.39	53.02	61.02	94.06	67.98	69.34	63.46	68.06	61.80
JSIA Wang et al. (2020)	AAAI'20	38.10	80.70	36.90	43.80	86.20	52.90	48.50	49.30	48.10	48.90
cm-SSFT (Lu et al., 2020)	CVPR'20	61.60	–	63.20	70.50	–	72.60	72.30	72.90	71.00	71.70
MMN Zhang et al. (2021)	ACMMM'21	70.60	96.20	66.90	76.20	97.20	79.60	**91.60**	84.10	87.50	80.50
GECNet Zhong et al. (2021)	TCSVT'21	53.37	89.86	51.83	60.60	94.29	62.89	82.33	78.45	78.93	75.58
LBA Park et al. (2021)	ICCV'21	55.41	–	54.14	58.46	–	66.33	74.17	67.64	72.43	65.46
MCLNet (Hao et al., 2021)	ICCV'21	65.40	–	61.98	72.56	–	76.58	80.31	73.07	75.93	69.49
SMCL (Wei et al., 2021b)	ICCV'21	67.39	–	61.78	68.84	–	75.56	83.93	79.83	83.05	78.57
CAJ (Ye et al., 2021b)	ICCV'21	69.88	–	66.89	76.26	–	80.37	85.03	77.82	84.75	77.82
TSME Liu et al. (2022)	TCSVT'22	64.23	95.19	61.21	76.20	96.75	54.36	87.35	76.94	86.41	75.70
FBP Wei et al. (2021a)	TNNLS'22	54.14	86.04	50.20	73.98	89.71	68.24	73.98	68.24	70.05	66.61
DTRM Ye et al. (2021a)	TIFS'22	63.03	93.82	58.63	66.35	95.58	71.76	79.09	70.09	78.02	69.56
FMCNet Zhang et al. (2022a)	CVPR'22	66.34	–	62.51	68.15	–	74.09	89.12	**84.43**	88.38	83.86
PIC (Zheng et al., 2022)	TIP'22	57.51	–	55.14	60.40	–	67.70	83.60	79.60	79.50	77.40
MSCLNet (Zhang et al., 2022c)	ECCV'22	76.99	–	71.64	78.49	–	81.17	84.17	80.99	83.86	78.31
PMT Lu et al. (2023)	AAAI'23	67.53	95.36	64.98	71.66	96.73	76.52	84.83	76.55	84.16	75.13
CMTR Liang et al. (2023)	TMM'23	65.45	94.47	62.90	71.46	97.16	76.67	88.11	81.66	84.92	80.79
DSCNet Zhang et al. (2022b)	TIFS'23	73.89	96.27	69.47	79.35	98.32	82.65	85.39	77.30	83.50	75.19
ACD Pan et al. (2024)	TIFS'24	74.44	–	71.17	78.98	–	82.75	84.71	83.28	87.18	**84.72**
CSM Yang et al. (2024)	TMM'24	70.59	96.22	67.49	75.98	98.12	80.24	85.34	76.39	83.88	75.16
IRL Wang et al. (2025)	ACMMM'25	76.23	96.75	72.08	81.05	98.22	84.11	90.35	81.90	87.82	79.29
WHU-MARS (Ground)	CVPR'26	50.22	–	50.66	56.30	–	61.43	71.17	68.96	71.89	67.18
USIL (Ours)	–	**77.00**	**97.33**	**73.18**	**83.12**	**98.51**	**85.58**	90.98	82.44	**88.54**	80.08

4.3 Ablation Study

We conduct ablation studies on SYSU-MM01 under the All-Search Single-Shot setting using the reproduced IRL model as the baseline (Tab. 3). USIM improves Rank-1/mAP by $+0.31\%/+0.87\%$, showing the benefit of content-adaptive bidirectional state propagation for preserving reliable cross-spectral identity cues. Adding UGLE yields a further $+0.30\%/+0.21\%$ by reusing spatial reliability to guide low-light compensation and descriptor aggregation. With DCRR, USIL achieves the best performance of 77.00% Rank-1 and 73.18% mAP, corresponding to $+0.77\%/+1.10\%$ over the baseline, supporting the effectiveness of upper-tail retrieval-risk regularization and the complete uncertainty-to-risk feedback flow. Additional sensitivity analysis in Fig. 6 shows that performance remains relatively stable when varying λ_{het}, λ^{+}, and ρ under the same 150-epoch protocol.

Table 2: Comparison with state-of-the-art methods on the LLCM dataset. The best and second-best results are highlighted in **bold** and underlined, respectively.

Methods	Venue	VIS to IR		IR to VIS	
		Rank-1	mAP	Rank-1	mAP
DDAG Ye et al. (2020)	ECCV'20	40.30	48.40	48.00	52.30
AGW Ye et al. (2021c)	TPAMI'21	51.50	55.30	43.60	51.80
CAJ Ye et al. (2021b)	ICCV'21	56.50	59.80	48.80	56.60
MMN Zhang et al. (2021)	ACM MM'21	59.90	62.70	52.50	58.90
DART Yang et al. (2022)	CVPR'22	60.40	63.20	52.20	59.80
DCPLNet Chan et al. (2024)	TII'24	60.50	63.20	53.40	59.80
IRL Wang et al. (2025)	ACM MM'25	63.91	66.05	55.06	61.39
USIL (Ours)	–	**64.38**	**66.48**	**55.63**	**61.91**

Qualitative and distributional analyses further explain these gains. Fig. 4 shows that USIL corrects baseline Rank-1 errors under low-light and cross-modal degradation, while USIM assigns higher reliability to stable identity cues such as body contour, structure, and carried objects and suppresses unreliable background, reflection, and modality-specific responses. In Fig. 5 (a), USIL reduces the mean intra-class distance from 0.304 to 0.301 and increases the mean inter-class distance from 0.638 to 0.643, enlarging their separation from 0.334 to 0.342. Fig. 5 (b) further shows more compact cross-modal clusters and shorter visible–infrared centroid distances for representative identities

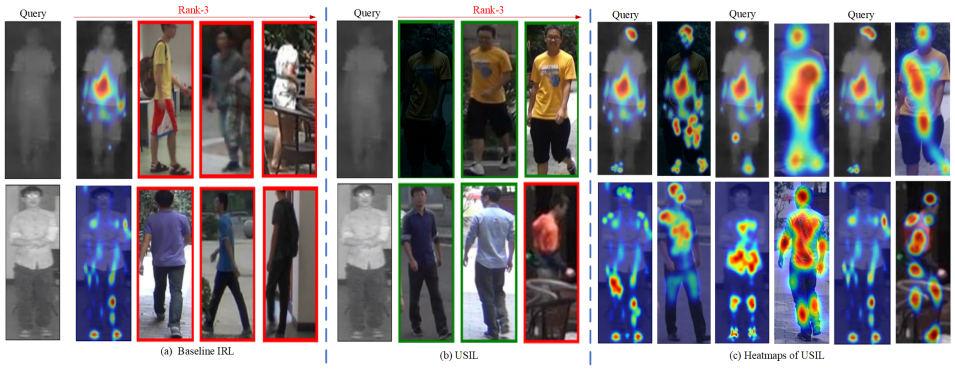


Figure 4: Qualitative comparison of retrieval results and spatial reliability maps. (a) Rank-3 retrieval results of the reproduced IRL baseline together with the USIM query-reliability map. (b) Rank-3 retrieval results of USIL. (c) USIM reliability maps for the USIL-retrieved gallery images.

selected from the 236 queries whose baseline Rank-1 errors are corrected by USIL, supporting improved cross-modal compactness and discriminability.

Table 3: Ablation studies of each component on SYSU-MM01 (Single-Shot All-Search mode).

Index	Base	USIM	UGLE	DCRR	SYSU-MM01	
					Rank-1	mAP
1	✓	×	×	×	76.23	72.08
2	✓	✓	×	×	76.54	72.95
3	✓	✓	✓	×	76.84	73.16
4	✓	✓	✓	✓	**77.00**	**73.18**

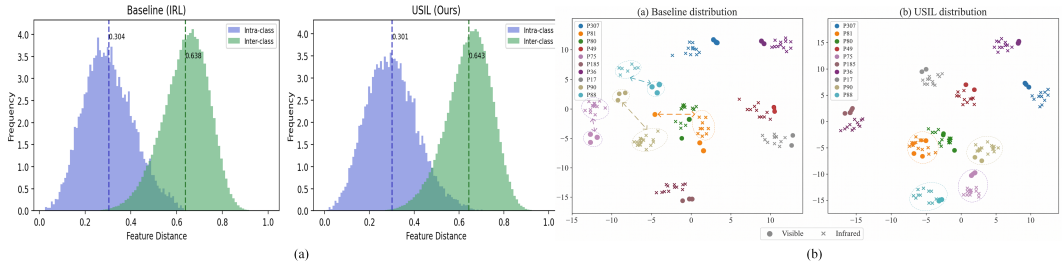


Figure 5: Feature-distance distributions and t-SNE visualizations on SYSU-MM01. (a) Distributions of intra-class and inter-class feature distances for the baseline (IRL) and USIL during inference, with dashed vertical lines indicating their respective mean distances. (b) t-SNE visualizations of the baseline and USIL. Colors represent different identities, while filled circles and crosses denote visible and infrared samples, respectively. Dashed ellipses highlight feature clusters.

5 Conclusion

We propose a novel Uncertainty-Aware Selective Invariant Learning (USIL) framework for low-light VI-ReID, which leverages an Uncertainty-Aware Selective Invariant Modeling (USIM) module to retain stable identity cues while predicting location-wise uncertainty, an Uncertainty-Guided Low-Light Enhancement (UGLE) module to modulate compensation and descriptor aggregation by the same reliability state, and a Distributional Cross-modal Retrieval-Risk Regularization (DCRR) to bound ranking violations and back-correct the uncertainty. By introducing this continuous uncertainty-to-risk feedback flow, USIL effectively suppresses illumination-induced identity-cue ambiguity, even under severe visible–infrared discrepancy. Extensive experiments demonstrate that USIL consistently outperforms existing rigid invariant-learning approaches by a significant margin.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Kaiwen Cai, Chris Xiaoxuan Lu, Xingyu Zhao, Wei Huang, and Xiaowei Huang. Risk controlled image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27224–27232, 2025.

Sixian Chan, Weihao Meng, Cong Bai, Jie Hu, and Shenyong Chen. Diverse-feature collaborative progressive learning for visible-infrared person re-identification. *IEEE Transactions on Industrial Informatics*, 20(5):7754–7763, 2024.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Wenbo Dai, Lijing Lu, and Zhihang Li. Diffusion-based synthetic data generation for visible-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11185–11193, 2025.

Neng Dong, Shuanglin Yan, Liyan Zhang, and Jinhui Tang. Richer semantics, better alignment: Aligning visual features with explicit and enriched semantics for visible-infrared person re-identification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 927–935, 2025.

Zhaopeng Dou, Zhongdao Wang, Weihua Chen, Yali Li, and Shengjin Wang. Reliability-aware prediction via uncertainty learning for person image retrieval. In *European Conference on Computer Vision*, pp. 588–605. Springer, 2022.

Zhihao Gong, Lian Wu, and Yong Xu. Base-detail feature learning framework for visible-infrared person re-identification. *arXiv preprint arXiv:2505.03286*, 2025.

RGB Ground. Whu-mars: A multispectral aerial-ground benchmark towards any-scenario person re-identification supplementary material.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Hongyang Gu, Qisong Yang, Lei Pu, Siming Han, and Yao Ding. Reidmamba: Learning discriminative features with visual state space model for person re-identification. *IEEE Transactions on Multimedia*, 2026.

Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 16403–16412, 2021.

Zhangyi Hu, Bin Yang, and Mang Ye. Empowering visible-infrared person re-identification with large foundation models. *Advances in Neural Information Processing Systems*, 37:117363–117387, 2024.

Xun Jiang, Yufan Gu, Disen Hu, Yuqing Hou, Yazhou Yao, Fumin Shen, Heng Tao Shen, and Xing Xu. Multimodal learning on low-quality data with conformal predictive self-calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23041–23050, 2026.

Yan Jiang, Xu Cheng, Hao Yu, Xingyu Liu, Haoyu Chen, and Guoying Zhao. Domain shifting: A generalized solution for heterogeneous cross-modality person re-identification. In *European Conference on Computer Vision*, pp. 289–306. Springer, 2024.

Yan Jiang, Hao Yu, Xu Cheng, Haoyu Chen, Zhaodong Sun, and Guoying Zhao. From laboratory to real world: A new benchmark towards privacy-preserved visible-infrared person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8828–8837, 2025.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18621–18632, 2023.

Xulin Li, Yan Lu, Bin Liu, Qinhong Yang, Qi Chu, Tao Gong, and Nenghai Yu. Mfen: Multi-frequency expert network for visible-infrared person re-id. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18471–18480, 2026.

Tengfei Liang, Yi Jin, Wu Liu, and Yidong Li. Cross-modality transformer with modality mining for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 25:8432–8444, 2023.

Jianan Liu, Jialiang Wang, Nianchang Huang, Qiang Zhang, and Jungong Han. Revisiting modality-specific feature compensation for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7226–7240, 2022.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1835–1843, 2023.

Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13379–13389, 2020.

Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017.

Honghu Pan, Wenjie Pei, Xin Li, and Zhenyu He. Unified conditional image generation for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 19: 9026–9038, 2024.

Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12026–12035. IEEE, 2021.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

Kaijie Ren and Lei Zhang. Implicit discriminative knowledge learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 393–402, 2024.

Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.

Jake C Snell and Thomas L Griffiths. Conformal prediction as bayesian quadrature. *arXiv preprint arXiv:2502.13228*, 2025.

Dengwen Wang, Guanyu Xing, and Yanli Liu. Low-light invariant representation learning for visible-infrared person re-identification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8645–8653, 2025.

Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12144–12151, 2020.

Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3623–3632, 2019.

Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Flexible body partition-based adversarial learning for visible infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4676–4687, 2021a.

Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 225–234, 2021b.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pp. 5380–5389, 2017.

Haoxuan Xu and Guanglin Niu. Bit: Matching-based bi-directional interaction transformation network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40386–40396, 2026.

Yunpeng Xu, Mufang Ying, Wenge Guo, and Zhi Wei. Two-stage risk control with application to ranked retrieval. *arXiv preprint arXiv:2404.17769*, 2024.

Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14288–14297. IEEE, 2022.

Xi Yang, Wenjiao Dong, Meijie Li, Ziyu Wei, Nannan Wang, and Xinbo Gao. Cooperative separation of modality shared-specific features for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 26:8172–8183, 2024.

Mang Ye, Jianbing Shen, David J. Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European conference on computer vision*, pp. 229–247. Springer, 2020.

Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE Transactions on Information Forensics and Security*, 17:386–398, 2021a.

Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13567–13576, 2021b.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021c.

Mang Ye, Zesen Wu, and Bo Du. Dual-level matching with outlier filtering for unsupervised visible-infrared person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11185–11195, 2023.

Jian Yu, Yujian Feng, Shuai You, Zhongkai Zhou, Fei Wu, Zhengjun Jing, and Yimu Ji. Spatial-frequency collaborative learning for occluded visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4343–4352, 2026.

Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7349–7358, 2022a.

Yiyuan Zhang, Yuhao Kang, Sanyuan Zhao, and Jianbing Shen. Dual-semantic consistency learning for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 18:1554–1565, 2022b.

Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *European conference on computer vision*, pp. 462–479. Springer, 2022c.

Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2153–2162, 2023.

Yukang Zhang, Yan Yan, Yang Lu, and Hanzi Wang. Towards a unified middle modality learning for visible-infrared person re-identification. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 788–796, 2021.

Xiangtao Zheng, Xiumei Chen, and Xiaoqiang Lu. Visible-infrared person re-identification via partially interactive collaboration. *IEEE Transactions on Image Processing*, 31:6951–6963, 2022.

Xian Zhong, Tianyou Lu, Wenxin Huang, Mang Ye, Xuemei Jia, and Chia-Wen Lin. Grayscale enhancement colorization network for visible-infrared person re-identification. *IEEE transactions on circuits and systems for video technology*, 32(3):1418–1430, 2021.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A Appendix

A Theoretical Preliminaries

A.1 Review of the Problem Setup

Cross-modal retrieval setting. For a mini-batch $\mathcal{B} = \{(x_b, y_b, m_b)\}_{b=1}^{N}$, let $f_b \in \mathbb{R}^d$ be the identity descriptor and $S_{bk} = \langle \hat{f}_b, \hat{f}_k \rangle \in [-1, 1]$ be the cosine similarity between normalized descriptors. The valid cross-modal positive and negative sets of anchor b are

$$\mathcal{P}_b = \{k : y_k = y_b,\ m_k \neq m_b\}, \qquad \mathcal{N}_b = \{k : y_k \neq y_b,\ m_k \neq m_b\}. \tag{19}$$

Loss function. When $\mathcal{P}_b \neq \varnothing$, define $h_b = \min_{p \in \mathcal{P}_b} S_{bp}$ and

$$r_b = \frac{1}{|\mathcal{N}_b|} \sum_{k \in \mathcal{N}_b} \sigma\left(\frac{S_{bk} - h_b}{\tau}\right), \qquad \tau > 0. \tag{20}$$

Consistent with the implementation, an anchor without a cross-modal positive or negative is assigned $r_b = 0$. This batch statistic is a smooth surrogate for ordering violations; it is not identical to Rank-1 or mAP.

Reliability-guided feature path. USIM predicts $s_{bq} = \sigma(a_{bq}) \in (0, 1)$ and $u_{bq} = 1 - s_{bq}$ at spatial position q. The UGLE modulation and pooling weights are

$$m_{bq} = 1 + \rho(s_{bq} - \tfrac{1}{2}), \qquad w_{bq} = \max\left\{1 + \alpha\left(\frac{s_{bq}}{\bar{s}_b + \epsilon} - 1\right), \eta_0\right\}, \tag{21}$$

where $\epsilon > 0$ is the numerical stabilizer and $\eta_0 > 0$ is the code-level pooling floor.

Bayesian quadrature of quantile functions. For a bounded random loss with distribution function F, its quantile function $K(t) = F^{-1}(t)$ is nondecreasing and its expectation can be written as $\int_0^1 K(t)\,dt$. Bayesian quadrature places uncertainty over this integral; random quantile spacings motivate symmetric Dirichlet weights Snell & Griffiths (2025). USIL uses only this computational idea inside a training regularizer and does not construct the calibration posterior required by the original result.

Dirichlet-reweighted risk surrogate. For each Monte Carlo draw, the implementation samples $U^{(j)} \sim \mathrm{Dirichlet}(\mathbf{1}_{N+1})$ and forms a nonnegative convex combination of the ordered batch risks and the worst-case pseudo-risk $r_{\max}$:

$$Q^{(j)} = \sum_{b=1}^{N} U_b^{(j)} r_{(b)} + U_{N+1}^{(j)} r_{\max}, \qquad L_M^+ = \mathrm{Quantile}_\eta\{Q^{(1)}, \ldots, Q^{(M)}\}, \tag{22}$$

where $r_{(1)} \le \cdots \le r_{(N)}$. No spacing is reused, and the final spacing is reserved for $r_{\max}$. This is the sampled Dirichlet-weighted sum in Eq. equation 11, without treating the mini-batch as an independent calibration set.

Scope of the analysis. The risks in Eq. equation 20 share samples and are generally dependent, and the descriptor changes throughout training. Therefore, the exchangeable calibration assumptions required for a deployment-time conformal or posterior risk guarantee are not imposed here. The results below establish properties of the implemented optimization surrogate only.

A.2 Background

Proposition A.1 (Logistic and cosine properties). For $\sigma(z) = (1 + e^{-z})^{-1}$, $0 < \sigma(z) < 1$, $\sigma'(z) = \sigma(z)(1 - \sigma(z)) \in (0, 1/4]$, and σ is strictly increasing. If $\hat{f}_b$ and $\hat{f}_k$ have unit norm, then $\langle \hat{f}_b, \hat{f}_k \rangle \in [-1, 1]$.

Proof. The sigmoid claims follow by direct differentiation. The cosine bound follows from Cauchy–Schwarz. □

Proposition A.2 (Moments of symmetric Dirichlet weights). If $U \sim \mathrm{Dirichlet}(\mathbf{1}_K)$, then

$$\mathbb{E}[U_i] = \frac{1}{K}, \quad \mathrm{Var}(U_i) = \frac{K-1}{K^2(K+1)}, \quad \mathrm{Cov}(U_i, U_j) = -\frac{1}{K^2(K+1)} \quad (i \neq j). \tag{23}$$

Proof. Substitution of $\alpha_i = 1$ and $\alpha_0 = \sum_i \alpha_i = K$ into the standard Dirichlet moments gives the result. □

Proposition A.3 (Consistency of the sampled quantile). Let $Q^{(1)}, Q^{(2)}, \ldots$ be i.i.d. draws from a distribution whose CDF is continuous and strictly increasing at its η-quantile q_η. Then the empirical quantile in Eq. equation 22 satisfies $L_M^+ \to q_\eta$ almost surely as $M \to \infty$.

Proof. The empirical CDF converges uniformly almost surely to the population CDF by the Glivenko–Cantelli theorem. Continuity and strict increase at q_η then imply convergence of the corresponding empirical quantile. □

B Proofs of Results from the Main Paper

B.1 Proof of Proposition 1: Uncertainty–Reliability Complementarity

Proposition 1. For every finite reliability logit a_{bq}, $s_{bq}, u_{bq} \in (0, 1)$, $s_{bq} + u_{bq} = 1$, and $0 < \partial s_{bq}/\partial a_{bq} \le 1/4$.

Proof. The sigmoid maps every finite real logit strictly into $(0, 1)$, so its complement does as well and their sum is one. Direct differentiation gives $\partial s_{bq}/\partial a_{bq} = s_{bq}(1 - s_{bq}) \in (0, 1/4]$. □

B.2 Proof of Proposition 2: Stability of UGLE

Proposition 2. For $s_{bq} \in (0, 1)$,

$$1 - \frac{|\rho|}{2} < m_{bq} < 1 + \frac{|\rho|}{2}. \tag{24}$$

Hence $m_{bq} > 0$ whenever $|\rho| < 2$. Moreover, $w_{bq} \geq \eta_0 > 0$. If the pooled activation is nonnegative and $p > 0$, its weighted GeM value lies between the minimum and maximum spatial activation in each channel.

Proof. Because $-\frac{1}{2} < s_{bq} - \frac{1}{2} < \frac{1}{2}$, multiplication by ρ and the triangle inequality yield Eq. equation 24. Positivity follows from $|\rho| < 2$. The pooling-weight result follows directly from the maximum in Eq. equation 21. Normalized positive weights form a weighted mean of $(G_{bq})^p$ between its minimum and maximum; applying the increasing map $x \mapsto x^{1/p}$ proves the GeM range. □

B.3 Proof of Proposition 3 and Theorem 1: State Stability and Linear Spatial Scaling

Proposition 3. Suppose that one directional recurrence satisfies $\|\bar{A}_t\|_{\mathrm{op}} \leq a < 1$, $\|\bar{B}_t\|_{\mathrm{op}} \leq b$, and $\|Q_t\| \leq q$ for every t. Then

$$\|o_t\| \leq a^t \|o_0\| + \frac{bq(1 - a^t)}{1 - a}. \tag{25}$$

Proof. The recurrence and submultiplicativity give $\|o_t\| \leq a\|o_{t-1}\| + bq$. Repeated substitution yields a geometric series, proving Eq. equation 25. The argument applies separately to the forward and backward scans; the triangle inequality bounds their fusion. □

Proposition B.1. For fixed channel width, state dimension, expansion factor, and convolution kernel, one selective scan over L tokens has time complexity $\mathcal{O}(L)$ with respect to L.

Proof. The recurrent selective update performs a fixed-dimensional parameter generation and state update once per token. The fallback implementation uses fixed-width projections, a fixed-kernel depth-wise convolution, and a pointwise gate. Each case therefore incurs constant work per token when model dimensions are fixed. □

Proposition B.2. Sequence reversal, elementwise fusion of two scan outputs, LayerNorm, and the location-wise reliability head each require $\mathcal{O}(L)$ work.

Proof. Each operation visits every token a constant number of times and performs fixed-dimensional work. □

Theorem 1. For fixed model dimensions, bidirectional USIM has $\mathcal{O}(L) = \mathcal{O}(HW)$ time and activation-memory complexity with respect to the number of spatial tokens.

Proof. USIM contains two selective scans plus the operations in Proposition B.2. Adding a constant number of $\mathcal{O}(L)$ terms remains $\mathcal{O}(L)$. The materialized sequence and its two scan outputs store a constant number of fixed-width vectors per token, yielding $\mathcal{O}(L)$ activation memory. □

B.4 Proof of Lemma 1: Soft Cross-modal Ranking Risk

Lemma 1. For a valid anchor with nonempty $\mathcal{P}_b$ and $\mathcal{N}_b$, $0 < r_b < 1$. At a point where the hardest positive is unique, r_b is strictly increasing in every S_{bk}, $k \in \mathcal{N}_b$, and strictly decreasing in h_b, with

$$0 < \frac{\partial r_b}{\partial S_{bk}} \leq \frac{1}{4\tau|\mathcal{N}_b|}, \qquad -\frac{1}{4\tau} \leq \frac{\partial r_b}{\partial h_b} < 0. \tag{26}$$

If $H_b = |\mathcal{N}_b|^{-1} \sum_{k \in \mathcal{N}_b} \mathbb{1}\{S_{bk} > h_b\}$ is the hard violation fraction, then $H_b \leq 2r_b$. In the absence of ties, $r_b \to H_b$ as $\tau \to 0^+$.

Proof. Every summand in Eq. equation 20 lies strictly between zero and one, so does their average. Differentiating gives

$$\frac{\partial r_b}{\partial S_{bk}} = \frac{\sigma'(z_{bk})}{\tau|\mathcal{N}_b|}, \qquad \frac{\partial r_b}{\partial h_b} = -\frac{1}{\tau|\mathcal{N}_b|} \sum_{k \in \mathcal{N}_b} \sigma'(z_{bk}),$$

where $z_{bk} = (S_{bk} - h_b)/\tau$. Proposition A.1 gives the signs and bounds. The minimum is differentiable when its minimizer is unique and is subdifferentiable at ties. Every hard violation contributes at least $1/2$ to the corresponding sigmoid term, hence $H_b \leq 2r_b$. Away from ties, sigmoid converges pointwise to the step function as $\tau \to 0^+$. □

B.5 Proof of Theorem 2: Dirichlet Surrogate Properties

Theorem 2. Conditioned on a fixed batch, let $a = (r_{(1)}, \ldots, r_{(N)}, r_{\max})$, $K = N + 1$, and $Q = U^\top a$ with $U \sim \mathrm{Dirichlet}(\mathbf{1}_K)$. Then

$$0 \leq Q \leq \|a\|_\infty \leq 1, \qquad \mathbb{E}[Q] = \frac{1}{K}\sum_{i=1}^{K} a_i, \tag{27}$$

and

$$\mathrm{Var}(Q) = \frac{K\sum_i a_i^2 - (\sum_i a_i)^2}{K^2(K+1)}. \tag{28}$$

Under the regularity condition of Proposition A.3, L_M^+ converges almost surely to the η-quantile of this conditional surrogate distribution.

Proof. Dirichlet weights are nonnegative and sum to one, hence Q is a convex combination of the entries of a, proving its range. Linearity of expectation and Proposition A.2 give Eq. equation 27. Expanding $\mathrm{Var}(\sum_i U_i a_i)$ and substituting the variances and covariances in Eq. equation 23 gives Eq. equation 28. The final claim is Proposition A.3. □

The theorem characterizes the random regularizer actually optimized by DCRR. It does not imply that L_M^+ upper-bounds unseen retrieval loss, nor does it provide a conformal, posterior, Rank-1, or mAP guarantee.

B.6 Proof of Corollary 1: Uncertainty-to-Risk Feedback

Corollary 1. Assume nonzero corresponding loss weights and differentiable non-saturated operations. The USIL objective contains (i) a path from observed ordering risk to the risk head through $\mathcal{L}_{\mathrm{risk}}$, (ii) a path from predicted and correctness-anchored risk to the USIM reliability through $\mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}}$, (iii) a path from detached local cross-modal identity margins to stripe reliability, and (iv) a path from L_M^+ through r_b to the identity descriptor.

Proof. In $\mathcal{L}_{\mathrm{risk}} = \mathrm{MSE}(\hat{r}, \mathrm{sg}[r])$, stop-gradient makes r a target while gradients flow through $\hat{r} = \sigma(\ell)$ and the risk head. In

$$\mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}} = \mathrm{MSE}(\bar{s}, \mathrm{sg}[t]) + \mathrm{MSE}(\hat{r}, 1 - \bar{s}),$$

$\bar{s}$ remains live in both terms. The local stripe reliability is instead anchored by a stripe-wise mean-centered ranking alignment, $\mathcal{L}_{\mathrm{rel}}^{\mathrm{loc}} = \mathrm{mean}_{b\in\mathcal{V}}(1 - \cos(\tilde{\mathbf{s}}_b^{\mathrm{stripe}}, \mathrm{sg}[\tilde{t}_b^{\mathrm{loc}}]))$, whose gradient flows into the reliability head through $\mathbf{s}^{\mathrm{stripe}}$ (an absolute-value MSE would admit a spatially constant minimizer and vanish). Finally, the implementation passes the non-detached r to L_M^+, so its gradient reaches pairwise similarities and the descriptors. □

This corollary establishes graph connectivity only; it does not assert a strictly nonzero gradient at every parameter value.

C Additional Optimization Details

The invariant backbone, USIM, UGLE, and the risk-prediction head are optimized jointly by Eq. equation C.8. The stop-gradient operators apply only to the observed targets in $\mathcal{L}_{\mathrm{risk}}$, $\mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}}$, and $\mathcal{L}_{\mathrm{rel}}^{\mathrm{loc}}$; the Dirichlet-reweighted L_M^+ remains differentiable with respect to the ordered batch risks and therefore regularizes the cross-modal similarities and identity descriptors.

D ALGORITHM OF USIL

Explanation of Algorithm 1. Algorithm 1 summarizes USIL as one continuous identity-cue-uncertainty-to-retrieval-risk learning flow. For each visible–infrared mini-batch, the invariant backbone first produces the third-stage pedestrian feature $\mathbf{F}$. The implementation uses separate feature-path and risk coefficients,

$$\nu_e = \mathrm{clip}_{[0,1]}\left(\frac{e-10}{20}\right), \qquad \omega_e = \mathrm{clip}_{[0,1]}\left(\frac{e-e_0}{e_1}\right), \tag{C.1}$$

which are deterministic optimization schedules rather than learnable parameters or assumptions of the theoretical results. USIM (lines 2–3) unfolds $\mathbf{F}$ into a spatial sequence, performs forward and reverse selective state propagation, and decodes both the refined feature and the shared reliability state:

$$\text{USIM}: \qquad \mathbf{F} \longrightarrow \mathbf{Q} \longrightarrow \mathbf{Y} \longrightarrow (\mathbf{F}^u, \mathbf{s}, \mathbf{u}), \qquad \mathbf{F}^u = \mathbf{F} + \nu_e \gamma\, \mathrm{Proj}_{\mathrm{out}}(\mathbf{Y}), \qquad \mathbf{u} = \mathbf{1} - \mathbf{s}. \tag{C.2}$$

Here, $\mathbf{s}$ measures the spatial support assigned to identity-consistent evidence, whereas $\mathbf{u}$ records its complementary identity-cue uncertainty. The two quantities are generated by one state rather than by independent attention branches. The released implementation uses Mamba when `mamba_ssm` is available and otherwise executes its gated depth-wise-convolution fallback; claims about selective SSM recurrence apply to the former path.

UGLE (lines 4–8) reuses $\mathbf{s}$ for both low-light compensation and descriptor aggregation:

$$\begin{aligned}
\mathbf{m} &= 1 + \nu_e \rho(\mathbf{s} - \tfrac{1}{2}),\\
\mathbf{A}^1 &= \mathrm{LayerATT}_0(\mathbf{F}^u, \mathbf{F}^{(0)}) + \delta \odot \mathbf{m} \odot \mathrm{NAF}_1(\mathbf{F}^u),\\
\mathbf{A}^2 &= \mathrm{LayerATT}_1(\mathbf{A}^1, \mathbf{F}^{(1)}) + 0.3\,\mathbf{m} \odot \mathrm{NAF}_2(\mathbf{F}^u),\\
\mathbf{A}^3 &= \mathrm{LayerATT}_2(\mathbf{A}^2, \mathbf{F}^{(2)}),\\
\mathbf{G} &= \mathrm{ResLayer}_4\big(\mathrm{LayerATT}_3(\mathbf{A}^3, \mathbf{F}^u)\big).
\end{aligned} \tag{C.3}$$

The first NAF branch uses the learnable channel-wise scale δ, whereas the second uses the fixed IRL coefficient 0.3. Both NAF branches read the same $\mathbf{F}^u$; the second does not consume $\mathbf{A}^1$. Reliability is bilinearly resized to $\mathbf{G}$ before weighted GeM, with

$$w_{bq} = \max\left\{1 + \nu_e \alpha \left(\frac{s_{bq}}{\bar{s}_b + 10^{-6}} - 1\right), 0.1\right\}, \quad f_{bc} = \left(\frac{\sum_q w_{bq} G^3_{bcq}}{\sum_q w_{bq}} + 10^{-12}\right)^{1/3}. \tag{C.4}$$

The descriptor then enters parallel identity and DCRR branches (lines 9–10):

$$\begin{aligned}
\mathbf{f}_b &\longrightarrow \mathrm{BN} \longrightarrow \mathbf{z}_b,\\
\mathbf{f}_b &\longrightarrow \psi \longrightarrow \ell_b \longrightarrow \widehat{r}_b,\\
\{\mathbf{f}_b\}_{b=1}^N &\longrightarrow \mathbf{S} \longrightarrow r_b.
\end{aligned} \tag{C.5}$$

Thus, the risk-prediction head does not consume the observed ranking risk. Instead, $\widehat{r}_b$ is predicted directly from $\mathbf{f}_b$, while r_b is independently computed from cross-modal positive–negative ordering violations. The two branches meet only through the calibration and reliability objectives.

Line 11 sorts the observed batch risks and applies one distinct Dirichlet spacing to each ordered risk, reserving the final spacing for $r_{\max}$:

$$\text{DCRR}: \qquad r \longrightarrow r_{(1:N)}, \quad U^{(j)} \sim \mathrm{Dirichlet}(\mathbf{1}_{N+1}) \longrightarrow Q^{(j)} \longrightarrow L^+. \tag{C.6}$$

This construction is a BQ-inspired stochastic upper-tail training surrogate. Because mini-batch ranking risks share descriptors and negatives, it is not a held-out calibration posterior and does not provide a deployment-time conformal guarantee.

Lines 12–15 combine the USIL constraints:

$$\begin{aligned}
\mathcal{L}_{\mathrm{het}} &\leftarrow (\mathbf{z}_b, \ell_b),\\
\mathcal{L}_{\mathrm{risk}} &\leftarrow (\widehat{r}_b, \mathrm{sg}[r_b]),\\
\mathcal{L}^{\mathrm{samp}}_{\mathrm{rel}} &\leftarrow (\bar{s}_b, \widehat{r}_b, \mathrm{sg}[t_b]),\\
\mathcal{L}^{\mathrm{loc}}_{\mathrm{rel}} &\leftarrow (\tilde{\mathbf{s}}^{\mathrm{stripe}}_b, \mathrm{sg}[\tilde{t}^{\mathrm{loc}}_b]),\\
L^+ &\leftarrow r_{(1:N)}.
\end{aligned} \tag{C.7}$$

Dataset	Training identities/images	Testing identities/images	Query/gallery construction	Aggregation
SYSU-MM01	395; 22,258 VIS / 11,909 IR	96; 6,775 VIS / 3,803 IR	3,803 IR queries; single-shot VIS gallery	10 gallery trials
RegDB	206; 2,060 VIS / 2,060 thermal per split	206; 2,060 VIS / 2,060 thermal per split	VIS↔thermal	10 official splits
LLCM	713; 16,946 VIS / 13,975 NIR	351; 8,680 VIS / 7,166 NIR	VIS↔NIR; single-shot gallery	10 gallery trials

Table 4: Verified statistics and evaluation construction of the three VI-ReID benchmarks.

The main paper groups these two reliability constraints as $\mathcal{L}_{\mathrm{rel}} = \mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}} + \kappa\mathcal{L}_{\mathrm{rel}}^{\mathrm{loc}}$. In particular, $\mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}}$ is not generated by the Dirichlet branch. Its correctness-aware target $t_b = \mathbb{1}[\widehat{y}_b = y_b](1 - r_b)$ and the consistency term $\mathrm{MSE}(\widehat{r}, 1 - \bar{s})$ feed retrieval difficulty back to the USIM reliability head. The stripe reliability is instead supervised by the ranking-alignment term $\mathcal{L}_{\mathrm{rel}}^{\mathrm{loc}}$ built on the detached six-stripe identity-margin target of Eq. equation 13: because the absolute reliability level is uninformative, an MSE against t^{loc} is minimized by a spatially constant map, so we mean-center both sequences across stripes and align their relative ordering, which cannot collapse to a constant. This term is scheduled by the module warm-up ν_e (active once USIM wakes), not by the DCRR risk ramp ω_e, so spatial reliability is shaped throughout training rather than only after the ramp opens; it does not assume pixel-level registration (only coarse-stripe semantic correspondence). The stop-gradient operator is applied to r_b only when it acts as a target in $\mathcal{L}_{\mathrm{risk}}$ or through t_b; the copy entering L^+ remains differentiable and regularizes pairwise similarities and descriptors. With the DCRR switch $I_{\mathrm{DCRR}} \in \{0, 1\}$, the exact implemented objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{task}} &+ I_{\mathrm{DCRR}}\omega_e \left(\lambda_{\mathrm{het}}\mathcal{L}_{\mathrm{het}} + \lambda_{\mathrm{risk}}\mathcal{L}_{\mathrm{risk}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}^{\mathrm{samp}} + \lambda^{+}L^{+}\right) \\ &+ I_{\mathrm{DCRR}}\nu_e\lambda_{\mathrm{rel}}^{\mathrm{loc}}\mathcal{L}_{\mathrm{rel}}^{\mathrm{loc}}.\end{aligned} \tag{C.8}$$

Here $\mathcal{L}_{\mathrm{task}} = \mathcal{L}_{\mathrm{id}} + \mathcal{L}_{\mathrm{tri}}$. Equation equation 18 in the main paper gives the concise full-scale grouped objective. Equation equation C.8 records the code-exact schedules; when $\omega_e = \nu_e = 1$, the two forms agree with $\kappa = \lambda_{\mathrm{rel}}^{\mathrm{loc}}/\lambda_{\mathrm{rel}}$.

Fixed quantities and optimized parameters. Identity labels, modality indicators, the induced positive/negative masks, the ranking temperatures τ and τ_{loc}, the GeM exponent p, the Dirichlet sample count M, the quantile level η, $r_{\max}$, the schedule values, and the loss coefficients are fixed during an optimization step. The invariant encoder, USIM projections and selective mixers, reliability head, UGLE modulation and pooling scales, identity classifier, and risk-prediction head are jointly optimized by Eq. equation C.8. The registered `sin` parameter and `NAF3` module are not used by the released forward graph and are therefore excluded from this optimized set.

Physical meaning of the loss decomposition. $\mathcal{L}_{\mathrm{het}}$ changes the contribution of each sample to identity classification according to its predicted log-risk, while retaining the ground-truth identity label. $\mathcal{L}_{\mathrm{risk}}$ fits the risk head to the stop-gradient observed ranking risk. The hierarchical $\mathcal{L}_{\mathrm{rel}}$ combines sample-level correctness–risk feedback with location-wise identity-margin ordering, so the local term is an internal reliability-learning process rather than a fifth top-level objective. L^+ changes the identity descriptor through differentiable cross-modal ranking violations and emphasizes their high-risk tail.

Fixed targets and optimized quantities. Identity labels and positive/negative membership are fixed by the mini-batch. The observed risk is detached only when serving as the target of $\mathcal{L}_{\mathrm{risk}}$ or $\mathcal{L}_{\mathrm{rel}}$, and the local stripe-margin target is also detached. In contrast, L^+ retains gradients through the observed ranking risks. Consequently, optimization changes the encoder, identity descriptor, USIM reliability head, enhancement path, classifier, and risk head, but it does not change labels, pair membership, detached targets, or sampled Dirichlet spacings.

E Additional Experimental and Implementation Details

E.1 Additional Dataset and Evaluation Details

We evaluate USIL on SYSU-MM01, RegDB, and LLCM. Their identity partitions are summarized in Tab. 4. These values follow the official splits used by the data loaders and evaluation scripts.

Algorithm 1 USIL: Uncertainty-Aware Selective Invariant Learning for Low-Light VI-ReID

Input: Visible–infrared mini-batch $\mathcal{B}=\{(x_b,y_b,m_b)\}_{b=1}^{N}$ with modality $m_b\in\{v,i\}$; deep pedestrian feature $\mathbf{F}\in\mathbb{R}^{N\times C\times H\times W}$ (IRL layer-3); low-light compensation branch $\mathrm{NAF}(\cdot)$; identity classifier ϕ; risk head ψ; ranking temperature τ; Dirichlet draws M and upper-quantile level η; local-margin temperature τ_{loc}; USIL loss weights $(\lambda_{\mathrm{het}},\lambda_{\mathrm{risk}},\lambda_{\mathrm{rel}},\lambda^{+})$; local reliability coefficient κ; feature-path coefficient $\nu_e\in[0,1]$.
Output: training objective $\mathcal{L}_{\mathrm{total}}$.
Note: identity-cue uncertainty $\mathbf{u}=\mathbf{1}-\mathbf{s}$; the ranking risk r is stop-gradient only when used as a prediction or reliability target, while its gradient is retained in L^{+}. The algorithm displays the full-scale objective; implementation schedules are given in Appendix C.

for each epoch $e=1,\dots,E$ and each mini-batch $\mathcal{B}$ **do**
// USIM: uncertainty-aware selective invariant state modeling (innov-1)
$\mathbf{q}\leftarrow \mathrm{Proj}_{\mathrm{in}}\big(\mathrm{LN}(\mathrm{flatten}(\mathbf{F}))\big)$; $\mathbf{y}\leftarrow \mathrm{LN}\big(\mathrm{SSM}_{\rightarrow}(\mathbf{q})+\mathrm{flip}(\mathrm{SSM}_{\leftarrow}(\mathrm{flip}(\mathbf{q})))\big)$
$\mathbf{F}^{u}\leftarrow \mathbf{F}+\nu_e\gamma\,\mathrm{Proj}_{\mathrm{out}}(\mathbf{y})$; $\mathbf{s}\leftarrow\sigma(\mathrm{Head}(\mathbf{y}))\in(0,1)^{H\times W}$, $\mathbf{u}\leftarrow\mathbf{1}-\mathbf{s}$ ▷ state refinement + spatial reliability/uncertainty
// UGLE: uncertainty-guided low-light enhancement and aggregation (innov-2)
$\mathbf{m}\leftarrow 1+\nu_e\rho(\mathbf{s}-\frac{1}{2})$
$\mathbf{A}^{1}\leftarrow \mathrm{LayerATT}_0(\mathbf{F}^{u},\mathbf{F}^{(0)})+\delta\odot\mathbf{m}\odot\mathrm{NAF}_1(\mathbf{F}^{u})$
$\mathbf{A}^{2}\leftarrow \mathrm{LayerATT}_1(\mathbf{A}^{1},\mathbf{F}^{(1)})+0.3\,\mathbf{m}\odot\mathrm{NAF}_2(\mathbf{F}^{u})$
$\mathbf{A}^{3}\leftarrow \mathrm{LayerATT}_2(\mathbf{A}^{2},\mathbf{F}^{(2)})$; $\mathbf{G}\leftarrow\mathrm{ResLayer}_4(\mathrm{LayerATT}_3(\mathbf{A}^{3},\mathbf{F}^{u}))$ ▷ implementation-faithful multi-stage enhancement
$w\leftarrow\big(1+\nu_e\alpha(\mathbf{s}/(\bar{\mathbf{s}}+10^{-6})-1)\big)_{\geq 0.1}$; $\mathbf{f}\leftarrow\big(\frac{\sum w\odot\mathbf{G}^{p}}{\sum w}+10^{-12}\big)^{1/p}$ ▷ reliability-weighted GeM
$\mathbf{z}\leftarrow\phi(\mathrm{BN}(\mathbf{f}))$; $\ell_b\leftarrow 4\tanh(\psi(\mathbf{f}_b)/4)$, $\hat{r}_b\leftarrow\sigma(\ell_b)$ ▷ identity logits; predicted per-sample risk
// DCRR: distributional cross-modal retrieval-risk regularization (innov-3)
$r_b\leftarrow\frac{1}{|\mathcal{N}_b|}\sum_{k\in\mathcal{N}_b}\sigma\Big(\frac{1}{\tau}\big(S_{bk}-\min_{p\in\mathcal{P}_b}S_{bp}\big)\Big)$ ▷ $\mathcal{P}_b,\mathcal{N}_b$: cross-modal positives/negatives
$h_{bk}\leftarrow\mathrm{sg}[\mathrm{StripePool}_k(\mathbf{F}_b)]$; $t^{\mathrm{loc}}_{bk}\leftarrow\sigma\Big(\frac{\min_{p\in\mathcal{P}_b}\langle\hat{h}_{bk},\hat{h}_{pk}\rangle-\max_{n\in\mathcal{N}_b}\langle\hat{h}_{bk},\hat{h}_{nk}\rangle}{\tau_{\mathrm{loc}}}\Big)$ ▷ spatial reliability target (no pixel-level registration)
$\mathbf{U}\sim\mathrm{Dirichlet}(\mathbf{1}_{N+1})^{\otimes M}$; $r_{(1:N)}\leftarrow\mathrm{sort}(r)$; $L^{+}\leftarrow\mathrm{Quantile}_{\eta}\big(\mathbf{U}_{:,1:N}\,r_{(1:N)}+\mathbf{U}_{:,N+1}\,r_{\max}\big)$ ▷ BQ-inspired upper-tail risk surrogate
$\mathcal{L}_{\mathrm{task}}\leftarrow\mathcal{L}_{\mathrm{id}}+\mathcal{L}_{\mathrm{tri}}$; $\mathcal{L}_{\mathrm{het}}\leftarrow\mathrm{mean}_b\big(\frac{1}{2}e^{-\ell_b}\mathrm{CE}(\mathbf{z}_b,y_b)+\frac{1}{2}\,\mathrm{softplus}(\ell_b)\big)$
$t_b\leftarrow\mathbb{1}[\hat{y}_b=y_b](1-r_b)$; $\mathcal{L}_{\mathrm{risk}}\leftarrow\mathrm{MSE}(\hat{r},\mathrm{sg}[r])$; $\mathcal{L}^{\mathrm{samp}}_{\mathrm{rel}}\leftarrow\mathrm{MSE}(\bar{s},\mathrm{sg}[t])+\mathrm{MSE}(\hat{r},1-\bar{s})$ ▷ sample-level risk prediction and reliability feedback
$\mathcal{L}^{\mathrm{loc}}_{\mathrm{rel}}\leftarrow\mathrm{mean}_{b\in\mathcal{V}}\big(1-\cos(\tilde{\mathbf{s}}^{\mathrm{stripe}}_{b},\mathrm{sg}[\tilde{t}^{\mathrm{loc}}_{b}])\big)$ ▷ stripe-wise mean-centered ranking alignment (no constant-map cheat)
$\mathcal{L}_{\mathrm{rel}}\leftarrow\mathcal{L}^{\mathrm{samp}}_{\mathrm{rel}}+\kappa\mathcal{L}^{\mathrm{loc}}_{\mathrm{rel}}$ ▷ hierarchical reliability learning
$\mathcal{L}_{\mathrm{USIL}}\leftarrow\lambda_{\mathrm{het}}\mathcal{L}_{\mathrm{het}}+\lambda_{\mathrm{risk}}\mathcal{L}_{\mathrm{risk}}+\lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}}+\lambda^{+}L^{+}$
$\mathcal{L}_{\mathrm{total}}\leftarrow\mathcal{L}_{\mathrm{task}}+\mathcal{L}_{\mathrm{USIL}}$
update {encoder, LPA/LayerATT, $\mathrm{NAF}_{1:2}$, USIM, UGLE, BN, classifier, ψ} with $\nabla\mathcal{L}_{\mathrm{total}}$
end for
return $\mathcal{L}_{\mathrm{total}}$

For SYSU-MM01, we use the official all-search and indoor-search protocols. Infrared images serve as queries and the visible gallery is sampled according to the single-shot protocol; reported results are averaged over ten random gallery trials. RegDB is evaluated in both visible-to-infrared and infrared-to-visible directions over its ten official train/test splits. LLCM uses its official bidirectional visible-to-infrared and infrared-to-visible protocols, and each reported direction is averaged over ten gallery trials. We report CMC accuracy and mAP, using Rank-1 and Rank-10 where specified in the main paper.

Each SYSU-MM01 all-search gallery trial contains 301 visible images from 96 identities; the indoor-search gallery contains 112 images from 56 identities. For LLCM, the infrared-to-visible direction uses 7,166 NIR queries and a 484-image visible gallery per trial, while the reverse direction uses 8,680 visible queries and a 619-image NIR gallery. The official evaluators apply their prescribed camera exclusions.

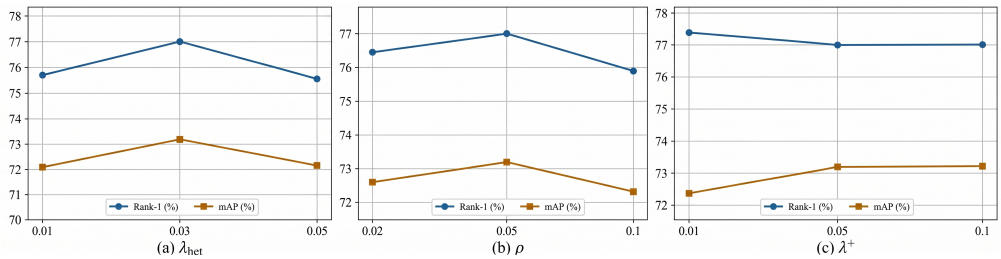


Figure 6: Parameter sensitivity on SYSU-MM01 (All-Search Single-Shot): (a) λ_{het}, (b) UGLE modulation scale ρ, and (c) λ^{+}.

E.2 Dataset-Specific Hyperparameter Settings

The common architecture uses a ResNet-50 AGW backbone, input resolution 384×192, six identities per mini-batch, and four visible plus four infrared images per identity. USIM uses state dimension 16, local convolution width 4, and expansion ratio 2. UGLE uses $p = 3$, while DCRR uses $\tau = \tau_{\text{loc}} = 0.1$, $M = 64$, $\eta = 0.9$, and $r_{\max} = 1$. All models are optimized for 150 epochs using SGD with initial learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, Nesterov acceleration, mixed precision, and triplet margin 0.7. The backbone parameter group uses one tenth of the nominal learning rate, while the BN bottleneck and classifier use the nominal rate. Each sampled source image produces two augmented views, so the four modal-view streams contain 96 images in one forward pass. The implementation uses PyTorch 2.1.1 and `mamba_ssm` 2.2.4.

The dataset-specific loss and modulation settings recovered from execution logs are listed in Tab. 5. For reproducibility, the feature-path interpolation (warm-up ν_e) reaches full scale from epoch 10 over 20 epochs; the local reliability term $\lambda_{\text{rel}}^{\text{loc}}\mathcal{L}_{\text{rel}}^{\text{loc}}$ is scheduled by this same ν_e. The other risk terms use a linear coefficient (ω_e) that starts at e_0 and reaches one after a further e_1 epochs; this optimization schedule is not an assumption of Propositions 1–3. The dashes for RegDB/LLCM indicate pre-fix logged runs that predate the local reliability supervision; the unified corrected retraining adds $\lambda_{\text{rel}}^{\text{loc}} = 0.5$ for all datasets.

Table 5: Recorded dataset-specific USIL configurations. RegDB and LLCM rows describe pre-fix executions and do not validate post-fix DCRR results.

Dataset	λ_{het}	λ_{risk}	λ_{rel}	$\lambda_{\text{rel}}^{\text{loc}}$	λ^{+}	ρ_{init}	e_0	e_1
SYSU-MM01 (corrected run)	0.03	0.01	0.01	0.5	0.05	0.05	35	15
RegDB (recorded pre-fix)	0.10	0.10	0.10	–	0.20	0.20	15	15
LLCM (recorded pre-fix)	0.10	0.10	0.10	–	0.30	0.20	15	15

E.3 Parameter Sensitivity Analysis

We further analyze the sensitivity of USIL to λ_{het}, ρ, and λ^{+} on SYSU-MM01 under the All-Search Single-Shot setting. Each parameter is varied independently while the others are fixed to the adopted settings, using the same 150-epoch training and ten-trial evaluation protocol. As shown in Fig. 6, Rank-1 and mAP remain relatively stable within the evaluated ranges.

E.4 Selective-State Reliability Diagnostic

We distinguish functional dependence from a monotonic coefficient interpretation. Let the bidirectional selective-state output be

$$\mathbf{Y} = \mathcal{S}(\mathbf{X}; \bar{\mathbf{A}}, \bar{\mathbf{B}}, \mathbf{C}, \mathbf{\Delta}), \qquad s_{bq} = \sigma(g(\mathbf{Y}_{bq})), \qquad u_{bq} = 1 - s_{bq}.$$

Thus, location-wise reliability and uncertainty depend on the input-dependent state propagation through $\mathbf{Y}$. For any scalar summary ξ of $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, $\mathbf{C}$, or $\mathbf{\Delta}$, the chain rule gives

$$\frac{\partial u_{bq}}{\partial \xi} = -\sigma'(g(\mathbf{Y}_{bq}))\, \mathbf{J}_g(\mathbf{Y}_{bq}) \frac{\partial \mathbf{Y}_{bq}}{\partial \xi}.$$

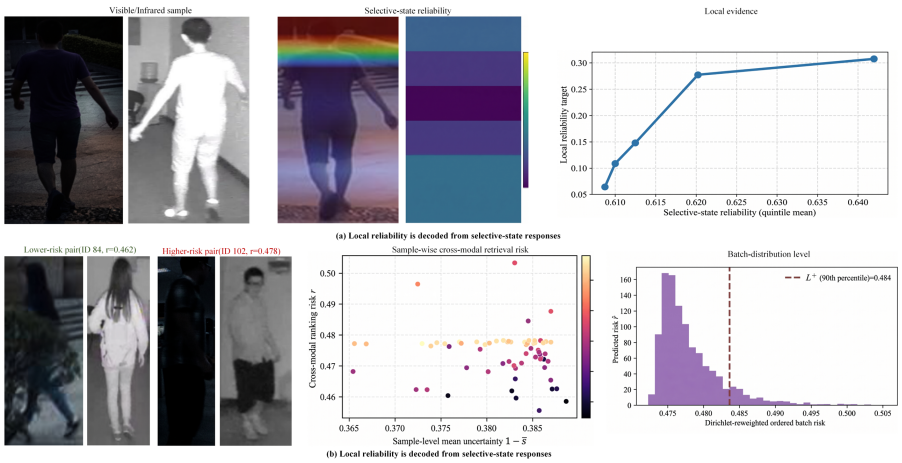


Figure 7: Real-data diagnostic of the complementary reliability–risk scales on SYSU-MM01. (a)At the local scale, reliability decoded from bidirectional selective-state representations follows the detached six-stripe identity-margin target ($\rho_s = 0.703$; top–bottom gap $= 0.228$). (b)At the sample scale, ranking violations define observed cross-modal retrieval risk and the learned risk head tracks it ($\rho_s = 0.769$). At the batch-distribution scale, Dirichlet quantile spacings reweight ordered risks to form L^+. Dirichlet spacings are not per-image uncertainty estimates, and the local result does not assign a monotonic uncertainty meaning to an individual raw Mamba coefficient.

The sign is not fixed because both the state sensitivity and learned reliability-head Jacobian may contain positive and negative components. Consequently, USIL does not interpret a larger raw Mamba coefficient as necessarily higher or lower uncertainty.

The empirical diagnostic instead evaluates whether the decoded six-stripe reliability follows the detached local identity-margin target in Eq. equation 13. On held-out SYSU-MM01 identities, we report Spearman's rank correlation between s^{stripe} and t^{loc}, the difference in mean target between the top and bottom reliability quantiles, and the spatial standard deviation of the predicted reliability. These quantities are computed from a fixed epoch-140 checkpoint without test-time optimization. Over 64 real visible–infrared identity pairs, the decoded reliability obtains a Spearman correlation of 0.703 with the local identity-margin target; the difference between the top and bottom 30% reliability groups is 0.228. The location-wise reliability range and standard deviation are 0.053 and 0.013, respectively, excluding a spatially constant solution. These results support the association between the *decoded selective-state reliability* and local identity consistency. They do not establish a monotonic interpretation for any individual raw state-space coefficient.

E.5 Reproducibility and Result-Aggregation Protocol

For SYSU-MM01 and LLCM, a checkpoint is evaluated on ten independently sampled gallery splits and the arithmetic mean of each metric is reported. For RegDB, training and evaluation are repeated on each of the ten official splits before averaging the metrics. All comparisons between a USIL configuration and the reproduced IRL baseline use the same image resolution, sampler, optimizer, backbone, retrieval direction, and trial aggregation. Parameter sensitivity experiments additionally use the same 150-epoch budget and strict ten-trial SYSU-MM01 evaluation for every point. The adopted midpoint $(\lambda_{\text{het}}, \rho, \lambda^+) = (0.03, 0.05, 0.05)$ reuses the reported main model, and the λ_{het} endpoints use the existing 150-epoch retrains.