BEYOND RANKED GALLERY: CONFORMAL IDENTITY SET FOR TEXT-BASED PERSON RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-based person retrieval enables users to search for individuals in an image gallery using natural-language descriptions. Existing methods primarily focus on cross-modal matching and ranking accuracy, whereas practical retrieval also requires deciding how many candidate identities to return. A fixed cutoff may exclude the target for ambiguous descriptions or return unnecessary candidates for distinctive ones. Conformal prediction offers a principled framework for constructing such sets by calibrating candidate-level scores. A natural approach is to derive these scores from softmax-normalized retrieval similarities. However, this normalization introduces a temperature dependence that can alter set sizes and coverage evenness without changing the retrieval ranking. To understand this dependence, we analyze the low-temperature behavior of representative softmax-based conformal scores. We prove that, after order-preserving transformations, these scores converge to a common similarity-margin score as temperature approaches zero. Motivated by this shared limit, we introduce a temperature-free method for constructing **Conformal Identity Sets** through direct margin calibration. Specifically, our method retains identities within a common calibrated similarity gap of the top-ranked identity, allowing set sizes to adapt to each query's similarity profile. The resulting sets provide finite-sample marginal coverage at a user-specified level under exchangeability, without softmax, temperature tuning, or retraining. Experiments with five frozen retrievers across three benchmarks demonstrate empirical coverage close to the target and a favorable trade-off between set size and coverage evenness.

1 INTRODUCTION

Text-based person retrieval (TBPR) ranks gallery images according to their relevance to a natural-language description (Li et al., 2017; Jiang & Ye, 2023; Qin et al., 2024). Advances in retrieval improve the ranking, but leave an important practical decision unresolved: how many distinct candidate identities should be returned? A fixed cutoff may omit the target for an ambiguous description or return unnecessary candidates for a distinctive one. As illustrated in Figure 1, a useful shortlist should adapt to query ambiguity while remaining manageable for a human operator to inspect. Conventional metrics such as Rank-1, Rank-5, and Rank-10 assess accuracy at fixed cutoffs (Yan et al., 2023), but provide no principled rule for this query-specific decision.

Conformal prediction provides a principled starting point for this decision. It calibrates a fixed conformal score on held-out data to construct prediction sets with finite-sample marginal coverage under exchangeabilitys (Vovk et al., 2005; Angelopoulos & Bates, 2021). Applied at the identity level, it can turn a retrieval ranking into a set of distinct candidates calibrated for a user-specified coverage level. Standard score choices include probability thresholding scores (Sadinle et al., 2019; Intrator et al., 2024) and adaptive scores (Romano et al., 2020; Angelopoulos et al., 2021; Huang et al., 2024), which use cumulative softmax probabilities, respectively. However, marginal coverage alone ensures neither compact sets nor even coverage across queries; these properties depend on the score used for calibration (Aleksandrova & Chertov, 2021; Akhavan Attar et al., 2026).

Despite their common coverage objective, these softmax-based constructions introduce a temperature dependence that can change candidate allocation. Our analysis reveals two distinct effects. (1) Temperature controls a query-wide adjustment to the probability thresholding score, with a larger

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.

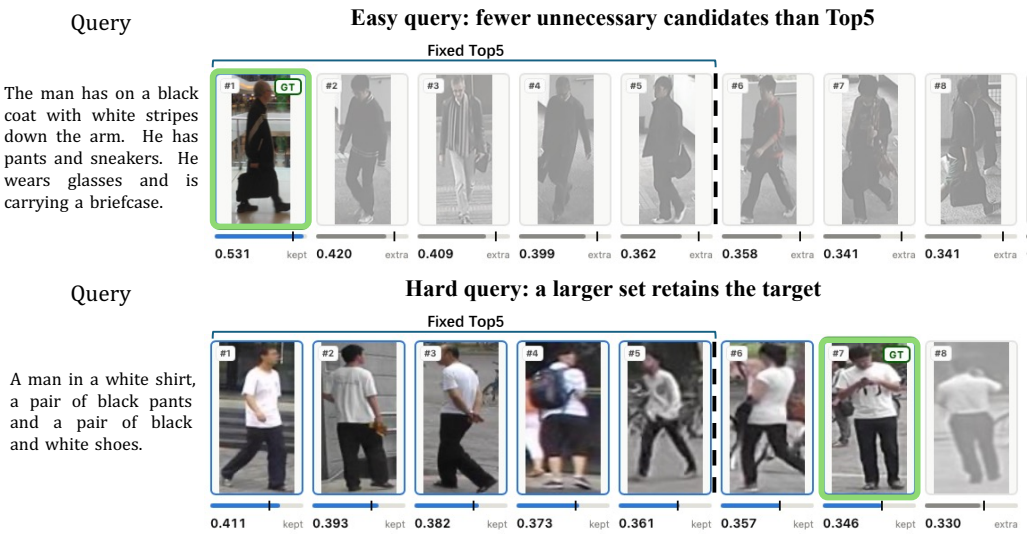


Figure 1: Conformal Identity Sets examples. As for easy query, fixed topk cutoff introduces more redundant answers. As for the hard query, fixed cutoff might exclude the correct answer.

adjustment tightening the allowable similarity margin under a common calibrated threshold. (2) As temperature increases, accumulated tail probabilities can delay stopping in the adaptive score, which approaches a rank-based rule that ignores similarity gaps. These observations raise a central question: *How does temperature shape conformal identity sets, and can a temperature-free score strike a favorable balance between set size and coverage evenness?*

To this end, we introduce a temperature-free method for constructing Conformal Identity Sets on top of frozen retrievers. Our key insight is that thresholding and adaptive scores, despite their different constructions, share a common low-temperature limit on the similarity scale. Specifically, we prove that, after order-preserving transformations, both scores converge to the similarity margin, the difference between the highest identity similarity and a candidate's similarity, as temperature approaches zero. This shared limit provides a theoretical basis for calibrating the margin directly, rather than selecting a temperature for either probability-based construction.

Our method calibrates the target identity's gap below the leader on held-out queries and uses the resulting conformal quantile as a common inclusion threshold. At inference, it returns all identities whose similarities lie within that calibrated gap of the top-ranked identity. Although the threshold is shared across queries, the set size adapts to the similarity profile: a well-separated leader can yield a singleton, whereas several closely competing identities can yield a longer shortlist. The method retains finite-sample marginal coverage under exchangeability and operates on existing retrievers without softmax, temperature tuning, or retraining.

Experiments with five frozen retrievers across three benchmarks demonstrate that our method offers a favorable trade-off between set size and coverage evenness without temperature tuning. Our method typically yields smaller average sets, while achieving more even coverage at comparable empirical coverage. Additional analyses support the predicted low-temperature convergence and examine how Margin reallocates candidates across queries of different retrieval difficulty.

Our contributions are as follows:

- We analyze how temperature affects conformal scores and prove that both scores converge to a common similarity-margin score as temperature approaches zero. This shared limit provides a theoretical foundation for our temperature-free construction.
- We develop an plug-and-play algorithm that constructs conformal identity sets for text-based person retrieval, with query-adaptive set sizes and finite-sample marginal coverage. It directly calibrates the similarity margin on top of any frozen retriever, without softmax, temperature tuning, or retraining.
- We conduct experiments on five retrieval backbones and three benchmarks, demonstrating a favorable trade-off at comparable empirical coverage: our method typically returns a more balanced

result than traditional conformal scores without temperature scaling. Additional experiments numerically verify the predicted low-temperature convergence to the similarity-margin score.

2 Related Work

Text-based Person Retrieval. Text-based person retrieval searches an image gallery for a target person using a natural-language description. Since the introduction of CUHK-PEDES (Li et al., 2017), research has primarily focused on improving retrieval accuracy through more discriminative visual–textual representations. Representative advances include cross-modal projection learning (Zhang & Lu, 2018), fine-grained correspondence discovery in CFine (Yan et al., 2023), and implicit relation reasoning with similarity distribution matching in IRRA (Jiang & Ye, 2023). More recent work addresses noisy training correspondences through robust dual embeddings in RDE (Qin et al., 2024) and explores ultra-fine-grained descriptions and alignment through UFineBench and CFAM (Zuo et al., 2024). Further advances improve generalization by modeling diverse human annotation styles in HAM (Jiang et al., 2025) and dialogue-based cross-modal alignment in DiaNA (Bai et al., 2025). Despite their different mechanisms, these approaches primarily improve the ordering of retrieved candidates, leaving a practical decision unresolved: how many answers should be returned for a particular query? Our work addresses this decision by constructing conformal identity sets with adaptive set size and guaranteed coverage.

Conformal Prediction. Conformal prediction wraps a pretrained model and returns prediction sets with finite-sample marginal coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2021), and the nonconformity score decides how those sets behave. The thresholding score LAC (Sadinle et al., 2019) keeps every label whose softmax probability clears one calibrated threshold, which gives compact sets but under-covers hard inputs. The adaptive score APS (Romano et al., 2020) keeps the shortest ranking prefix whose probability mass reaches a calibrated level, which covers more evenly but lets the softmax tail inflate the sets. RAPS and SAPS mitigate this by penalizing or discarding the tail (Angelopoulos et al., 2021; Huang et al., 2024). Both failures trace back to the softmax. LAC's size optimality and APS's conditional coverage are stated for true conditional probabilities (Sadinle et al., 2019; Romano et al., 2020), which deep networks do not provide (Guo et al., 2017). In retrieval, moreover, the softmax over gallery similarities is a post-hoc normalization with an arbitrary temperature. Our score removes the softmax: it is the raw similarity margin, the gap between the top similarity and a candidate's similarity.

3 From Softmax Scores to Similarity Margins

In this section, we first introduce the retrieval and calibration setup. We then analyze the temperature dependence of LAC and APS, derive their common transformed low-temperature limit, and develop direct margin calibration.

3.1 Preliminaries

Fixed top-K retrieval returns the same number of candidates regardless of query ambiguity, and the cutoff alone provides no prescribed coverage guarantee. We therefore seek query-adaptive identity sets with guaranteed marginal coverage. Let $\mathcal{G}$ be a fixed gallery partitioned into M nonempty identity groups $\mathcal{G}_y$, where $y \in \mathcal{Y} = \{1, \dots, M\}$. Given a text query q, a frozen retriever assigns similarity $s(q, x)$ to each gallery image x. We score each identity by its best-matching image:

$$u(q, y) = \max_{x \in \mathcal{G}_y} s(q, x). \tag{1}$$

Write $\mathbf{u}(q) = (u(q, 1), \dots, u(q, M))$ for the similarity profile and $y_{(r)}(q)$ for the rank-r identity under fixed tie-breaking, with $u_{(r)}(q) = u(q, y_{(r)}(q))$. The top-ranked identity is the *leader*. Each episode (Q, Y^*) pairs a query with a target identity enrolled in the gallery. Instead of retaining a fixed number of ranked identities, we aim to construct a query-dependent set $C_\alpha(q)$ satisfying $\Pr\{Y^* \in C_\alpha(Q)\} \geq 1 - \alpha$.

Conformal calibration provides a way to meet this requirement. Let $A(q, y)$ be a nonconformity score fixed before calibration, with smaller values indicating greater compatibility. From n calibra-

tion episodes, compute the target scores $A_i = A(Q_i, Y_i^*)$ and define

$$k_\alpha = \lceil (n+1)(1-\alpha) \rceil, \qquad \widehat{\tau}_\alpha = A_{(k_\alpha)}, \tag{2}$$

$$C_\alpha(q) = \{y \in \mathcal{Y} : A(q,y) \le \widehat{\tau}_\alpha\}, \tag{3}$$

where $A_{(k)}$ is the kth smallest calibration score and $A_{(n+1)} := +\infty$. Although the threshold is shared across queries, the number of identities satisfying it depends on the query. Scores nondecreasing along the ranking yield prefix sets, replacing the fixed cutoff with $K_\alpha(q) = |C_\alpha(q)|$. Under the exchangeability condition in Appendix A, this construction guarantees finite-sample marginal coverage of at least $1-\alpha$ over calibration data and a new episode.

The remaining choice is how to construct the nonconformity score from the similarity profile. Two standard choices, LAC and APS, use softmax weights at temperature $T > 0$:

$$p_T(q,y) = \frac{\exp(u(q,y)/T)}{\sum_{z\in\mathcal{Y}} \exp(u(q,z)/T)}, \tag{4}$$

with $p_{T,(j)}(q) = p_T(q, y_{(j)}(q))$. LAC (Sadinle et al., 2019) scores each identity individually using $A_T^{\mathrm{LAC}}(q,y) = 1 - p_T(q,y)$, whereas APS (Romano et al., 2020) accumulates weights along the ranking:

$$A_T^{\mathrm{APS}}(q, y_{(r)}(q); U) = \sum_{j=1}^{r} p_{T,(j)}(q) - U p_{T,(r)}(q), \tag{5}$$

where $U \sim \mathrm{Unif}(0,1)$ is shared across candidates within a query and drawn independently across episodes and independently of the data. The coverage guarantee follows from conformal calibration and does not require the softmax weights to be calibrated probabilities. However, temperature changes these weights without changing the ranking, and can therefore affect the resulting set sizes.

3.2 How Temperature Changes the Scores

Temperature preserves the retrieval order but changes the weights used for calibration. We define the *similarity margin* $d(q,y)$ as candidate y's gap below the leader, and the *soft count* $N_T(q)$ as a temperature-dependent weighted count of identities, including the leader:

$$d(q,y) = u_{(1)}(q) - u(q,y), \qquad d_r(q) = d(q, y_{(r)}(q)), \tag{6}$$

$$N_T(q) = \sum_{y\in\mathcal{Y}} \exp(-d(q,y)/T) = \frac{1}{\max_y p_T(q,y)}. \tag{7}$$

Here $d(q,y) \ge 0$ and $N_T(q) \in [1, M]$, with $p_T(q,y) = \exp(-d(q,y)/T)/N_T(q)$. A large soft count can arise from a crowded top or many lower-ranked identities contributing small weights.

For LAC, applying the strictly increasing map $a \mapsto -T\log(1-a)$ to calibration and test scores preserves their ordering and threshold comparisons, yielding identical prediction sets and exposing this decomposition. With the log-sum-exp reference $\ell_T(\mathbf{u}(q)) = T\log\sum_z \exp(u(q,z)/T)$, we obtain

$$\begin{aligned} S_T(q,y) &:= -T\log\big(1 - A_T^{\mathrm{LAC}}(q,y)\big) \\ &= \ell_T(\mathbf{u}(q)) - u(q,y) = d(q,y) + T\log N_T(q). \end{aligned} \tag{8}$$

The query-wide adjustment $T\log N_T(q)$ changes the gap allowed below each query's leader under a common calibration threshold. To connect the temperature endpoints, subtract the common constant $T\log M$ from all transformed scores, preserving their sets. The centered reference $\ell_T(\mathbf{u}(q)) - T\log M$ lies between the mean similarity $\bar{u}(q) = M^{-1}\sum_y u(q,y)$ and the top similarity, approaching the mean as $T \to \infty$ and the top as $T \downarrow 0$.

At high temperature, LAC therefore approaches the mean-centered score $\bar{u}(q) - u(q,y)$, yielding the rule $\{y : u(q,y) \ge \bar{u}(q) + c\}$ for a calibration-dependent constant c. APS instead approaches $(r-U)/M$, ignoring similarity gaps and randomizing its set size between adjacent integers . These are limiting rules; finite-temperature set agreement requires care at threshold boundaries. Figure 2 shows nearly overlapping set-size distributions at low temperature. At high temperature , APS concentrates on sizes three and four, whereas LAC retains a broader distribution.

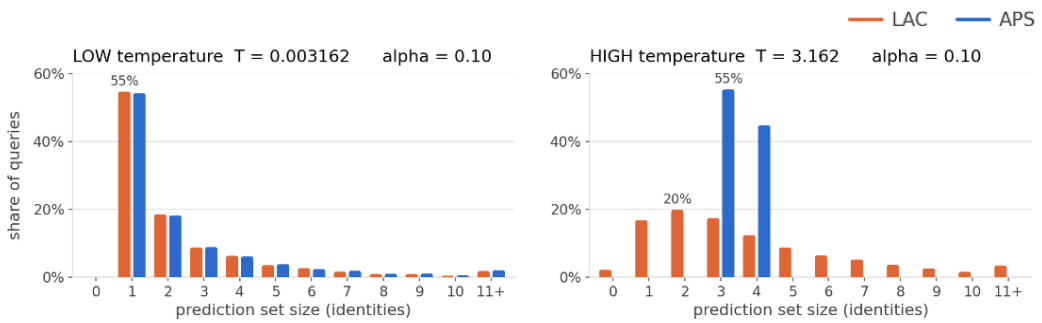


Figure 2: Temperature dependence of LAC and APS set sizes on [RDE / CUHK-PEDES] at target coverage $1 - \alpha = 0.9$.

3.3 The Low-Temperature Limit: The Similarity Margin

Although LAC and APS have different high-temperature limits, their transformed scores share the same low-temperature limit. For LAC, Eq. equation 8 and $1 \le N_T(q) \le M$ give

$$0 \le S_T(q, y) - d(q, y) \le T \log M.$$

Thus S_T converges uniformly over candidates to the similarity margin. For APS, we apply the same order-preserving transformation.

Proposition 1 (**Margin limit of APS**). Let $B_T = -T \log(1 - A_T^{\mathrm{APS}})$ be APS on the similarity scale. For a finite gallery with finite similarities, every query, rank r, and fixed $U \in (0, 1)$ satisfy

$$\left| B_T(q, y_{(r)}; U) - d_r(q) \right| \le T \left(\log M + \log(1/U) \right). \tag{9}$$

Consequently, $B_T(q, y_{(r)}; U) \to d_r(q)$ as $T \downarrow 0$. At each $T > 0$, calibrating B_T and A_T^{APS} yields identical prediction sets.

For a fixed finite calibration sample, fixed auxiliary draws in $(0, 1)$, and $k_\alpha \le n$, the transformed thresholds also converge to the margin threshold (Proofs are in Appendix B). This limit motivates calibrating $A^{\mathrm{mar}}(q, y) = d(q, y)$ directly via Eq. equation 2. We next examine how this temperature-free score changes the calibrated sets.

3.4 What Changes When We Calibrate the Margin

We compare the sets after calibrating each score on the same episodes, fixing α and omitting it from the notation. Let $\widehat{\tau}_T^{\mathrm{LAC}}$, $\widehat{\tau}_T^{\mathrm{APS}}$, and $\widehat{\tau}^{\mathrm{mar}}$ denote the thresholds of transformed LAC (S_T), original APS, and the margin, respectively, with corresponding sets C_T^{LAC}, C_T^{APS}, and C^{mar}. Proofs are in Appendix C.

Proposition 2 (**LAC sets**). For every $T > 0$ and every query q:

(i) $C_T^{\mathrm{LAC}}(q) = \{y : d(q, y) \le \widehat{\tau}_T^{\mathrm{LAC}} - T \log N_T(q)\}$.

(ii) $|C_T^{\mathrm{LAC}}(q)| \le \exp(\widehat{\tau}_T^{\mathrm{LAC}}/T)$; no identity ranked beyond this bound is included.

(iii) $C_T^{\mathrm{LAC}}(q) = \emptyset$ if and only if $N_T(q) > \exp(\widehat{\tau}_T^{\mathrm{LAC}}/T)$.

(iv) If $k_\alpha \le n$, then $N^* = \exp((\widehat{\tau}_T^{\mathrm{LAC}} - \widehat{\tau}^{\mathrm{mar}})/T) \in [1, M]$, and

$$\begin{aligned} C_T^{\mathrm{LAC}}(q) &\subseteq C^{\mathrm{mar}}(q) \quad \text{if } N_T(q) \ge N^*, \\ C_T^{\mathrm{LAC}}(q) &\supseteq C^{\mathrm{mar}}(q) \quad \text{if } N_T(q) \le N^*. \end{aligned} \tag{10}$$

Part (i) shows that a larger soft count permits a smaller gap below the leader, although set size also depends on the candidate margins. Part (ii) gives a query-independent cap on shortlist length: every included identity must carry weight at least $\exp(-\widehat{\tau}_T^{\mathrm{LAC}}/T)$, and the total weight is one. Part (iii) shows that LAC returns an empty set when even the leader falls below this cutoff; a sufficiently

large soft count can therefore exclude all candidates. Part (iv) accounts for recalibration: LAC and Margin allow the same gap at N^*. Replacing LAC with Margin can only increase target inclusion for $N_T(q) \geq N^*$ and can only decrease it for $N_T(q) \leq N^*$. This can help undercovered high-soft-count queries without guaranteeing more even coverage overall.

For APS, candidate inclusion depends on cumulative weights rather than a query-dependent gap allowance. With $k_\alpha \leq n$, we have $\widehat{\tau}_T^{\mathrm{APS}} \leq 1$, and Eq. equation 5 includes the rank-r identity exactly when

$$\sum_{j>r} p_{T,(j)}(q) + U p_{T,(r)}(q) \geq 1 - \widehat{\tau}_T^{\mathrm{APS}}.$$

Proposition 3 (**APS set sizes**). Fix $T > 0$, a query q, and $U \in [0,1]$, and let $K = |C_T^{\mathrm{APS}}(q)|$. If $K < M$, then

$$\sum_{r>K+1} p_{T,(r)}(q) < 1 - \widehat{\tau}_T^{\mathrm{APS}}.$$

In every case, $K \geq \lfloor \widehat{\tau}_T^{\mathrm{APS}} N_T(q) \rfloor$.

Many individually small tail weights can therefore prevent APS from stopping early, even when the leader is well separated . For a fixed threshold, the lower bound grows with the soft count. It does not establish monotonicity in temperature after recalibration, since the threshold changes as well.

Direct margin calibration removes both LAC's query-dependent gap adjustment and APS's cumulative weights, applying a common gap allowance $\widehat{\tau}^{\mathrm{mar}}$ below the leader. A crowded top can thus yield a longer set; for $M \geq 2$, a singleton is returned exactly when $d_2(q) > \widehat{\tau}^{\mathrm{mar}}$. Whether this rule improves efficiency or coverage evenness depends on how margins reflect query difficulty and is evaluated in Section 4; conditional coverage is not guaranteed.

3.5 Conformal Identity Sets with the Similarity Margin

We implement this common gap rule by calibrating the target margins once, as detailed in Algorithm 1. For each new query, the method returns

$$C_\alpha^{\mathrm{mar}}(q) = \{y : u(q,y) \geq u_{(1)}(q) - \widehat{\tau}_\alpha^{\mathrm{mar}}\}.$$

The sets are nonempty prefixes containing all top-scoring identities and are nested as the target coverage increases. Direct calibration requires neither temperature selection nor auxiliary randomization and avoids cumulative-softmax saturation at very low temperatures (Dabah & Tirer, 2025). Changing the gallery or retriever requires recalibration.

Algorithm 1 Conformal identity sets with the similarity margin

Require: frozen retriever s; gallery $\mathcal{G}$ with identities $\mathcal{Y}$; calibration episodes $\{(Q_i, Y_i^*)\}_{i=1}^n$; miscoverage level $\alpha \in (0,1)$

Calibration (once)

for $i = 1, \ldots, n$ **do**
  compute $u(Q_i, y)$ for all $y \in \mathcal{Y}$ by Eq. equation 1
  $a_i \leftarrow \max_y u(Q_i, y) - u(Q_i, Y_i^*)$ ▷ target's gap from the leader
end for
$k_\alpha \leftarrow \lceil (n+1)(1-\alpha) \rceil$
$\widehat{\tau}_\alpha^{\mathrm{mar}} \leftarrow a_{(k_\alpha)}$ if $k_\alpha \leq n$, and $+\infty$ otherwise

Prediction (for each test query q)

compute $u(q,y)$ for all $y \in \mathcal{Y}$
return $\{y : u(q,y) \geq \max_z u(q,z) - \widehat{\tau}_\alpha^{\mathrm{mar}}\}$, in decreasing order of $u(q,y)$ with the fixed tie-breaking rule

4 Experiments

Our experiments connect the score analysis to practical retrieval performance through three questions. First, how does Margin compare with probability-based scores and Calibrated rank in set

Table 1: CUHK-PEDES results at 90% target coverage.

Retriever	Method	Cov.	Mean size ↓	SSCV ↓	Singleton	P90
APTM Rank-1 76.17%	LAC	89.86 ± 1.02	2.49 ± 0.16	31.38 ± 23.96	50.09 ± 1.25	5.70 ± 0.63
	APS	90.91 ± 0.63	11.85 ± 1.33	6.30 ± 0.50	54.60 ± 1.00	26.51 ± 1.67
	Calibrated rank	90.83 ± 1.05	4.80 ± 0.43	–	0.00 ± 0.00	4.80 ± 0.43
	Margin (Ours)	89.90 ± 1.02	3.11 ± 0.25	11.70 ± 2.70	58.40 ± 1.30	7.12 ± 0.84
IRRA Rank-1: 73.15%	LAC	89.97 ± 0.95	2.41 ± 0.11	74.60 ± 30.97	40.95 ± 1.53	4.90 ± 0.36
	APS	90.81 ± 0.60	4.59 ± 0.24	6.80 ± 0.50	45.00 ± 1.20	11.13 ± 0.68
	Calibrated rank	90.76 ± 0.90	4.19 ± 0.39	–	0.00 ± 0.00	4.19 ± 0.39
	Margin (Ours)	90.00 ± 1.04	2.60 ± 0.14	11.90 ± 2.90	50.00 ± 1.70	5.70 ± 0.48
RDE Rank-1 75.84%	LAC	89.86 ± 0.93	2.25 ± 0.10	63.82 ± 33.07	45.06 ± 1.64	4.33 ± 0.47
	APS	90.07 ± 0.69	4.48 ± 0.24	7.00 ± 0.50	48.10 ± 1.30	11.45 ± 0.71
	Calibrated rank	90.83 ± 0.45	4.01 ± 0.10	–	0.00 ± 0.00	4.01 ± 0.10
	Margin (Ours)	89.87 ± 0.91	2.39 ± 0.11	15.50 ± 3.40	54.20 ± 1.30	5.17 ± 0.40
NAFS Rank-1 58.79%	LAC	89.91 ± 1.08	8.22 ± 0.55	11.19 ± 1.61	3.58 ± 0.55	14.94 ± 1.14
	APS	90.32 ± 0.99	12.01 ± 0.86	7.00 ± 0.70	15.50 ± 1.10	28.24 ± 2.32
	Calibrated rank	90.27 ± 1.05	11.84 ± 0.80	–	0.00 ± 0.00	11.84 ± 0.80
	Margin (Ours)	89.87 ± 1.06	8.48 ± 0.60	9.10 ± 0.90	9.00 ± 1.20	17.98 ± 1.23
TIPCB Rank-1 63.44%	LAC	90.02 ± 0.98	5.85 ± 0.41	14.46 ± 3.26	14.81 ± 1.12	11.75 ± 0.95
	APS	90.55 ± 0.72	8.98 ± 0.42	6.70 ± 0.50	27.50 ± 1.10	23.11 ± 1.22
	Calibrated rank	90.42 ± 0.94	9.26 ± 0.75	–	0.00 ± 0.00	9.26 ± 0.75
	Margin (Ours)	90.04 ± 0.96	6.16 ± 0.39	9.10 ± 1.30	23.70 ± 1.20	14.46 ± 1.07

size and coverage evenness, accounting for achieved coverage (Section 4.2) Second, how does temperature affect LAC and APS, and do their transformed scores approach the common margin limit predicted by our theory (Section 4.3) Finally, beyond aggregate performance, does Margin adapt shortlist lengths to retrieval difficulty We examine this by comparing its set sizes and coverage with Calibrated rank across ground-truth rank groups (Section 4.4).

4.1 Experimental Setup

We evaluate five frozen retrievers: APTM (Yang et al., 2023), IRRA (Jiang & Ye, 2023), NAFS (Gao et al., 2021), RDE (Qin et al., 2024), and TIPCB (Chen et al., 2022) on CUHK-PEDES (Li et al., 2017), ICFG-PEDES (Ding et al., 2021), and RSTPReid (Zhu et al., 2021); NAFS and TIPCB are evaluated only on CUHK-PEDES. These retrievers span three backbone families: APTM uses Swin Transformer, IRRA and RDE use ViT, and NAFS and TIPCB use ResNet. All methods share the identity scores in Eq. equation 1 and follow the identity-level protocol in Section 3.1, using every caption as a query. We primarily compare Margin with Calibrated rank, which provides the same marginal coverage guarantee using a fixed, calibrated rank cutoff, to assess whether query-adaptive set sizes improve efficiency at the target coverage level. We additionally compare with LAC (Sadinle et al., 2019) and APS (Eq. equation 5) to examine temperature-dependent softmax calibration. All methods use the same calibration episodes and quantile rule; LAC and APS also select their temperatures on these episodes. We report empirical coverage, mean set size, the 90th percentile of set size (P90), singleton rate, and size-stratified coverage violation (SSCV) (Angelopoulos et al., 2021), interpreting efficiency alongside achieved coverage. Configuration details and additional baselines appear in Appendices E, D, and G. For the results in Table 1, the temperatures of LAC and APS are fitted on the calibration episodes using standard temperature scaling (Guo et al., 2017); Margin and Calibrated rank require no temperature tuning. We randomly partition identities into disjoint calibration and evaluation sets in a 50:50 ratio and repeat the experiments over 100 random partitions.

4.2 Coverage and Shortlist Efficiency

Table 1 shows that Margin performs consistently better than calibrated rank and offers a balance between LAC's compact sets and APS's lower SSCV. **Cov.** is the percentage of queries whose returned identity sets contain the target identity. **Mean size** is the average number of distinct identities returned per query. **SSCV** (size-stratified coverage violation) is the largest absolute deviation from

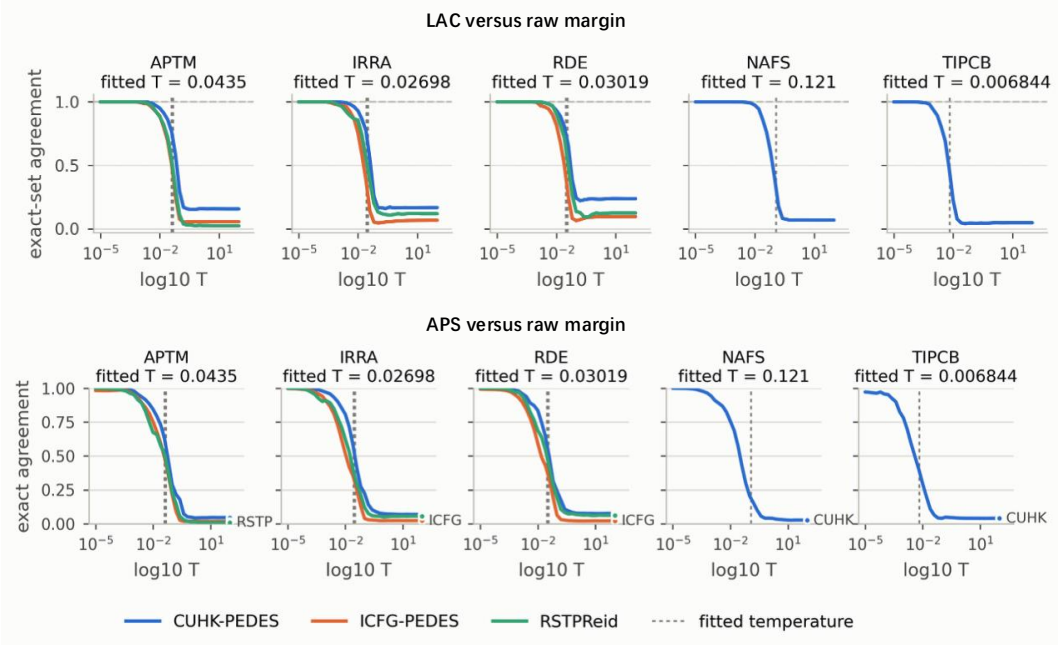


Figure 3: Agreement with Margin on five methods.(RDE/CUHK-PEDES, 90% coverage)

target coverage across five fixed set-size bins (1, 2–3, 4–6, 7–10, and 11 or more.), measured in percentage points, with unpopulated bins assigned zero coverage. Since calibrated rank isn't query-adaptive, we do not we report the SSCV of calibrated rank. **Singleton** is the percentage of queries for which exactly one identity is returned. **P90** is the 90th percentile of the returned identity-set size, characterizing the size of larger shortlists. Additional results are reported in Appendix F.

4.3 Testing the Low-Temperature Limit

We test whether the transformed LAC and APS scores converge to the margin as predicted in Section 3.3. For each partition, we fix the calibration and evaluation episodes and auxiliary draws, and recalibrate each score separately as temperature decreases. We compute S_T (Eq. equation 8) and $B_T = -T\log(1 - A_T^{\mathrm{APS}})$ using stable log-sum-exp expressions, and measure their mean absolute differences from the margin and the absolute differences between their calibrated thresholds. For $s \in \{\mathrm{LAC}, \mathrm{APS}\}$, we also measure exact-set disagreement:

$$D^s_{\mathrm{set}}(T) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathbf{1}\{C^s_{\alpha,T}(q_i) \neq C^{\mathrm{mar}}_{\alpha}(q_i)\}\,. \tag{11}$$

Figure 3 plots the corresponding agreement, $1 - D^s_{\mathrm{set}}(T)$. To distinguish convergence away from the threshold from boundary effects, we fix $\epsilon > 0$ before examining the results and define $\mathcal{I}_\epsilon$ as the candidate pairs (i, y) satisfying $|A^{\mathrm{mar}}(q_i, y) - \widehat{\tau}^{\mathrm{mar}}_{\alpha}| > \epsilon$. We report candidate-inclusion disagreement on $\mathcal{I}_\epsilon$ and the fraction of candidates outside it. Across all evaluated retriever–dataset pairs, exact-set agreement with Margin approaches one as temperature decreases for both LAC and APS.

4.4 Query-adaptive set sizes

Figure 4 compares Margin with Calibrated rank on RDE and CUHK-PEDES, grouping queries by the rank of their ground-truth identity under the frozen retriever. Calibrated rank returns a fixed number of candidates within each split, whereas Margin returns progressively larger sets across groups with worse ground-truth ranks. For queries whose target ranks first, Margin substantially reduces set size while preserving full coverage. For targets ranked beyond the fixed cutoff, it expands the candidate list and recovers some identities that Calibrated rank misses, most visibly in the rank-> 10 group at 95% target coverage. This redistribution does not improve coverage in every group: Margin has lower coverage in some intermediate-rank groups. The results demonstrate query-adaptive

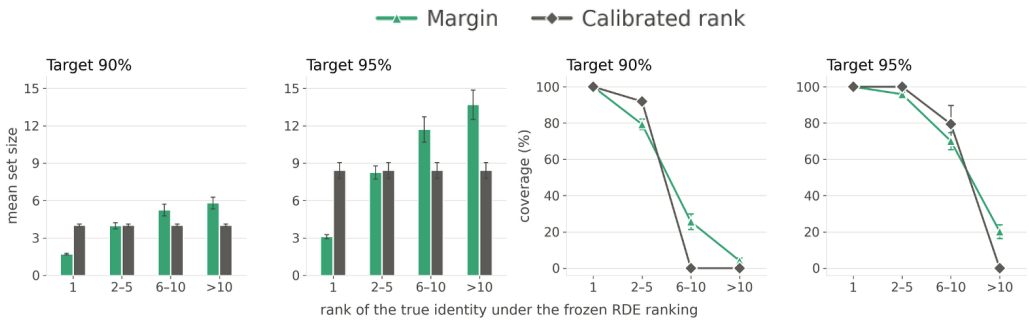


Figure 4: Query adaptivity on RDE and CUHK-PEDES: mean set size and coverage by ground-truth rank at 90% and 95% target coverage. Error bars show standard deviations over 100 splits.

candidate allocation, with shorter lists for easier retrieval cases and additional coverage for some harder cases, rather than a uniform improvement in groupwise coverage.

4.5 Coverage Across Soft-Count Groups

Figure 5 compares LAC and Margin across soft-count groups on RDE / CUHK-PEDES at 90% target coverage. Consistent with Proposition 2(iv), Margin returns smaller sets with slightly lower coverage when $N_T(q)/N^* < 1$, and larger sets with higher coverage above this crossover. The largest gains occur in the highest soft-count group, where LAC undercovers most severely. Margin thus alleviates undercoverage in these queries at the cost of larger sets, without ensuring 90% coverage within each group.

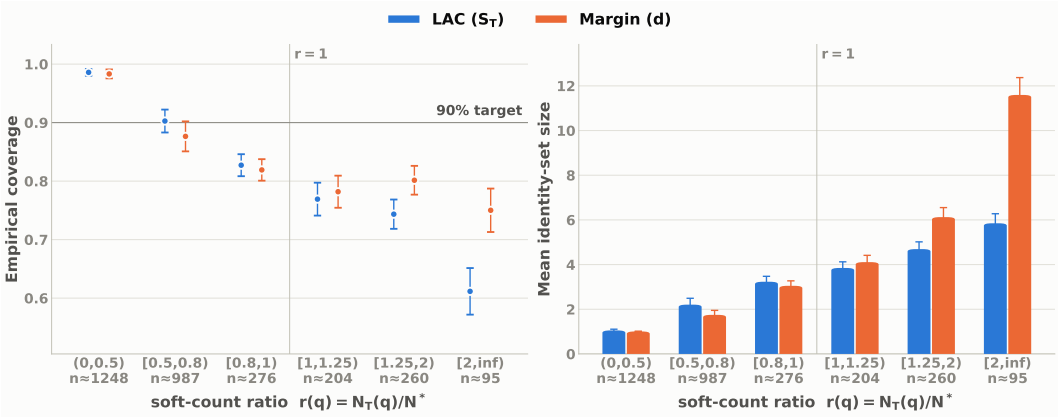


Figure 5: Coverage (left) and mean set size (right) across soft-count groups on RDE / CUHK-PEDES at 90% target coverage. Margin increases set size and coverage above the crossover $r = 1$.

5 Conclusion

We introduced a temperature-free method for constructing conformal identity sets in text-based person retrieval. Our analysis shows that suitably transformed LAC and APS scores converge to a common similarity-margin score as temperature approaches zero, motivating direct calibration of this margin. The resulting method produces query-adaptive shortlists with finite-sample marginal coverage under exchangeability, without softmax, temperature tuning, or retraining. Experiments across multiple retrievers and benchmarks demonstrate a favorable trade-off between set size and coverage evenness. Future work will investigate extending margin calibration to distribution shifts and improving coverage across query groups.

AI USE STATEMENT

LLMs did not play a significant role in this work and were only used for grammar polishing in writing.

ETHICS STATEMENT

Our study evaluates uncertainty calibration on existing person-retrieval benchmarks, and the resulting candidate sets should not be treated as definitive identity decisions. The coverage guarantee is marginal and relies on exchangeability; it does not ensure correctness for individual queries or equal coverage across demographic groups. Practical deployment would require appropriate data authorization, privacy safeguards, human oversight, and evaluation of potential biases inherited from the data and frozen retrievers.

REPRODUCIBILITY STATEMENT

We specify the identity-level scoring rule, nonconformity scores, and conformal quantile construction in the main text. The appendix provides the theoretical assumptions and proofs, including the exchangeability conditions underlying the coverage guarantee. We describe the datasets, frozen retrievers, calibration protocol, temperature selection, and evaluation metrics in the experimental section and accompanying appendix. Results are reported as means and standard deviations over 100 matched random identity partitions to characterize variability across calibration and evaluation splits.

REFERENCES

Navid Akhavan Attar, Hesam Asadollahzadeh, Ling Luo, and Uwe Aickelin. Softmax is not enough (for adaptive conformal classification). In *The Fourteenth International Conference on Learning Representations*, 2026.

Marharyta Aleksandrova and Oleg Chertov. Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: An extensive study. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 151–170. PMLR, 2021.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3952–3962, 2025.

Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494: 171–181, 2022. doi: 10.1016/j.neucom.2022.04.081.

Lahav Dabah and Tom Tirer. On temperature scaling and conformal prediction of deep classifiers. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11813–11845. PMLR, 2025.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20331–20347. PMLR, 2024.

Yotam Intrator, Regev Cohen, Ori Kelner, Roman Goldenberg, Ehud Rivlin, and Daniel Freedman. Streamlining conformal information retrieval via score refinement. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 186–191. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.fever-1.22.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2787–2797, 2023.

Jiayu Jiang, Changxing Ding, Wentao Tan, Junhong Wang, Jin Tao, and Xiangmin Xu. Modeling thousands of human annotators for generalizable text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9220–9230, 2025.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1970–1979, 2017.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27197–27206, 2024.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019. doi: 10.1080/01621459.2017.1395341.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. CLIP-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 2023. doi: 10.1109/TIP.2023.3327924.

Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4492–4501. ACM, 2023. doi: 10.1145/3581783.3611709.

Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision*, 2018.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. DSSL: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 209–217. ACM, 2021. doi: 10.1145/3474085.3475369.

Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. UFineBench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22010–22019, 2024.

A Exchangeability and Marginal Coverage

This section states the sampling condition behind the coverage guarantee of Section 3.1. Recall that $\mathcal{H}$ collects the fixed gallery and the frozen retriever; the gallery stays complete, so identities assigned to every block remain retrieval candidates.

Assumption 1 (Exchangeable query episodes)**.** *Conditional on $\mathcal{H}$, the calibration episodes $(Q_1, Y_1^*), \ldots, (Q_n, Y_n^*)$ and a test episode (Q_{n+1}, Y_{n+1}^*) are exchangeable.*

Sampling protocol. On benchmarks, we partition identities uniformly at random into calibration (50%), and evaluation blocks (50%), so no identity contributes captions to more than one block. Every caption of an identity becomes a query episode in that identity's block, and the episodes within each block are ordered at random, so n counts calibration captions. This prevents identity overlap but does not by itself establish episode-level exchangeability for pooled captions; coverage under this protocol is assessed empirically.

Parameter selection. The fixed-score guarantee requires score parameters to be chosen before calibration. Parameters selected on separate design data can be treated as fixed conditional on that information, provided conditional exchangeability holds. Selecting parameters on the same calibration episodes requires a separate validity argument.

Scores and prefix structure. A nonconformity score may depend on the whole similarity profile, provided the same rule, fixed given $\mathcal{H}$, is applied to calibration and test episodes; for randomized scores, each episode carries an independent auxiliary draw from a common distribution. The set of Eq. equation 3 is a prefix of the ranking, with $K_\alpha(q) = 0$ denoting the empty set, whenever the score is nondecreasing along the ranking for every realization of the auxiliary randomness. Prefix structure is thus a property of the score, whereas coverage follows from calibration alone.

Proposition 4 (Split-conformal marginal coverage). Under Assumption 1, the set in Eq. equation 3 satisfies

$$\Pr\{Y_{n+1}^* \in C_\alpha(Q_{n+1}) \,|\, \mathcal{H}\} \geq 1 - \alpha. \tag{12}$$

The probability averages over the calibration and test episodes and any auxiliary randomization.

Proof. Conditional on $\mathcal{H}$, applying the same fixed score to exchangeable episodes produces exchangeable scores. If $k_\alpha \leq n$, the standard split-conformal rank argument (Vovk et al., 2005; Angelopoulos & Bates, 2021) gives coverage at least $k_\alpha/(n+1) \geq 1 - \alpha$; ties preserve this lower bound because inclusion in Eq. equation 3 is non-strict. If $k_\alpha = n + 1$, the threshold is infinite and the set contains every identity. □

The guarantee is marginal: it does not assert coverage conditional on a query, an identity, or the realized calibration sample.

B Low-Temperature Limits of the Softmax Scores

This section proves Proposition 1, establishes threshold convergence and inclusion agreement away from the margin threshold, and bounds the resulting difference in set size.

B.1 Proof of Proposition 1

Proof. Fix a query, rank r, and $U \in (0, 1)$, and write $d_j = u_{(1)} - u_{(j)}$. The residual probability mass is

$$1 - A_T^{\mathrm{APS}}(q, y_{(r)}; U) = \frac{U e^{-d_r/T} + \sum_{j>r} e^{-d_j/T}}{\sum_{j=1}^{M} e^{-d_j/T}}.$$

Set $D_T = \sum_j e^{-d_j/T}$ and $R_T = U + \sum_{j>r} e^{-(d_j - d_r)/T}$. Since $d_j \geq d_r$ for $j > r$,

$$1 \leq D_T \leq M, \qquad U \leq R_T \leq M.$$

Hence

$$B_T(q, y_{(r)}; U) - d_r = T \log D_T - T \log R_T$$

lies between $-T \log M$ and $T \log(M/U)$, which implies Eq. equation 9, including for profiles with tied scores.

For calibration equivalence, the map $\varphi_T(a) = -T \log(1-a)$ is strictly increasing on $[0, 1)$, with $\varphi_T(1) = +\infty$. For $k_\alpha \le n$, it commutes with order statistics, so the threshold of B_T is $\varphi_T(\widehat{\tau}_T^{\mathrm{APS}})$ and preserves every inclusion comparison. For $k_\alpha = n+1$, both procedures use an infinite threshold and return the full gallery. The same argument applies to original LAC and S_T in Eq. equation 8; $\widehat{\tau}_T^{\mathrm{LAC}}$ in the main text denotes the threshold of S_T. □

B.2 Calibrated Thresholds and Inclusion Decisions

For fixed calibration episodes and auxiliary draws, let $a_i = A^{\mathrm{mar}}(Q_i, Y_i^*)$, and let $b_i(T)$ be the calibration scores of either S_T (Eq. equation 8) or B_T (Proposition 1). Since $N_T \in [1, M]$, Eq. equation 8 gives $0 \le S_T - d \le T \log M$; together with Eq. equation 9, this implies $\max_i |b_i(T) - a_i| \to 0$. Order statistics are nonexpansive in the supremum norm, so for $k_\alpha \le n$,

$$\left| b_{(k_\alpha)}(T) - a_{(k_\alpha)} \right| \le \max_i |b_i(T) - a_i| \to 0.$$

For a fixed test candidate whose margin differs from $a_{(k_\alpha)}$, score and threshold convergence preserve the sign of their difference for all sufficiently small T. This establishes eventual agreement away from the boundary, and full-set agreement for a finite test collection if none of its candidates lies on the boundary. No such claim follows at equality. In particular, top-scoring candidates always have margin zero; when the calibrated threshold is zero, their boundary status must be retained in any set-agreement analysis.

B.3 Set-Size Stability of Low-Temperature APS

Proposition 1 controls scores; this subsection controls the number of returned identities. Fix the calibration episodes and their auxiliary draws $U_1, \ldots, U_n \in (0, 1)$, and assume $k_\alpha \le n$. Write

$$e_T(U) = T\left(\log M + \log(1/U)\right), \qquad \varepsilon_T = \max_{i \le n} e_T(U_i), \qquad \tau = \widehat{\tau}_\alpha^{\mathrm{mar}},$$

and let $K_{\alpha,T}^{\mathrm{APS}}(q; U)$ and $K_\alpha^{\mathrm{mar}}(q)$ be the sizes of the APS set (calibrated at temperature T) and the margin set for a test query q with auxiliary draw U.

Proposition 5 (Set-size stability). For every $T > 0$, test query q, and $U \in (0, 1)$,

$$\left| K_{\alpha,T}^{\mathrm{APS}}(q; U) - K_\alpha^{\mathrm{mar}}(q) \right| \le \sum_{r=1}^{M} \mathbf{1}\{|d_r(q) - \tau| \le e_T(U) + \varepsilon_T\}. \tag{13}$$

If, in addition, the test query almost surely has no candidate with $d_r(Q_{n+1}) = \tau$, then, conditional on the calibration sample and its auxiliary draws, $\mathbb{E}\,|K_{\alpha,T}^{\mathrm{APS}} - K_\alpha^{\mathrm{mar}}| \to 0$ as $T \downarrow 0$; in particular, the mean sizes converge.

Proof. Let τ_T be the calibrated threshold of B_T. By the preceding subsection, $|\tau_T - \tau| \le \varepsilon_T$, and calibrating B_T yields the APS sets (Proposition 1).

Consider the rank-r candidate of q. By Proposition 1, $|B_T(q, y_{(r)}; U) - d_r(q)| \le e_T(U)$. If $d_r(q) < \tau - e_T(U) - \varepsilon_T$, then $B_T(q, y_{(r)}; U) < \tau - \varepsilon_T \le \tau_T$, so both sets include it; if $d_r(q) > \tau + e_T(U) + \varepsilon_T$, then $B_T(q, y_{(r)}; U) > \tau + \varepsilon_T \ge \tau_T$, so both exclude it. Hence the two sets can disagree only on candidates with $|d_r(q) - \tau| \le e_T(U) + \varepsilon_T$. Both sets are prefixes of the same ranking: the APS scores increase strictly along it (proof of Proposition 3 in Appendix C), and the margin is nondecreasing along it. The difference of two prefix lengths equals the number of candidates in exactly one of them, which gives Eq. equation 13.

For the mean, $e_T(U) + \varepsilon_T \to 0$ as $T \downarrow 0$ for every $U \in (0, 1)$, so each indicator converges to $\mathbf{1}\{d_r(q) = \tau\}$, and under the additional condition the right-hand side converges to zero almost surely. It is bounded by M, so dominated convergence gives the claim. □

The condition fails when $\tau = 0$, since the leader always has margin zero; this is the boundary case noted in the preceding subsection. The bound also predicts where disagreement concentrates: on candidates near the margin threshold, which the off-boundary disagreement of Section 4.3 excludes by construction.

C Proofs for Section 3.4

Proof of Proposition 2. By Eq. equation 8, $S_T(q, y) \le \widehat{\tau}_T^{\mathrm{LAC}}$ if and only if $d(q, y) \le \widehat{\tau}_T^{\mathrm{LAC}} - T \log N_T(q)$, which is (i). Every $y \in C_T^{\mathrm{LAC}}(q)$ has

$$p_T(q, y) = \exp(-S_T(q, y)/T) \ge \exp(-\widehat{\tau}_T^{\mathrm{LAC}}/T).$$

Summing over the set and using $\sum_y p_T(q, y) = 1$ gives

$$|C_T^{\mathrm{LAC}}(q)| \exp(-\widehat{\tau}_T^{\mathrm{LAC}}/T) \le 1.$$

Since the set is a prefix of the ranking, it contains no identity of rank greater than $\exp(\widehat{\tau}_T^{\mathrm{LAC}}/T)$, which is (ii). The smallest score is the leader's, $S_T(q, y_{(1)}(q)) = T \log N_T(q)$, so the set is empty exactly when $T \log N_T(q) > \widehat{\tau}_T^{\mathrm{LAC}}$, which is (iii).

For part (iv), Eq. equation 8 and $N_T \in [1, M]$ give $A^{\mathrm{mar}} \le S_T \le A^{\mathrm{mar}} + T \log M$ on every calibration episode. Order statistics preserve pointwise inequalities, so $\widehat{\tau}^{\mathrm{mar}} \le \widehat{\tau}_T^{\mathrm{LAC}} \le \widehat{\tau}^{\mathrm{mar}} + T \log M$ when $k_\alpha \le n$, and hence $N^* \in [1, M]$. By (i), $C_T^{\mathrm{LAC}}(q)$ is the set of identities whose margin is at most $\widehat{\tau}_T^{\mathrm{LAC}} - T \log N_T(q)$, and this threshold is at most $\widehat{\tau}^{\mathrm{mar}}$ if and only if $N_T(q) \ge N^*$. Sets of the form $\{y : d(q, y) \le t\}$ are nested in t, which gives both inclusions. □

Proof of Proposition 3. Write $p_r = p_{T,(r)}(q)$, $A_r = A_T^{\mathrm{APS}}(q, y_{(r)}(q); U)$, and $\widehat{\tau} = \widehat{\tau}_T^{\mathrm{APS}}$, with $k_\alpha \le n$ as in the main text. Since $A_{r+1} - A_r = U p_r + (1 - U) p_{r+1} > 0$, the set is a ranking prefix. If $K < M$, exclusion of rank $K + 1$ gives

$$\sum_{j > K+1} p_j \le \sum_{j > K+1} p_j + U p_{K+1} < 1 - \widehat{\tau},$$

proving the first claim. Equivalently, $\sum_{j \le K+1} p_j > \widehat{\tau}$. Since $p_j \le p_1 = 1/N_T(q)$, we obtain $K > \widehat{\tau} N_T(q) - 1$, hence $K \ge \lfloor \widehat{\tau} N_T(q) \rfloor$. If $K = M$, this bound follows from $\widehat{\tau} N_T(q) \le N_T(q) \le M$. □

D Rank-Based Baselines and Temperature Limits

APS at high temperature. For a fixed finite profile, $p_{T,(r)} \to 1/M$ as $T \to \infty$. Thus

$$A_T^{\mathrm{APS}}(q, y_{(r)}; U) \longrightarrow \frac{r - U}{M}. \tag{14}$$

Thresholding this limiting score at t gives $K = \min\{M, \max\{0, \lfloor Mt + U \rfloor\}\}$, which takes at most two adjacent integer values as U varies. The cutoff does not depend on the similarity gaps, although the returned identities still follow the query's ranking.

SAPS. Using the same convention for U as in Eq. equation 5, the SAPS score (Huang et al., 2024) with fixed $\lambda > 0$ is

$$A_T^{\mathrm{SAPS}}(q, y_{(r)}; U) = \begin{cases} (1 - U) p_{T,(1)}(q), & r = 1, \\ p_{T,(1)}(q) + \lambda(r - 1 - U), & r \ge 2. \end{cases}$$

This convention replaces the uniform variable in the original formula by $1 - U$, leaving its distribution unchanged. If the leading identity is unique, $p_{T,(1)} \to 1$ as $T \downarrow 0$, whereas $p_{T,(1)} \to 1/M$ as $T \to \infty$. At either endpoint, substitution gives a score depending only on rank, auxiliary randomness, and fixed parameters. At finite temperature, the maximum probability is its only profile-dependent quantity beyond rank. This does not rule out useful adaptation at finite temperatures.

Table 2: Results on ICFG-PEDES for the three retrievers available on this benchmark. Coverage is a percentage; set sizes count distinct identities.

Retriever	Method	Cov.	Mean size
APTM	Calibrated rank	90.27 ± 0.81	9.91 ± 0.75
	APS	90.44 ± 0.47	25.41 ± 1.79
	Margin	89.94 ± 0.68	5.63 ± 0.24
IRRA	Calibrated rank	90.25 ± 0.77	10.99 ± 0.75
	APS	90.27 ± 0.52	8.85 ± 0.31
	Margin	89.91 ± 0.73	6.53 ± 0.27
RDE	Calibrated rank	90.30 ± 0.80	9.34 ± 0.71
	APS	90.74 ± 0.52	8.28 ± 0.31
	Margin	89.93 ± 0.76	5.45 ± 0.28

RAPS. For a fixed rank cutoff $k_{\mathrm{reg}} \in \{1, \dots, M\}$ and fixed $\lambda > 0$, consider the penalized RAPS score (Angelopoulos et al., 2021)

$$A_T^{\mathrm{RAPS}}(q, y_{(r)}; U) = A_T^{\mathrm{APS}}(q, y_{(r)}; U) + \lambda(r - k_{\mathrm{reg}})_+.$$

With a unique leading identity, its low-temperature limit is

$$\begin{cases} 1 - U, & r = 1, \\ 1 + \lambda(r - k_{\mathrm{reg}})_+, & r \geq 2. \end{cases}$$

At high temperature the limit is $(r-U)/M+\lambda(r-k_{\mathrm{reg}})_+$. These limits assume fixed penalties rather than penalties that vanish or otherwise vary with temperature. If the top similarity is tied, the low-temperature probabilities also depend on its multiplicity. Moreover, plateaus in a limiting score, such as unpenalized RAPS ranks, can coincide with a limiting calibration threshold; the limiting scores then do not determine all finite-temperature inclusion decisions. The same boundary caution applies to the other endpoint limits.

E Experimental Configuration

All methods use the same identity-score matrices and 100 matched random identity partitions; the gallery remains complete throughout. The main comparison targets $1 - \alpha = 0.90$, and the query-adaptivity analysis also considers 0.95. Each temperature in the limit experiment receives a separate conformal threshold, with episodes and auxiliary draws held fixed across temperatures.

For the main LAC and APS baselines, temperature is selected by minimizing target-identity negative log-likelihood on the calibration episodes. This reuse of calibration data is subject to the qualification in Appendix A. SSCV follows the five fixed size bins and empty-bin convention of Section 4.1. Score exports containing missing-score placeholders are excluded unless complete similarities can be recovered.

F Results on ICFG-PEDES and RSTPReid

Tables 2 and 3 report coverage and mean identity-set size for APTM, IRRA, and RDE at 90% target coverage, as mean $\pm$ standard deviation over 100 matched partitions. Only available measurements are shown. Margin has smaller mean sets than APS and Calibrated rank for every retriever on both benchmarks; APS generally achieves higher coverage. These comparisons should therefore be interpreted jointly with achieved coverage.

G Results for a Refinement Baseline

Table 4 reports a refinement-score baseline under the protocol, partitions, and target coverage of Section 4.2, with Margin repeated from Table 1 and Appendix F.

Table 3: Results on RSTPReid for the three retrievers available on this benchmark. Coverage is a percentage; set sizes count distinct identities.

Retriever	Method	Cov.	Mean size
APTM	Calibrated rank	90.39 ± 2.15	7.98 ± 1.44
	APS	90.13 ± 1.37	9.51 ± 1.14
	Margin	89.91 ± 2.02	6.15 ± 0.81
IRRA	Calibrated rank	90.54 ± 2.73	8.77 ± 1.48
	APS	90.51 ± 2.41	7.35 ± 0.86
	Margin	89.74 ± 2.75	6.30 ± 1.02
RDE	Calibrated rank	90.50 ± 2.65	7.49 ± 1.57
	APS	90.01 ± 2.21	6.37 ± 0.69
	Margin	89.65 ± 2.63	4.90 ± 0.69

Table 4: Coverage and mean identity-set size of the refinement baseline and Margin at target coverage $1 - \alpha = 0.90$, as mean $\pm$ standard deviation across 100 matched partitions. Coverage is a percentage, and set sizes count distinct identities.

		Refinement		Margin	
Benchmark	Retriever	Cov.	Mean size	Cov.	Mean size
CUHK-PEDES	APTM	89.86 ± 0.98	3.08 ± 0.18	89.90 ± 1.02	3.11 ± 0.25
	IRRA	89.99 ± 0.97	2.98 ± 0.16	90.00 ± 1.04	2.60 ± 0.14
	RDE	89.82 ± 0.93	2.76 ± 0.17	89.87 ± 0.91	2.39 ± 0.11
	NAFS	89.93 ± 1.08	9.20 ± 0.67	89.87 ± 1.06	8.48 ± 0.60
	TIPCB	90.02 ± 0.99	8.33 ± 0.68	90.04 ± 0.96	6.16 ± 0.39
ICFG-PEDES	APTM	89.95 ± 0.79	6.84 ± 0.45	89.94 ± 0.68	5.63 ± 0.24
	IRRA	89.92 ± 0.80	8.41 ± 0.54	89.91 ± 0.73	6.53 ± 0.27
	RDE	89.92 ± 0.82	6.90 ± 0.46	89.93 ± 0.76	5.45 ± 0.28
RSTPReid	APTM	89.93 ± 2.24	5.29 ± 0.82	89.91 ± 2.02	6.15 ± 0.81
	IRRA	89.93 ± 2.84	6.94 ± 1.22	89.74 ± 2.75	6.30 ± 1.02
	RDE	89.86 ± 2.88	5.65 ± 1.07	89.65 ± 2.63	4.90 ± 0.69