SOCCERCANON: A CHALLENGING BENCHMARK AND THE CASE FOR STITCHING VISION EXPERTS OVER VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Soccer player recognition in broadcast imagery remains challenging under blur, occlusion, small scale, and extreme viewpoints, where identity cues such as faces, appearance, and jersey numbers are often incomplete or unreliable. Existing benchmarks provide limited support for systematically evaluating recognition under such heterogeneous visual evidence. To address this limitation, we introduce *SoccerCanon*, a benchmark of 1,108 labeled player instances, comprising 613 validation and 495 test instances. Face recognizers that identify clean player photographs with about 96% R@1 reach only 40–47% on SoccerCanon, and accuracy collapses on the small faces that dominate broadcast imagery. We evaluate specialized vision models and general-purpose vision-language models (VLMs) under controlled, modality-aware settings. Our experiments reveal that different visual cues exhibit complementary strengths, while no single recognition pathway remains consistently reliable across challenging cases. Motivated by this observation, we propose a training-free expert-stitching framework that combines face, appearance, and number experts according to their query-specific reliability. Our framework reaches 60.94% R@1 on the face-available test observations without match metadata, whereas the strongest VLM reaches 29.24% even when given match-restricted candidates, and reliability-aware stitching improves most where face evidence is weak.

1 INTRODUCTION

Identifying a professional soccer player from broadcast imagery is fundamentally different from assigning a temporary track identity. The goal is to resolve an observed player to a persistent, real-world identity—for example, *Son Heung-min*—that remains semantically valid across different actions, matches, and seasons. This setting is especially challenging in professional broadcasts: a player may occupy only a small image region, while motion blur, partial occlusion, compression, and side or rear viewpoints can severely degrade the cues that normally reveal identity. A face may be small, blurred, or seen only from the side, uniform appearance may narrow the relevant team without uniquely identifying the player, and a jersey number may be unreadable or ambiguous across teams and seasons. Consequently, canonical player identification cannot reliably depend on a single visual recognition pathway.

Existing soccer player recognition benchmarks and systems capture important parts of this problem, but typically under different identity scopes or observation settings. Early studies use facial appearance, jersey numbers, and textual broadcast cues for player recognition (Bertini et al., 2006b;a; Ballan et al., 2007), while later approaches use spatial trajectories, jersey-number recognition, tracking, or external player information (Gerke et al., 2017; Diop et al., 2022; Solberg et al., 2024). Other work emphasizes action-local visual re-identification or long-term identification from multi-frame track fragments (Comandur, 2025; Dusov et al., 2026). These settings are valuable, but they do not jointly capture the case where a *single challenging broadcast observation* must be linked to a *canonical real-world player identity* using heterogeneous reference evidence. Recognizing a team, jersey number, or match-local identity alone does not establish who the observed player is.

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.

To study this setting, we introduce **SoccerCanon**, a benchmark for canonical soccer player identification from challenging professional broadcast imagery derived from SoccerNet (Deliege et al., 2021). SoccerCanon contains **1,108 manually labeled player observations**, comprising **613 validation** and **495 test** instances, and covers **424 canonical player identities**, **213 matches**, **99 teams**, **five professional competitions**, and **three seasons**. The benchmark is intentionally built from diverse match footage rather than controlled portrait imagery, preserving the variability of real broadcast production across competitions, seasons, viewpoints, player scales, and visual conditions. Each query corresponds to a single localized player observation, from which face, upper-body appearance, and jersey-number regions are automatically derived as complementary views of the same query. Apart from a 75-observation validation pilot, observations are chosen by an identity-agnostic selector that requires a weakly detectable face but no readable number. Canonical identities are then manually verified using match metadata, team–season rosters, visible player cues, and surrounding video context.

SoccerCanon also exposes a central modeling challenge: the reliability of each identity cue changes from query to query. A clear face can provide highly discriminative player-level evidence, whereas the same pathway may be nearly unusable for a small, blurred, or side-view face. Appearance can remain informative when the face is weak, but it primarily constrains team or team–season identity rather than the individual player. Jersey-number evidence can sharply reduce ambiguity within a roster, yet OCR may fail under blur, occlusion, or extreme scale. This motivates an approach that preserves the specialization of each cue while allowing its contribution to vary with the observation.

We therefore propose a **training-free reliability-aware expert-stitching framework**. The face expert provides direct player-level identity evidence, while the appearance (kit) and number experts produce structured team–season and jersey-number evidence that is mapped to canonical players through team–season rosters. The resulting heterogeneous signals are projected into a shared player space and combined according to their query-specific reliability. Rather than requiring a single general-purpose multimodal model to jointly infer every cue, the framework allows each specialized expert to contribute only when its corresponding evidence is informative.

Finally, we ask whether this explicit specialization remains useful in the presence of modern visual encoders such as CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024) and vision-language models (VLMs) (Alayrac et al., 2022; Dai et al., 2023; Bai et al., 2025; Li et al., 2023; Liu et al., 2023; Wu et al., 2024; Zhu et al., 2024). We therefore compare dedicated face-recognition models, appearance-based retrieval approaches, and VLM-based recognition and reranking baselines (Li et al., 2026; Bai et al., 2025; Wang et al., 2025) with the proposed framework under a common evaluation protocol. Our contributions are:

- **SoccerCanon**, a benchmark for canonical real-world soccer player identification from single challenging broadcast observations (1,108 instances, 424 identities, 213 matches, 99 teams, five competitions, three seasons).
- A **training-free expert-stitching framework** that maps specialized expert outputs into a shared canonical player space and combines them according to their query-specific reliability.
- **Controlled evaluations** against face-recognition, appearance, and VLM baselines, with cue- and component-level ablations and qualitative analyses.

2 Related Work

Reliability-aware multi-view fusion. Multi-view learning models each view's prediction as evidence and fuses views according to their estimated reliability (Han et al., 2021; Zhang et al., 2023). Previous work improves conflict handling and aggregation (Xu et al., 2024; Liu et al., 2024; Shi et al., 2026; Xu et al., 2025; Liang et al., 2025) and robustness under low-quality or open-world inputs (Liu et al., 2025; Jiang et al., 2026). Most of these methods assume that all views predict over a single shared label space and learn view-specific networks end-to-end. Our framework instead keeps the experts frozen, and their outputs live in heterogeneous label spaces. Before any evidence can be fused, these spaces must be *stitched*[1] into a common candidate set through structured ros-

[1]We use *stitching* for mapping expert outputs into a shared candidate space, not in the sense of model stitching between network layers (Lenc & Vedaldi, 2019; Bansal et al., 2021).

Table 1: **Comparison of soccer player identification settings and data.** Columns: the target is a persistent real-world player (not a track or action-local ID); queries are linked to identities from other matches; each query is one image; references are photographs from outside the broadcast; more than one identity cue is used per query; evaluation data, identities, and matches as reported in each paper (seq.: sequences, fr.: frames, q.: queries, trk.: tracklets, frag.: track fragments, obs.: observations); n/r: unclear from the paper.

	Identity Scope		Query	Reference	Evidence	Scale		
Study / Benchmark	**Real-World Identity**	**Cross-Match Linkage**	**Single Observation**	**Clean External Visual Ref.**	**Multi-Cue Evidence**	**Eval. data**	**Iden-tities**	**Mat-ches**
Bertini et al. (2006b)	✓	✓	×	×	n/r	78 seq.	5	4
Ballan et al. (2007)	✓	✓	n/r	×	×	∼2,000 fr.	20	10
Gerke et al. (2017)	✓	n/r	×	×	✓	14 shots	n/r	6
Comandur (2025)	×	×	✓	×	×	11,777 q.	n/r	n/r
Diop et al. (2022)	✓	×	✓	✓	n/r	n/r	22	1
Solberg et al. (2024)	✓	×	×	×	✓	511 trk.	n/r	n/r
A. Correia et al. (2025)	✓	×	✓	✓	×	1,500 fr.	n/r	8
Dusov et al. (2026)	n/r	×	×	×	✓	3,087 frag.	n/r	1
SoccerCanon (Ours)	✓	✓	✓	✓	✓	**1,108 obs.**	**424**	**213**



Figure 1: **Representative samples from SoccerCanon.** Each tile is a crop from one benchmark observation, illustrating the variation in player scale, blur, occlusion, and viewpoint across matches, competitions, and seasons.

ters. We therefore extend reliability-aware evidence fusion to this stitching setting without training, estimating each expert's reliability per query.

Player identification. Prior work on soccer player recognition exploits soccer-specific cues such as faces, jersey numbers, superimposed text, and spatial constellations (Bertini et al., 2006b; Ballan et al., 2007; Gerke et al., 2017; Diop et al., 2022). Other work uses external face references (A. Correia et al., 2025) or combines tracking, team appearance, OCR, and roster lookup (Solberg et al., 2024). Recent work studies same-player re-identification within a match (Comandur, 2025) and long-term identification from fragmented tracks (Dusov et al., 2026). These studies differ from SoccerCanon in their combination of identity scope, query unit, reference evidence, and scale (Table 1): they evaluate a few dozen players, a single match, or identities local to an action or a track. SoccerCanon jointly considers a single challenging observation, persistent canonical identity across 424 players and 213 matches, heterogeneous reference resources, and potentially incomplete visual evidence. General-purpose VLMs (Bai et al., 2025; Wang et al., 2025) and multimodal rerankers (Li et al., 2026) can instead verify a query against reference images or choose among candidate names; we evaluate both uses against an explicit composition of frozen experts.

3 SoccerCanon: Benchmark for Canonical Player Identification

We define the task (Section 3.1), describe the automatic and manual construction of **SoccerCanon** (Section 3.2; examples in Figure 1), and summarize its statistics, splits (Section 3.3), and reference resources (Section 3.4).

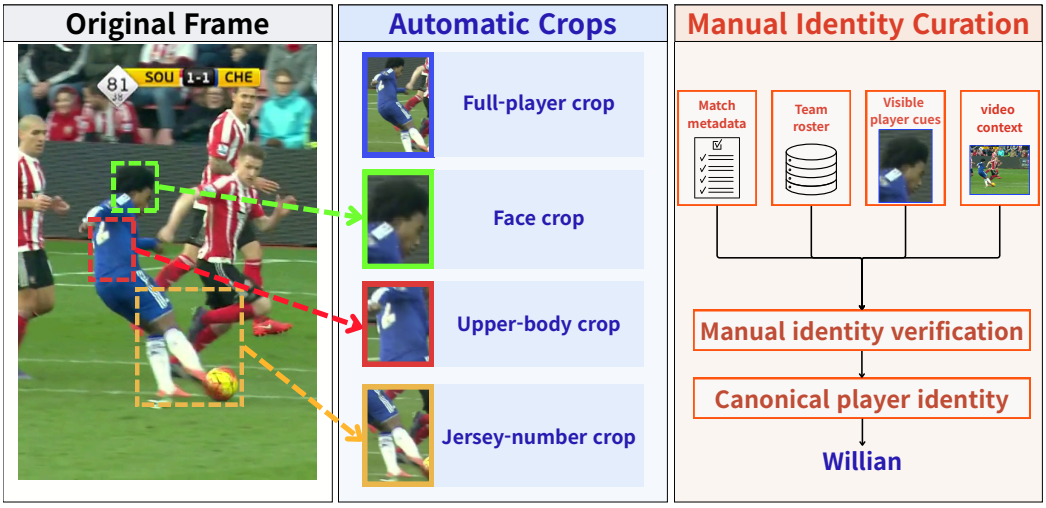


Figure 2: **Dataset construction pipeline of SoccerCanon.** Starting from a professional soccer broadcast frame, player regions are automatically localized and converted into a full-player crop together with face, upper-body, and jersey-number crops. A fixed, identity-agnostic selector chooses all test observations and most validation observations, while a 75-observation validation pilot predates the selector. Canonical player identities are then manually verified using match metadata, team–season rosters, visible player cues, and surrounding video context. Example broadcast imagery and derived crops are from SoccerNet (Deliege et al., 2021).

3.1 Benchmark Task and Scope

SoccerCanon evaluates *canonical real-world player identification*: given one localized player observation in professional broadcast imagery, predict the real-world player it depicts. The target is not a track identifier, an action-local identity, a team label, or a jersey number, but a canonical identity that remains consistent across actions, matches, and seasons. The benchmark does not evaluate detection or tracking, and it provides no temporally adjacent observations at inference time; face, upper-body, and jersey-number regions derived from an observation are complementary views of the same query.

3.2 Dataset Construction

The source images are the action and replay frames of SoccerNet (Deliege et al., 2021) released with SoccerNet-v3 (Cioppa et al., 2022), so observations preserve broadcast-specific variation in scale, viewpoint, blur, occlusion, and image quality; no synthetic image is introduced. As summarized in Figure 2, *automatic preprocessing* (Section 3.2.1) extracts and selects observations and their cue crops without using the ground-truth identity, and *manual curation* (Section 3.2.2) afterwards assigns the canonical identity.

3.2.1 Automatic Preprocessing

Players are localized with YOLOv8n-Pose (Jocher et al., 2023), and each retained box is enlarged by 10% on every side to form the full-player crop, from which all cue-specific regions are derived. The face crop is the largest face detected by the InsightFace SCRFD detector (Guo et al., 2021), with pose-keypoint and fixed-region fallbacks; the upper-body crop follows the shoulder and hip keypoints; and the jersey-number crop is the highest-scoring digit region found by EasyOCR (Kittinaradorn, 2020) over five body regions, with a fixed shorts region as fallback, so a cue-oriented input is always kept even when the cue itself is weak or absent (Appendix A.1). Full-player crops are then filtered by a fixed selector with no access to identity labels, lineups, or match metadata (Appendix A.2). It combines lenient structural and face-attachment checks, a CLIP-based (Radford et al., 2021) player-role check, and a DINOv2 (Oquab et al., 2024) few-shot check, and rejects the most degenerate cases (e.g., referees and spectators, rear views with no visible face, and tiny or severely blurred players). All 495 test and 538 validation observations were selected this way; the

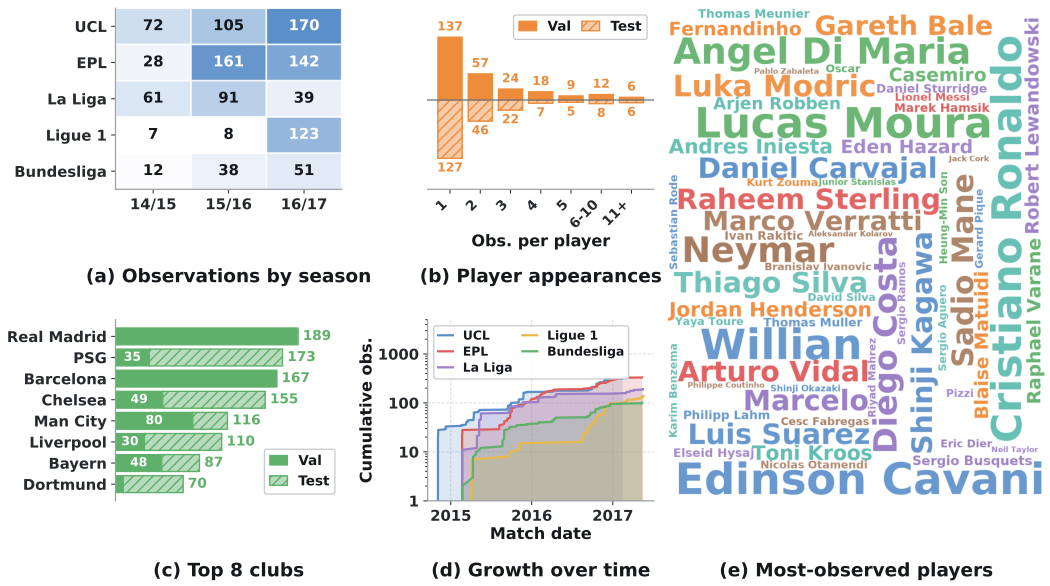


Figure 3: **Statistical overview of SoccerCanon.** (a) Observation coverage across competitions and seasons. (b) Distribution of the number of observations per player. (c) Number of validation and test observations from matches involving each of the eight most frequent clubs (an observation is counted for both participating clubs). (d) Cumulative number of observations over match date, per competition (log scale). (e) Most frequently observed player identities in the benchmark.

remaining 75 validation observations form a manually selected pilot subset collected before the selector was developed. Among the evaluated test faces, 94.7% are smaller than the recognizer's input resolution and must be upsampled (Appendix A.5).

3.2.2 Manual Identity Curation

The automatically generated crops are not used to assign ground-truth identities by themselves. Annotators first establish the fixture context from the participating teams, competition, match date, and team–season roster, using Flashscore (Flashscore, 2026) for fixture, lineup, and squad information. They then check the kit against the fixture archive, match jersey or shorts numbers to the lineup, recognize well-known players by face, and, when a frame remains ambiguous, inspect neighboring broadcast video. Observations whose identity cannot be verified are discarded: of the 3,599 selector-approved candidates from the test competitions, 495 (13.8%) were retained. One of the authors finally reviewed all retained labels (Appendix A.3).

3.3 Dataset Statistics and Splits

SoccerCanon contains **1,108 labeled player observations**, comprising **613 validation** and **495 test** instances. Overall, the benchmark covers **424 canonical player identities**, **213 distinct matches**, **99 teams**, **five professional competitions**, and **three seasons** spanning 2014–15 to 2016–17.

The validation split contains 263 unique player identities from 88 matches and 58 teams, whereas the test split contains 221 identities from 125 matches and 59 teams. Splits are match-disjoint and largely competition-disjoint: the validation split contains all UEFA Champions League (UCL; 347) and LaLiga (191) observations plus the 75-observation English Premier League (EPL) pilot subset, whereas the test split contains EPL (256), Ligue 1 (138), and Bundesliga (101) observations; all Ligue 1 test observations come from Paris Saint-Germain fixtures, as do all Ligue 1 source candidates. Because all hyperparameters are selected on the validation split, test evaluation also involves a competition-level shift. The two splits share 60 canonical identities but no matches.

Figure 3 summarizes the competition, temporal, player-frequency, and club-level composition of the benchmark; Appendix A.5 reports the competition-, season-, and split-level counts.

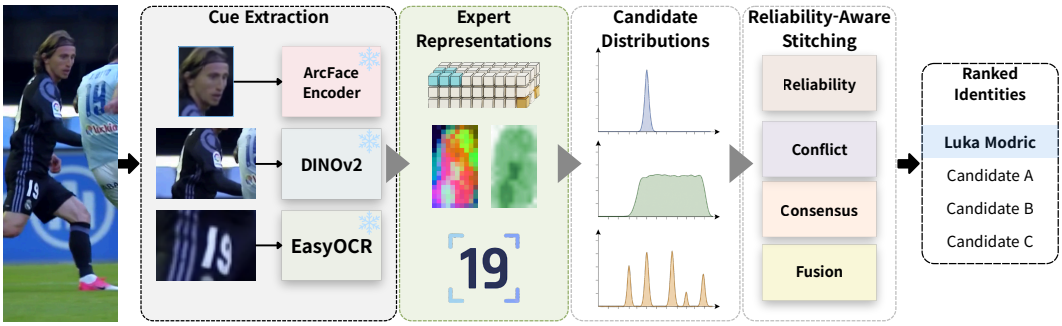


Figure 4: **Overview of the proposed expert-stitching framework.** Given a challenging broadcast player observation, three vision experts extract face, appearance, and number cues, encoded as expert representations by an ArcFace (Deng et al., 2019) encoder, DINOv2-based (Oquab et al., 2024) kit matching, and EasyOCR (Kittinaradorn, 2020). Each cue is converted into a candidate distribution over a shared player candidate set, where the appearance and number cues are projected through structured team–season rosters. Reliability-aware stitching then estimates the reliability of each cue, discounts conflicting cues, adds face–structured consensus, and fuses all evidence into the final ranking over candidate players (right).

3.4 Reference Resources

Methods may use three fixed reference resources, all built independently of the benchmark frames and shared by every query (Appendix A.4). *Face gallery*: 9,366 portrait photographs of 7,074 players from SoccerWiki (Rao et al., 2025) (Flashscore and Wikipedia images); the ground-truth player is present for 454 of the 489 evaluated test observations. *Kit references*: 208 clean images, one per team–season kit (home, away, or third). *Rosters*: 75 Wikipedia team–season squad lists (2,591 entries, 1,905 players, with jersey numbers) of the 48 clubs from the three test competitions, covering 72 of the 92 team–seasons in the test fixtures. None of them encodes which match or team a query comes from.

4 Method

4.1 Overview

Our goal is to resolve a broadcast observation q to a canonical player identity without forcing all evidence through a single recognition pathway. Following Figure 4, **cue extraction** (Section 4.2) turns q into expert representations z_F, z_A, z_N; **candidate distributions** (Section 4.3) express them over a shared candidate set $\mathcal{C}$ as P_F, P_A, P_N, and P_{AN}; and **reliability-aware stitching** (Section 4.4) fuses these into the final ranking over $\mathcal{C}$.

4.2 Cue Extraction and Expert Representations

Face expert. The face expert provides the most direct player-level identity evidence. For each available player or face crop, we use the InsightFace SCRFD detector (Guo et al., 2021) to re-detect the face, estimate five facial landmarks, and align the face to a canonical 112×112 crop. The aligned face is encoded by the frozen ArcFace (Deng et al., 2019) recognizer of InsightFace (`buffalo_l`, a ResNet-50 (He et al., 2016) trained on WebFace600K (Zhu et al., 2021)). To improve robustness to clutter and imperfect localization, we encode three views of the same aligned face: the original view, a DINOv2-guided soft-masked view (Oquab et al., 2024), and a horizontally flipped view. The view embeddings are combined with consistency-dependent weights, so that an auxiliary view contributes less when it substantially changes the original identity embedding. When several crops are available, their embeddings are further pooled by face quality and cross-crop consistency into the face embedding z_F.

Appearance (kit) expert. The appearance expert describes *team–season* kit appearance rather than player identity. It builds several upper-body views of the query and suppresses skin and back-

ground regions after color normalization. Each view is described by color statistics, DINOv2 local and global features (Oquab et al., 2024), and pattern and layout descriptors; the set of these view descriptors forms the appearance representation z_A.

Number expert. The number expert applies EasyOCR (Kittinaradorn, 2020) to multiple number-oriented crops and deterministic preprocessing variants. Its representation z_N is the set of scored readings and their supporting crops. Within the same crop, preprocessing variants are treated as alternative measurements rather than independent evidence, whereas support across independent crops is explicitly rewarded.

4.3 Candidate Distributions

All cues are expressed over one shared candidate set $\mathcal{C} = \mathcal{C}_{\text{roster}} \cup \mathcal{C}_{\text{face-only}}$: all 1,905 players of the 75 team–season rosters, together with those of the top-100 face candidates that appear in no roster ($\bar{K} = 1{,}983$ candidates per query). No match, team, or league metadata of the query is used.

Face. P_F is a softmax over the gallery scores of the top-100 identities for z_F; all other candidates receive zero face mass.

Appearance and number through rosters. z_A is compared with the 208 clean team–season kit references (Section 3.4) using color, local-feature, global-feature, pattern, and layout similarities whose weights adapt to query quality, which yields a team–season distribution $P_A(t, s)$. z_N yields a jersey-number distribution P_N^{num}, in which an *unknown* state absorbs the remaining mass when no stable number is read. Neither identifies players, so both are projected through the rosters. With $\mathcal{R}_i$ the roster entries (t, s) of player c_i, $n_{i,t,s}$ its number there, and $d_{t,s,n}$ the number of players of roster (t, s) wearing n,

$$P_A(c_i) \propto \max_{(t,s)\in\mathcal{R}_i} \frac{P_A(t,s)}{|\mathcal{R}_{t,s}|}, \qquad P_{AN}(c_i) \propto \max_{(t,s)\in\mathcal{R}_i} \frac{P_A(t,s)\, P_N^{\text{num}}(n_{i,t,s})}{d_{t,s,n_{i,t,s}}}. \tag{1}$$

P_N divides the mass of each number equally among the roster players who wear it; a player in several rosters keeps only its strongest entry.

4.4 Reliability-Aware Stitching

Reliability. Each cue $m \in \{F, A, N\}$ is assigned a query-specific reliability $\rho_m \in [0, 1]$. The face cue combines the concentration of P_F with a face-quality score that combines detection confidence, face size, sharpness, head pose, and multi-view consistency. The appearance cue uses the concentration of its team–season distribution, adjusted by how well its team–season labels match the available rosters. The number cue uses a number-quality estimate that is capped by the number of independent crops supporting the top reading and their agreement, and scaled by the sharpness and top-rank margin of P_N^{num}. When the appearance or number cue is reliable, face-only candidates outside every roster are softly down-weighted, but never removed.

Conflict. Two cues conflict only when their candidate distributions differ *and* share no strongly supported candidate; the less reliable cue is then discounted to $\widetilde{\rho}_m \le \rho_m$, down to a non-zero floor. The reliability ρ_{AN} follows from $\widetilde{\rho}_A$, $\widetilde{\rho}_N$, and the compatibility of P_A and P_N.

Consensus. A face–structured consensus P_{cons} with reliability ρ_{cons} rewards candidates supported by both P_F and the roster-based P_A and P_{AN}, without removing one-sided candidates.

Fusion. All channels are accumulated as Dirichlet evidence over $\mathcal{C}$, with $K = |\mathcal{C}|$ and fixed cue weights λ:

$$\boldsymbol{\alpha} = \alpha_0 \mathbf{1} + K \sum\nolimits_{m\in\{F,A,N\}} \lambda_m \widetilde{\rho}_m P_m + K\lambda_{AN}\rho_{AN} P_{AN} + K\lambda_{\text{cons}}\rho_{\text{cons}} P_{\text{cons}}. \tag{2}$$

Candidates are ranked by $0.4\,\hat{s}_{\text{Dir}} + 0.6\,\hat{s}_{\text{RRF}}$, the max-normalized posterior $s_{\text{Dir}}(c_i) = \alpha_i / \sum_j \alpha_j$ and reciprocal rank fusion (Cormack et al., 2009) ($k = 60$) of the posterior and consensus rankings, with 0.4 selected on the validation split. Appendix D gives all equations and constants.

Table 2: **SoccerCanon is hard for existing recognizers.** (a) R@1 of the same frozen models with the same 7,074-identity gallery when the query is a clean face-gallery portrait (leave-one-out, 3,200 queries of 908 players) or a SoccerCanon test observation of a gallery player (454 queries). (b) R@1 on FA-489 by condition for the face cue alone and our method (full candidate set); *read*: our number expert reads the correct squad number (444 queries with a known team–season).

(a) Clean vs. broadcast query

Model	Clean	SoccerCanon
ArcFace (Deng et al., 2019)	95.97	40.53
AdaFace (Kim et al., 2022)	96.09	41.85
KP-RPE (Kim et al., 2024)	96.00	47.14
CLIP (Radford et al., 2021)	9.81	6.17
DINOv2 (Oquab et al., 2024)	1.38	0.44
OSNet (Zhou et al., 2019)	1.06	0.22
TransReID (He et al., 2021)	2.56	2.42

(b) R@1 by condition

Condition		N	Face	Ours
Face size	$<$32 px	91	28.57	37.36
	32–47 px	245	52.65	60.00
	$\geq$48 px	153	71.90	76.47
Number	read	58	41.38	58.62
	not read	386	58.03	63.21
League	EPL	252	48.02	54.37
	Ligue 1	136	58.82	67.65
	Bundesliga	101	63.37	68.32

5 Experiments

5.1 Experimental Setup

Evaluation population. For six of the 495 test observations our face expert detects no face (Appendix A.2), so all methods are evaluated on the 489-observation *face-available* subset (FA-489); counting the six as failures for ours and as correct for every VLM still gives 60.20 versus at most 30.10% R@1 (Appendix B.1).

Protocol and metrics. All experts are frozen, all stitching parameters are fixed on the validation split, and we report R@1, R@5, R@10, and mean reciprocal rank (MRR); Table 8 lists each method's candidate space.

5.2 Benchmark Analysis

Table 2 shows that existing recognizers do not solve SoccerCanon, that the difficulty lies in the broadcast query rather than in the gallery or the models, and that it concentrates on small faces, where complementary cues matter most.

Clean versus broadcast identification. We query the same frozen models and gallery with the face-gallery portraits of Section 3.4 (leave-one-out over the 908 players with at least two). Face recognizers identify these portraits almost perfectly (about 96% R@1) but reach only 40–47% on the 454 SoccerCanon observations whose player is in the gallery (Table 2a). The two query sets differ, and photographs of a player may share a source session, so the clean setting is slightly optimistic (Appendix C.1). Generic visual encoders and re-ID models fail in both settings, since they are not trained to discriminate thousands of player identities.

What makes SoccerCanon hard. Test faces are small (median 42 px; 18.6% below 32 px), and the face cue alone falls to 28.57% R@1 below 32 px (Table 2b). The number cue is rarely available: our number expert reads the correct squad number for only 58 of 444 queries (13.1%). Yet these are the queries where the face is weakest (41.38 vs. 58.03% R@1 when the number is not read), likely because a readable number usually comes with a side or rear view, and combining cues raises R@1 on them to 58.62%: the cues are complementary rather than redundant. EPL is the hardest competition; the Ligue 1 subset, all Paris Saint-Germain fixtures, is not easier than the Bundesliga (67.65 vs. 68.32% R@1), so club concentration does not inflate the results (Appendix C.1).

5.3 Comparison with Existing Model Families

General-purpose VLMs used as identity verifiers reach at most 29.24% R@1 although they rank only 100 match-restricted candidates, and scaling Qwen3-VL from 8B to 32B does not help (Table 3).

Table 3: **Comparison with existing model families on FA-489.** VLMs rank 100 candidates per query drawn from the two teams' season squads (ground truth reachable for 442/489); face recognition and re-ID models rank the full 7,074-identity gallery (454/489); our method ranks roster players and face candidates without match metadata (485/489). Appendix B.4 evaluates all methods on identical candidates.

Family	Method	R@1	R@5	R@10	MRR
VLM	Qwen3-VL-Reranker-8B (Li et al., 2026)	29.24	58.49	69.53	0.423
	Qwen3-VL-8B (Bai et al., 2025)	25.36	54.81	65.64	0.386
	Qwen3-VL-32B (Bai et al., 2025)	25.36	53.58	65.44	0.383
	InternVL3.5-8B (Wang et al., 2025)	14.31	36.40	49.08	0.257
Face recognition	ArcFace (Deng et al., 2019)	37.63	43.76	46.01	0.406
	AdaFace (Kim et al., 2022)	38.85	45.81	47.24	0.420
	KP-RPE (Kim et al., 2024)	43.76	50.51	53.17	0.471
Re-ID / encoders	OSNet (Zhou et al., 2019)	0.20	0.41	1.02	0.009
	TransReID (He et al., 2021)	2.25	6.95	8.59	0.048
	CLIP (Radford et al., 2021)	5.73	13.91	21.27	0.104
	DINOv2 (Oquab et al., 2024)	0.41	1.02	1.84	0.013
Ours (stitching)		**60.94**	**72.60**	**76.28**	**0.663**

Table 4: **How the cues are combined matters more than which cues are used.** Results on FA-489 with the same frozen expert outputs and candidate sets. *Correct*: queries ranked correctly at rank 1; Δ: R@1 change relative to the face cue alone.

	Cues used									
Method	**Face**	**App.**	**Num.**	**How cues are combined**	**Correct**	**R@1**	**R@5**	**R@10**	**MRR**	**Δ R@1**
Face only	✓				265	54.19	64.83	68.92	0.591	
Equal weights	✓	✓	✓	Same weight for every cue	246	50.31	58.69	61.76	0.546	−3.89
Fixed weights	✓	✓	✓	One preset weight per cue, same for all queries	270	55.21	66.26	70.14	0.603	+1.02
Ours	✓	✓	✓	**Reliability-aware aggregation + rank fusion** (Sec. 4.4)	**298**	**60.94**	**72.60**	**76.28**	**0.663**	**+6.75**

Face recognizers reach 37.6–43.8%, and re-ID models and visual encoders stay below 6%. Our framework reaches 60.94% without match metadata, 61.35% when restricted to gallery identities (Appendix B.6), and keeps the lead when every method ranks the same candidates (Appendix B.4).

5.4 Effect of Reliability-Aware Stitching

As shown in Table 4, reliability-aware stitching raises R@1 from 54.19 to 60.94 (paired bootstrap 95% CI of the gain: $[4.29, 9.20]$), correcting 36 observations while breaking only three relative to the face cue alone (exact McNemar $p < 10^{-7}$). The gains concentrate where the face cue is unreliable: 32 of the 33 net corrections fall in the two lower terciles of face reliability, whereas R@1 barely changes in the top tercile (Appendix C.2), and the largest gains occur on faces below 32 px and when the number is read (Table 2b). Most of the gain comes from face–structured consensus and rank-level fusion; reliability and conflict discounting alone do not improve R@1 ($270 \rightarrow 269$ correct; Appendix C.2).

6 Conclusion

We introduced **SoccerCanon**, a benchmark for identifying canonical real-world soccer players from single broadcast observations. Face recognizers that are nearly perfect on clean photographs lose more than half of their accuracy on it, most severely on small faces, and general-purpose VLMs used as identity verifiers stay below 30% R@1. We further proposed a training-free framework that stitches face, appearance, and number experts according to their query-specific reliability; it outperforms face recognition, re-ID, and VLM baselines, with the largest gains where the face cue is unreliable. We hope SoccerCanon encourages recognition methods that weigh heterogeneous and unreliable evidence rather than relying on a single visual pathway.

AI USE STATEMENT

In this work, we used generative AI tools to assist with English-language editing, manuscript organization, LaTeX formatting, code drafting and debugging, internal consistency checking, and feedback on the manuscript, including suggestions for additional analyses. The block illustrating embeddings in the "Expert Representations" panel of Figure 4 was generated with a generative AI tool; it is illustrative only and does not depict data or results. All identity labels were assigned and verified manually. A brief trial of AI-assisted labeling on part of the validation split was abandoned because of its poor accuracy, and none of its outputs was used for any benchmark label. Generative AI tools were not used to modify benchmark labels or to select experimental results.

All AI-assisted text, code, analyses, and manuscript revisions were reviewed by the authors. Experimental results reported in the paper were obtained from our evaluation pipelines and were checked against the corresponding outputs. The authors take full responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

SoccerCanon associates broadcast images of professional soccer players with their real-world identities, which is a form of face and person identification. All depicted individuals are professional athletes appearing in publicly broadcast matches, and identity labels refer only to their public professional role; no private attributes are annotated. The benchmark is intended for research on robust fine-grained recognition in sports analytics, and we discourage its use for surveillance or for identifying non-public individuals. Broadcast frames and videos were obtained from SoccerNet under its non-disclosure agreement, and we do not redistribute them; the released annotations will consist of identity labels, SoccerNet frame identifiers, and crop coordinates, so that users who obtain SoccerNet access under its terms can reconstruct the benchmark. Reference photographs from Flashscore and Wikipedia and kit images from Football Kit Archive remain subject to their respective terms and are not redistributed. The pretrained InsightFace `buffalo_l` models used by our face expert are licensed for non-commercial research only, and we use them accordingly. No human-subject study was conducted; identity annotation was performed by two undergraduate research assistants of our group, followed by a final manual review of all retained labels by one of the authors.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe every benchmark-construction stage, including all automatic preprocessing and selection thresholds (Tables 5 and 6), the annotation procedure, and the split composition (Table 7). Section 4 and Appendix D give the full scoring rule of our method together with its constants and how each was set (during development, as a standard value, or by a validation sweep; Table 21). Appendix B specifies, for every baseline, the model weights, inputs, candidate construction, prompts, and ranking rule, and Table 8 summarizes the candidate space and ground-truth coverage of each method. All reported numbers are computed from frozen per-query rankings produced before test labels are read. We will release the code, the frozen per-query rankings, and the benchmark annotations upon publication.

REFERENCES

Helena A. Correia, Diogo Pontes, and José Henrique Brito. Face recognition in online soccer streaming for piracy detection. *Multimedia Tools and Applications*, 84(26):30737–30756, 2025. doi: 10.1007/s11042-024-20389-3. URL https://doi.org/10.1007/s11042-024-20389-3.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

L. Ballan, M. Bertini, A. Del Bimbo, and W. Nunziati. Soccer players identification based on visual local features. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, CIVR '07, pp. 258–265, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937339. doi: 10.1145/1282280.1282321. URL https://doi.org/10.1145/1282280.1282321.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Marco Bertini, Alberto Del Bimbo, and Walter Nunziati. Matching faces with textual cues in soccer videos. In *2006 IEEE International Conference on Multimedia and Expo*, pp. 537–540, 2006a. doi: 10.1109/ICME.2006.262444.

Marco Bertini, Alberto Del Bimbo, and Walter Nunziati. Automatic detection of player's identity in soccer videos using faces and text cues. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, pp. 663–666, New York, NY, USA, 2006b. Association for Computing Machinery. ISBN 1595934472. doi: 10.1145/1180639.1180778. URL https://doi.org/10.1145/1180639.1180778.

Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up SoccerNet with multi-view spatial localization and re-identification. *Scientific Data*, 9(1):355, 2022. doi: 10.1038/s41597-022-01469-1. URL https://doi.org/10.1038/s41597-022-01469-1.

Bharath Comandur. Sports re-id: Improving re-identification of players in broadcast videos of team sports, 2025. URL https://arxiv.org/abs/2206.02373.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pp. 758–759, New York, NY, USA, 2009. Association for Computing Machinery. ISBN

9781605584836. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. doi: 10.52202/075280-2142. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4508–4519, June 2021.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Charles-Alexandre Diop, Baptiste Pelloux, Xinrui Yu, Won-Jae Yi, and Jafar Saniie. Soccer player recognition using artificial intelligence and computer vision. In *2022 IEEE International Conference on Electro Information Technology (eIT)*, pp. 477–481, 2022. doi: 10.1109/eIT53891.2022.9813788.

Murad Dusov, Vasiliy Chelpanov, Andrey Sakhovskiy, Vadim Linkov, Oleg Durygin, Mikhail Moiseev, Matvey Isupov, Konstantin Mitin, and Semen Budennyy. Ltpi: A benchmark for long-term player identification from single-camera football video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 9919–9929, June 2026.

Flashscore. Flashscore: Live scores, results, and sports statistics. https://www.flashscore.com/, 2026. Accessed: 2026-09-22.

Football Kit Archive. Football kit archive. https://www.footballkitarchive.com/, 2026. Accessed: 2026-09-22.

Sebastian Gerke, Antje Linnemann, and Karsten Müller. Soccer player recognition using spatial constellation features and jersey number recognition. *Computer Vision and Image Understanding*, 159:105–115, 2017. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2017.04.010. URL https://www.sciencedirect.com/science/article/pii/S1077314217300723. Computer Vision in Sports.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection, 2021. URL https://arxiv.org/abs/2105.04714.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification, 2021. URL https://arxiv.org/abs/2102.02051.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Los Alamitos, CA, USA, June 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.90. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based Object Re-Identification . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14993–15002, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.01474. URL https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01474.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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, 2026. URL https://arxiv.org/abs/2605.03820.

Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. URL https://github.com/ultralytics/ultralytics. Software, version 11.0.0.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL https://github.com/ultralytics/ultralytics. Software, version 8.0.0.

Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18729–18738, 2022. doi: 10.1109/CVPR52688.2022.01819.

Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. Keypoint relative position encoding for face recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 244–255, 2024. doi: 10.1109/CVPR52733.2024.00031.

Rakpong Kittinaradorn. Easyocr, 2020. URL https://github.com/JaidedAI/EasyOCR. Open-source optical character recognition software.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *Int. J. Comput. Vision*, 127(5):456–476, May 2019. ISSN 0920-5691. doi: 10.1007/s11263-018-1098-y. URL https://doi.org/10.1007/s11263-018-1098-y.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720.

Xinyan Liang, Shijie Wang, Yuhua Qian, Qian Guo, Liang Du, Bingbing Jiang, Tingjin Luo, and Feijiang Li. Trusted multi-view classification with expert knowledge constraints. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37409–37426. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liang25p.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Wei Liu, Yufei Chen, and Xiaodong Yue. Enhancing multi-view classification reliability with adaptive rejection. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i18.34088. URL https://doi.org/10.1609/aaai.v39i18.34088.

Ying Liu, Lihong Liu, Cai Xu, Xiangyu Song, Ziyu Guan, and Wei Zhao. Dynamic evidence decoupling for trusted multi-view learning, 2024. URL https://arxiv.org/abs/2410.03796.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Jiayuan Rao, Zifeng Li, Haoning Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-agent system for comprehensive soccer understanding. In *ACM Multimedia 2025*, 2025.

Long Shi, Chuanqing Tang, Huangyi Deng, Cai Xu, Lei Xing, and Badong Chen. Generalized trusted multi-view classification framework with hierarchical opinion aggregation. *IEEE Transactions on Multimedia*, 28:3396–3408, 2026. doi: 10.1109/TMM.2026.3651098.

Hakon Solberg, Mehdi H. Sarkhoosh, Sushant Gautam, Saeed S. Sabet, Pal Halvorsen, and Cise Midoglu. Playertv: Advanced player tracking and identification for automatic soccer highlight clips. In *2024 International Symposium on Multimedia (ISM)*, pp. 93–97. IEEE, December 2024. doi: 10.1109/ism63611.2024.00019. URL http://dx.doi.org/10.1109/ISM63611.2024.00019.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Wikipedia contributors. Wikipedia, the free encyclopedia. https://www.wikipedia.org/, 2026. Accessed: 2026-09-22.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning, 2024. URL https://arxiv.org/abs/2402.16897.

Cai Xu, Ziqi Wen, Jie Zhao, Wanqing Zhao, Jinlong Yu, Haishun Chen, Ziyu Guan, and Wei Zhao. Beyond equal views: Strength-adaptive evidential multi-view learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 1278–1287, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755092. URL https://doi.org/10.1145/3746027.3755092.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3701–3711, 2019. doi: 10.1109/ICCV.2019.00380.

Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/50623630a2372839c078474efa6c0cb8-Paper-Conference.pdf.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10487–10497, 2021. doi: 10.1109/CVPR46437.2021.01035.

APPENDIX CONTENTS

A SoccerCanon Construction and Statistics **17**

A.1 Automatic Preprocessing 17

A.2 Automatic Observation Selection 17

A.3 Identity Annotation and Quality Control 18

A.4 Reference Resources 18

A.5 Statistics and Cue Availability 19

B Baselines and Candidate-Controlled Comparison **19**

B.1 Evaluation Protocol 19

B.2 Face Recognition, Re-ID, and Visual Encoder Baselines 20

B.3 VLM Baselines 21

B.4 Candidate-Controlled Comparison 21

B.5 MiniCPM-V-4.6 with LoRA 22

B.6 Our Method on the Baseline Gallery 22

C Additional Results **22**

C.1 Benchmark Analysis Details 22

C.2 Ablations 23

D Method Details **25**

D.1 Problem Formulation and Notation 25

D.2 Candidate Space Construction 25

D.3 Face Expert 26

D.3.1 Multi-View Face Representation 26

D.3.2 Face Quality 26

D.3.3 Face Retrieval 26

D.4 Appearance Expert 26

D.5 Number Expert 27

D.6 Projection into a Shared Player Space 27

D.7 Reliability 27

D.8 Conflict 28

D.9 Consensus 28

D.10 Dirichlet Fusion 28

D.11 Final Ranking 29

E Qualitative Analysis **29**

Table 5: **Automatic preprocessing and selection**, with the fallback used when a cue is unavailable.

Stage	Primary method	Fallback	Output
Player localization	YOLOv8n-Pose (Jocher et al., 2023) (conf. ≥ 0.25)	None; boxes $< 0.3\%$ of frame area discarded	Full-player crop (top-6, 10% enlargement)
Face localization	SCRFD (Guo et al., 2021) (largest face)	Pose head keypoints $\rightarrow$ fixed upper region	Face crop
Appearance localization	Shoulder/hip keypoints	Fixed torso region	Upper-body crop
Number localization	EasyOCR (Kittinaradorn, 2020) (digits; original + enhanced)	Fixed lower-right (shorts) region	Jersey-number crop
Observation selection	Structural, face, CLIP, and DINOv2 checks (Table 6)	Rejected (not annotated)	Selected observation

A SoccerCanon Construction and Statistics

A.1 Automatic Preprocessing

The input frames are the action and replay frames of SoccerNet (Deliege et al., 2021) released with SoccerNet-v3 (Cioppa et al., 2022), covering the English Premier League, LaLiga, Ligue 1, Bundesliga, and UEFA Champions League from 2014–15 to 2016–17; the corresponding match videos are kept only to resolve ambiguous identities during annotation. Players are localized with YOLOv8n-Pose (Jocher et al., 2023) (confidence $\geq$ 0.25); boxes smaller than 0.3% of the frame are discarded, up to six of the largest remaining boxes are kept per frame, and each is enlarged by 10% on every side to form the full-player crop. The pose model is re-applied to this crop, and keypoints with confidence ≥ 0.2 from the largest person are used below. The face crop is the largest face found by the SCRFD detector (Guo et al., 2021) of InsightFace (`buffalo_l`, 640×640 input), enlarged by 30%. If no face is found, a square head region of side $\max(0.55\, w_s, 40\,\text{px})$ (w_s: shoulder width; without shoulders, $\max(0.22W, 0.16H, 40\,\text{px})$) is centered on the nose, eye, and ear keypoints and enlarged by 20%; without these keypoints, the region $x \in [0.30W, 0.70W]$, $y \in [0.02H, 0.28H]$ of the $W \times H$ player crop is used. The upper-body crop spans the shoulder and hip keypoints (extrapolated downward from the shoulders when the hips are missing), with the fallback $x \in [0.15W, 0.85W]$, $y \in [0.18H, 0.65H]$. For the jersey-number crop, EasyOCR (Kittinaradorn, 2020) restricted to digits is run on five regions (upper jersey, chest, lower body, and its left and right halves), each in its original form and after grayscale conversion, $2\times$ bicubic upsampling, and histogram equalization. A one- or two-digit reading of length ℓ is scored by its confidence times $1 + 0.2\min(\ell, 2)$; the best box is enlarged by 65%, and a fixed lower-right (shorts) region is used when no digit is found. A number crop is therefore always produced, but a readable number is never required. Table 5 summarizes the stages.

A.2 Automatic Observation Selection

The benchmark was built in two phases. Before the selector was developed, 75 EPL validation observations were chosen by hand from the same automatically generated crop pool; all other observations, namely the 538 UEFA Champions League and LaLiga validation observations and *all* 495 test observations, were chosen by the selector with one fixed configuration (Table 6). Hence 1,033 of the 1,108 observations (93.2%), including the entire test split, were chosen automatically. The selector sees only the image, with no identity label, lineup, or match metadata, and later manual annotation could only *remove* observations; no selector output was replaced and no image was added to the test split by hand. It approved 3,599 candidates from the three test competitions (1,453 EPL, 1,205 Ligue 1, and 941 Bundesliga); the total number of crops passed to it was not recorded.

Checks. The structural check requires the target to occupy enough of the crop with at least one shoulder or hip keypoint, and rejects crops in which another person is more than $1.35\times$ larger than the target. A face proposed by SCRFD at score 0.20 must lie in the upper part of the person box, agree with the head keypoints, and pass a minimum score of 0.25 and lenient size, sharpness (Laplacian variance), and eye-distance checks. The CLIP check contrasts three positive prompts ("a professional soccer player on the pitch wearing a team uniform", "a full body football player during a competitive match", "a soccer player with a visible face and jersey number on the field") with seven negative prompts (referee, coach or staff, spectator, crowd, a blurry tiny distant player, a back view with no visible face, and overlapping players with an occluded face), using the sigmoid of the difference of the log-mean-exp positive and negative logits. The DINOv2 check compares the concatenated CLS and mean patch tokens with three positive and five negative exemplars (mean of the top-two cosine similarities), with thresholds calibrated from leave-one-out statistics of the exemplars (5th percentile minus 0.15).

Table 6: **Selector parameters**, used unchanged for all automatically selected observations. Person and pose evidence come from YOLO11n-Pose (Jocher & Qiu, 2024), face evidence from SCRFD (Guo et al., 2021), semantic filtering from CLIP ViT-B/32 (Radford et al., 2021), and few-shot filtering from DINOv2-small (Oquab et al., 2024).

Parameter	Value
Minimum image side	160 px
Person confidence / min. height ratio / min. area ratio	0.25 / 0.35 / 0.06
Visible-keypoint threshold / min. visible keypoints	0.15 / 3
Maximum other-to-target person area ratio	1.35
Face-detector threshold / minimum face score	0.20 / 0.25
Minimum face side / face-area ratio	16 px / 0.0006
Minimum face blur score / normalized eye distance	5.0 / 0.03
CLIP player-role threshold	0.38
DINOv2 min. positive similarity / pos.–neg. margin	0.615 / -0.057
Jersey-number gate	Disabled

Face availability. Every selector-approved observation contains an attached face that is detectable at this lenient threshold. Because the selector's face detector is more permissive than the detectors used downstream, some of these faces are not re-detected either in the crop-generation stage (SCRFD default threshold 0.5 at 640×640) or by the face-recognition pipeline (threshold 0.25 at 1024×1024); for such observations the face cue is effectively unavailable to recognition models. In the final benchmark this affects 8 of the 1,108 observations (2 validation and 6 test observations); for 5 further test observations our face expert falls back to a center crop of the face crop, and these remain in the evaluation. The 8 observations are retained in the released benchmark, but because our pipeline and the face-recognition baselines require a re-detected face, all reported test results use the remaining 489-observation face-available subset (FA-489); Appendix B.1 bounds the results on all 495 test observations.

A.3 Identity Annotation and Quality Control

Canonical identities were assigned by two undergraduate research assistants of our group. For each observation, they identified the fixture and consulted its archived Flashscore (Flashscore, 2026) lineup, which restricts the plausible identities but never determines them alone. The team is established from the kit, checked against the kit worn in that fixture according to match photographs from reports and news articles (home, away, or third kit); the kit references of our appearance expert carry no such fixture information. The identity is then established by at least one of three routes: (i) a visible shorts or jersey number, checked first and matched to the lineup; (ii) the face, for well-known players; or (iii) the SoccerNet broadcast video around the frame, when the static frame is ambiguous. Hairstyle, body appearance, and other characteristics served only as supporting cues checked against the lineup and never established an identity on their own. Observations whose identity could not be established by any route were discarded rather than assigned an uncertain label. Across the test competitions, the selector approved 1,453 EPL, 1,205 Ligue 1, and 941 Bundesliga candidates from 77, 37, and 37 fixtures, of which 256 (17.6%), 138 (11.5%), and 101 (10.7%) were retained from 65, 32, and 28 fixtures; all 1,205 Ligue 1 candidates already come from Paris Saint-Germain fixtures, so the Ligue 1 composition reflects the source matches rather than annotation.

After the initial annotation stage, one of the authors performed a final manual review of all retained observations and their canonical identity labels, re-checking each identity against the visual evidence and match-level information, including fixture lineups, match-specific kit evidence, and surrounding broadcast video when necessary. Observations whose identities could not be maintained with sufficient confidence were excluded, and three players that had been labeled under two spellings each (e.g., "Messi" and "Lionel Messi") were merged into a single canonical identity.

A.4 Reference Resources

All reference resources are built independently of the benchmark frames, frozen before test evaluation, and shared by every query.

Table 7: **Composition of SoccerCanon.** Observations per competition and split; matches, teams, and identities per split.

	Validation	Test	Total
UEFA Champions League	347	–	347
English Premier League	75	256	331
LaLiga	191	–	191
Ligue 1	–	138	138
Bundesliga	–	101	101
Observations	613	495	1,108
Matches	88	125	213
Teams	58	59	99
Identities	263	221	424

Face gallery. The gallery is a subset of SoccerWiki (Rao et al., 2025), which collects player photographs from Flashscore (Flashscore, 2026) and Wikipedia (Wikipedia contributors, 2026). A verified base cache contains 7,001 players with 9,185 images (one Flashscore image per player and 2,184 Wikipedia images). For the 874 SoccerWiki players whose Flashscore image is missing (a list determined by SoccerWiki, not by benchmark labels), Wikipedia images are added when at least two of them form a consistent ArcFace cluster (pairwise cosine similarity ≥ 0.36 and centroid similarity ≥ 0.40, at most five images per player), which adds 73 players and 181 images. The final gallery contains 9,366 portrait photographs of 7,074 players and is used by the face and re-ID baselines, the VLM baselines (through the Fixed-100 references), and our face expert; the MiniCPM baseline uses a larger SoccerWiki CLIP index (Appendix B.5).

Kit references and rosters. The appearance expert compares every query with 208 clean images from Football Kit Archive (Football Kit Archive, 2026), one per team–season kit type. Our method uses 75 team–season squad lists from the Wikipedia (Wikipedia contributors, 2026) season articles (2,591 entries covering 1,905 players, with jersey numbers); they are historical season squads rather than matchday lineups. The database contains the 48 clubs of the three test competitions, covers 72 of the 92 team–seasons in the test fixtures, and contains no LaLiga club; the ground-truth player is listed in at least one roster for 478 of the 489 FA-489 queries. The same database is loaded for every query, and no match, team, or league metadata of the query is used to select a subset.

A.5 Statistics and Cue Availability

The splits are match-disjoint and share 60 identities (Table 7); they were assigned by competition and collection phase rather than at random, so test evaluation also measures transfer across competitions. The 1,108 observations come from the 2014–15 (180), 2015–16 (403), and 2016–17 (525) seasons. Identity frequency is long-tailed: of the 424 identities, 207 occur once, 181 two to five times, 20 six to ten times, and 16 more than ten times (median 2, maximum 23).

Cue quality varies widely. On the 489 FA-489 test observations, 463 faces (94.7%) are smaller than the 112×112 recognition input and must be upsampled, 5 come only from a forced center crop, and 21 are used at native resolution; for 383 observations only one of the player and face crops yields a usable face. The median shorter side of the face is 42 px, and 91 faces (18.6%) are below 32 px. The calibrated number reliability of our number expert (Appendix D) is below 0.1 for 332 test observations (67.9%) and at least 0.3 for only 96 (19.6%).

B Baselines and Candidate-Controlled Comparison

B.1 Evaluation Protocol

All methods are evaluated on the same 489 face-available test observations (252 EPL, 136 Ligue 1, and 101 Bundesliga). Let r_i be the rank of the ground truth for query i, with $r_i = 0$ if it is not in the ranked list; then $\mathrm{R@}k = \frac{1}{N}\sum_i \mathbb{I}(0 < r_i \leq k)$ and $\mathrm{MRR} = \frac{1}{N}\sum_i \mathbb{I}(r_i > 0)/r_i$ with $N = 489$ throughout, so a query whose ground truth is outside a method's candidate space counts as a failure. Every ranking is frozen before the test labels are read. The ground truth is in the 7,074-identity

Table 8: **Inputs, candidate spaces, and ranking rules on FA-489.** *GT cov.*: queries whose ground truth is in the candidate space (an upper bound on R@k hits). *Match meta.*: whether the query's match or teams are used to build the candidate space. Independently of the query, our roster database covers the clubs of the three test competitions (Appendix A.4).

Family	Query input	Candidate space	Match meta.	GT cov.	Ranking rule
Face recognition	aligned face crop	full gallery, 7,074 IDs	no	454	cosine similarity to mean prototype
Re-ID / encoders	whole-player crop	full gallery, 7,074 IDs	no	454	cosine similarity to mean prototype
VLM	whole-player crop	100 identities per query (two teams' squads + same-league fillers)	yes	442	max. yes/no logit margin over references
MiniCPM	whole-player crop	two teams' squad names ∪ CLIP top-100 names	yes	481	candidate-constrained generation
Ours	face, upper-body, number, and player crops	1,905 roster players ∪ top-100 face candidates ($\bar{K}$ = 1,983)	no	485	reliability-aware stitching

Table 9: **Bounds on all 495 test observations**, obtained from the FA-489 results without re-running any model.

Method	Bound	R@1	R@5	R@10	MRR
Qwen3-VL-Reranker-8B	upper	30.10	58.99	69.90	0.430
Qwen3-VL-8B	upper	26.26	55.35	66.06	0.393
Qwen3-VL-32B	upper	26.26	54.14	65.86	0.390
InternVL3.5-8B	upper	15.35	37.17	49.70	0.266
Ours	lower	**60.20**	**71.72**	**75.35**	**0.655**

gallery for 454 of the 489 queries; the remaining 35 (27 players) are failures for every gallery-based method. Table 8 summarizes the candidate space of each method family.

All 495 test observations. Table 9 rescales the results to all 495 test observations under the assumption least favorable to our method: the six observations without a re-detected face count as failures for our method and as rank-1 hits for every VLM.

B.2 Face Recognition, Re-ID, and Visual Encoder Baselines

ArcFace (Deng et al., 2019) and AdaFace (Kim et al., 2022) (IR-101) and KP-RPE (Kim et al., 2024) (ViT-Base) use the public CVLFace weights trained on WebFace4M. Each receives the face crop of the query, re-detected with SCRFD (Guo et al., 2021) at threshold 0.25 and input size 1024×1024 (the setting of our face expert) and aligned to 112×112 with the standard five-landmark similarity transform; if no face is detected, the crop is padded to a square and resized. Gallery images are processed identically, and the query is ranked against the mean L2-normalized embedding of each of the 7,074 identities. Our face expert shares the detector, detection setting, and alignment, and uses a smaller recognizer (InsightFace ResNet-50 trained on WebFace600K (Zhu et al., 2021)); it also searches the player crop, encodes three views of each face, pools embeddings across crops, and scores an identity by its maximum similarity over gallery images. Re-scoring the frozen baseline embeddings with the maximum over gallery images gives 37.83 (ArcFace), 38.45 (AdaFace), and 42.13% R@1 (KP-RPE), within 1.7 points of the prototype results in Table 3; ranking the same 7,074 identities with our face expert gives 54.40% R@1.

CLIP ViT-B/32 (Radford et al., 2021), DINOv2 ViT-B/14 (Oquab et al., 2024), and the official Market-1501 checkpoints of OSNet x1.0 (Zhou et al., 2019) and TransReID ViT-B/16 (He et al., 2021) encode the whole-player crop and the gallery images; features are L2-normalized, averaged per identity into a prototype, and ranked by cosine similarity. Because these encoders reach at most 9.81% R@1 even with clean gallery photographs as queries (Table 2a), their failure mainly reflects that they are not trained to discriminate thousands of player identities.

Table 10: **All methods on identical candidates**: every method ranks the same frozen Fixed-100 pool of each FA-489 query with the same gallery references (ground truth reachable for 442/489, an upper bound of 90.39%).

Method	Query input	R@1	R@5	R@10	MRR
Qwen3-VL-Reranker-8B	player crop	29.24	58.49	69.53	0.423
Qwen3-VL-8B	player crop	25.36	54.81	65.64	0.386
Qwen3-VL-32B	player crop	25.36	53.58	65.44	0.383
InternVL3.5-8B	player crop	14.31	36.40	49.08	0.257
ArcFace (Deng et al., 2019)	face crop	49.69	56.44	60.53	0.537
AdaFace (Kim et al., 2022)	face crop	50.10	57.26	59.71	0.539
KP-RPE (Kim et al., 2024)	face crop	53.78	59.51	63.80	0.575
Our face expert alone	face crop	72.39	**82.62**	**84.87**	0.768
Ours (stitching)	face, kit, number	**74.23**	80.57	81.60	**0.774**

B.3 VLM Baselines

We evaluate Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025), and Qwen3-VL-Reranker-8B (Li et al., 2026) at pinned revisions. Each model verifies the benchmark query, the whole-player crop that contains the face together with the kit and the number, against the reference photographs of each candidate, which is how a multimodal reranker scores a query against documents.

Fixed-100 candidate pools. For each query, we use the match metadata to load the season squad lists of the two participating teams and link each name to the gallery after normalizing case, accents, and punctuation (all gallery identities sharing a name are kept); these form a core set of 42.5 identities on average. The remaining slots are filled with players of other teams in the same league and season, ordered by a fixed SHA-256 hash of the seed string `TABLE3_FIXED100_V3`, league, season, and identity (other seasons of the same league for 27 queries), and the 100 identities are shuffled by a separate fixed hash so that position carries no information. The ground truth, query image, and model scores are never used, and the pool is frozen before inference and shared by all VLMs. The ground truth is in the pool for 442 of the 489 queries. The pools use 1,075 gallery identities with 1 to 6 (on average 1.85) reference images each; cached face-box references are used for 938 identities and raw references otherwise. Eleven of the 12 gallery-supported misses come from name-linkage differences between sources (eight roster aliases and three name-order mismatches); counting all 11 as rank-1 hits would add at most 2.25 points to any metric, leaving the strongest VLM below 31.5% R@1.

Scoring. The query is the whole-player crop and the document one candidate reference image; names and file paths are never shown to the model. The system prompt is "Judge whether the Document meets the requirements based on the Query and the Instruct provided. Note that the answer can only be "yes" or "no"." and the instruction is "Determine whether the person in the Document is the same individual as the person in the Query. The Query is a broadcast crop of one player and the Document is a reference photograph. Use visible identity evidence. Different clothing or surroundings alone do not imply different individuals." For query q and reference r, the score is the raw next-token logit margin $m(q, r) = z_{\text{yes}}(q, r) - z_{\text{no}}(q, r)$, a candidate c with references $\mathcal{R}_c$ scores $s(q, c) = \max_{r \in \mathcal{R}_c} m(q, r)$, and the 100 identities are ranked by $s(q, c)$.

B.4 Candidate-Controlled Comparison

Table 10 ranks, for every method, exactly the Fixed-100 pool of each query with the same gallery references, so the methods differ only in their input and recognizer. Face-specialized models, which see an aligned face crop, clearly outperform the VLMs, which see the whole-player crop, and our method remains the strongest. Within this roster-restricted pool the face expert already reaches 72.39% R@1, so stitching adds less here (1.84 points) than on the full candidate set (6.75 points, Table 4), where the structured cues must also exclude candidates outside the two teams. Appendix C.1 breaks these results down by face size, number legibility, and competition.

Table 11: **MiniCPM-V-4.6 on FA-489** with identical candidate lists (ground truth in the list for 481/489).

Method	R@1	R@5	R@10	MRR
MiniCPM-V-4.6 (zero-shot)	2.45	9.20	14.93	0.071
MiniCPM-V-4.6 + LoRA	**7.36**	**19.84**	**26.99**	**0.140**

B.5 MiniCPM-V-4.6 with LoRA

This baseline asks a VLM to select the player's name from a candidate list. The base model is MiniCPM-V-4.6 (Yao et al., 2024) (SigLIP2-400M vision encoder and Qwen3.5-0.8B language model, about 1.3B parameters). Candidates are the squad names of the query's two teams (match metadata) together with the top-100 names retrieved by CLIP (Radford et al., 2021) from a larger SoccerWiki index (9,296 identities, 24,210 images); the ground truth is in the list for 481 of the 489 test queries. The prompt is:

```
<image> Identify the professional soccer player shown in
this broadcast image.  Select exactly one identity from the
candidate list below.  Use only the image and the provided
candidate identities.  Do not answer with a player outside
the list.  Return only the player's name exactly as written
in the candidate list.  Do not provide any explanation.
Candidate identities (N): 1.  <name> 2.  <name> ...
```

LoRA (Hu et al., 2021) ($r = 16$, $\alpha = 32$, dropout 0.05, on the q/k/v/o and gate/up/down projections of the language model; vision tower and projector frozen) is trained for one epoch (192 steps, learning rate 3×10^{-5}, cosine schedule, effective batch size 8, bf16, seed 42) on the 512 face-available validation queries whose ground truth is in their candidate list, each with three candidate orders (1,536 records). There is no image or match overlap with the test split, but 51 of the 60 identities shared by the two splits occur among the training queries. At test time, names are decoded with a candidate trie (beam size 10), the top five are removed from the trie, and decoding is repeated to build the ranking.

B.6 Our Method on the Baseline Gallery

To test whether our advantage stems from its candidate space, we restrict our final ranking stage to those of its candidates that are among the 7,074 gallery identities (1,021 per query on average) and recompute the posterior, consensus, reciprocal rank fusion, and final score inside this set; all expert outputs are unchanged, and 453 of the 489 ground truths remain reachable. Applying the same procedure to the unrestricted set reproduces the original ranking for all 489 queries. The restricted ranking reaches 61.35% R@1, 73.01% R@5, 76.69% R@10, and 0.667 MRR, essentially matching the unrestricted result, so the improvement over the gallery-based baselines does not stem from larger candidate coverage.

C Additional Results

C.1 Benchmark Analysis Details

Clean identification. Table 2a uses every gallery photograph of the 908 players with at least two photographs (3,200 photographs) as a query against all 7,074 identity prototypes, recomputing the query player's prototype without the query photograph. The models, gallery, and ranking rule are those of Table 3; the broadcast column uses the 454 FA-489 queries whose player is in the gallery. Some photographs of a player may come from the same source session. Over the full leave-one-out set, ArcFace, AdaFace, and KP-RPE reach 0.964, 0.965, and 0.963 MRR on clean photographs, against 0.438, 0.452, and 0.507 on SoccerCanon.

Breakdown on identical candidates. Table 12 complements Table 2b with the methods of Table 10. Every method loses accuracy on smaller faces, and the VLM changes least with face size

Table 12: **R@1 by condition on identical Fixed-100 candidates** (FA-489). VLM: Qwen3-VL-Reranker-8B; the two queries without an output in its per-query file count as failures (29.04 instead of 29.24 overall). Number legibility uses our number expert's reading of the squad number on the 444 queries with a known team–season.

Variable	Level	N	VLM	KP-RPE	Our face	Ours
Face size	<32 px	91	24.18	42.86	54.95	60.44
	32–47 px	245	28.57	55.10	71.43	72.65
	≥48 px	153	32.68	58.17	84.31	84.97
Number	read correctly	58	22.41	53.45	70.69	75.86
	not read	386	30.57	56.48	75.39	76.94
Competition	EPL	252	27.38	48.81	64.68	65.48
	Ligue 1 (PSG)	136	31.62	58.09	80.88	85.29
	Bundesliga	101	29.70	60.40	80.20	81.19
All		489	29.04	53.78	72.39	74.23

Table 13: **Cue ablation** with fixed weights. Rows with face and a structured cue use the roster-gated face distribution.

F	A	N	R@1	R@5	R@10	MRR
✓			54.19	64.83	68.92	0.591
	✓		0.61	2.04	5.93	0.028
		✓	0.41	0.82	2.04	0.014
✓	✓		55.01	65.64	69.33	0.598
✓		✓	54.81	65.24	69.33	0.597
	✓	✓	3.07	7.16	10.84	0.057
✓	✓	✓	**55.21**	**66.26**	**70.14**	**0.603**

Table 14: **Native-level accuracy** of the kit and number experts on the 444 queries whose team–season is in the roster database.

Expert	Metric (queries)	Acc.
Kit	team–season top-1 (444)	10.36
Kit	team–season top-5 (444)	30.63
Number	reading, all (444)	13.06
Number	reading, $\rho_N \geq 0.3$ (91)	34.07
Number	reading, $\rho_N < 0.1$ (299)	8.03

(24–33%). EPL is the hardest competition for every method, and the Ligue 1 subset, which consists of Paris Saint-Germain fixtures, stays within five points of the Bundesliga for every method.

C.2 Ablations

Tables 13–18 share the same frozen expert outputs, the same 489 queries, and the same per-query candidate set (ground truth present for 485); only the aggregation component in each row changes, ties follow Appendix D, and the labels are read only after all rankings are produced. One query corresponds to 0.20 points of R@k, and the 95% confidence interval of an R@1 near 60% spans about ± 4.3 points, so we only interpret differences supported by paired tests.

Cues. Table 13 combines the retained cues with fixed weights $(\lambda_F, \lambda_A, \lambda_N, \lambda_{AN}) = (4, 2, 0.4, 0.6)$. The face is by far the strongest single cue, whereas appearance and number alone reach below 1% R@1, because they predict team–seasons and numbers rather than players; with fixed weights, adding both changes R@1 only from 54.19% to 55.21%. At their native level (Table 14), both are weak but informative: the kit expert ranks the correct team–season among 75 first for 10.36% of the queries (chance 1.33%), and the number expert's accuracy rises from 8.03% to 34.07% as its reliability ρ_N rises, so its reliability indicates when to trust it. Part of the number errors stem from localization: in both qualitative cases (Appendix E), the shorts number is visible in the player crop but the number crop lands on the torso.

Structured evidence. Table 15 shows that adding kit and number to the face without coupling changes R@1 by one query (54.19 to 54.40%); the kit–number interaction mainly raises R@10 (69.12 to 70.14%), and the soft roster gate raises R@1 to 55.21%. The gate never removes a candidate: it activates only when $c_S = \max(\rho_A, \rho_N) \geq 0.28$ and then multiplies the face probability of out-of-roster candidates by $1 - c_S(1 - 0.72) \geq 0.72$; applied to the face alone, it raises R@1 to 54.60%.

Table 15: **Structured evidence ablation** with fixed cue weights. The last row equals the all-cue row of Table 13.

Configuration	Kit-Num.	Gate	R@1	R@5	R@10	MRR
Face + roster gate only		✓	54.60	65.24	69.12	0.594
Face + Kit + Number (no coupling)			54.40	65.44	69.12	0.595
+ Structured Kit-Number	✓		54.60	66.05	70.14	0.599
+ Roster gate	✓	✓	**55.21**	**66.26**	**70.14**	**0.603**

Table 16: **Aggregation component ablation.** *Correct*: queries ranked correctly at rank 1. The shaded row is our final model.

Aggregation	Reliability	Conflict	Consensus	Rank fusion	Correct	R@1	R@5	R@10	MRR
Fixed cue weights ($\rho_m = 1$)					270	55.21	66.26	70.14	0.603
+ Query-specific reliability	✓				269	55.01	65.24	69.94	0.600
+ Conflict discounting	✓	✓			269	55.01	65.85	70.35	0.602
+ Face–structured consensus	✓	✓	✓		275	56.24	67.28	70.96	0.611
+ Reciprocal rank fusion	✓	✓	✓	✓	**298**	**60.94**	**72.60**	**76.28**	**0.663**

Table 17: **Interpolation weight λ.** Validation sweep (611 queries; $\lambda = 0.4$ selected) and test sensitivity (FA-489; not used for selection).

λ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Val. R@1	47.63	47.79	47.95	47.63	**50.57**	49.26	48.12	46.81	46.48	45.83	45.34
Val. MRR	0.513	0.514	0.517	0.522	**0.539**	0.532	0.524	0.514	0.508	0.501	0.496
Test R@1	61.76	–	60.94	60.94	60.94	60.53	58.69	–	–	–	56.24
Test MRR	0.670	–	0.665	0.665	0.663	0.657	0.644	–	–	–	0.611

Aggregation components. Table 16 starts from fixed weights ($\rho_m = 1$ for every cue) and adds the components of Appendix D one at a time. Query-specific reliability and conflict discounting do not change R@1 by themselves (270, 269, 269 correct queries); they set the per-query weights from which the consensus is formed. Face–structured consensus yields 275 correct queries (paired bootstrap 95% CI of the R@1 gain over fixed weights: $[0.20, 2.04]$ points), and reciprocal rank fusion adds 23 more (CI $[2.86, 6.75]$), raising MRR from 0.611 to 0.663, so the posterior and the consensus ranking carry complementary information.

Interpolation weight. Table 17 sweeps the weight λ of Eq. (21) on the 611 face-available validation observations, where $\lambda = 0.4$ maximizes both R@1 and MRR; it is not tuned on the test set. Recomputed on the test set for robustness only, R@1 stays between 60.53% and 61.76% for $\lambda \in [0, 0.5]$, against 56.24% for the posterior alone ($\lambda = 1$).

Gains by face reliability. Table 18 splits FA-489 into terciles of the face reliability ρ_F computed by our method, which uses no ground truth, and compares the face-only and final rankings. Of the 33 net corrections, 32 fall in the two lower terciles, where the face cue alone reaches only 32–39% R@1. Because ρ_F is computed on the roster-gated face distribution, we repeat the split with the face-quality score q_F, computed before any other cue is used: the net corrections are 15, 12, and 6 from the lowest to the highest tercile (face-only R@1 30.67%, 55.83%, and 76.07%). Over all queries, the R@1 gain of 6.75 points has a paired bootstrap 95% CI of $[4.29, 9.20]$ (10,000 resamples), and the exact McNemar test on the 36 corrected and 3 broken queries gives $p < 10^{-7}$.

Computational Cost Table 19 reports the wall-clock cost logged during the test runs on a single NVIDIA A100 in Colab. The VLM baselines were run on 493 test queries (including the 489 of FA-489), scoring about 197 query–reference pairs per query, one pair per forward pass in NF4 precision; their times cover preprocessing and forward passes and exclude model loading. For our method, run on the 489 FA-489 queries, the face and number experts run on an A100 machine, with the face detector and ArcFace recognizer executed on the CPU through ONNX Runtime. Excluding the kit expert, the face, number, and stitching stages of our method take 17.9 s per query in total, compared with 37–71 s per query for the VLM baselines under the protocol above.

Table 18: **R@1 by tercile of face reliability** ρ_F (163 queries each). *Corrected/Broken*: queries that our method ranks correctly while the face cue does not, and vice versa.

Tercile	ρ_F range	Face only	Ours	Corrected	Broken
Low	0.069–0.209	32.52	41.72	15	0
Mid	0.209–0.258	39.26	49.69	18	1
High	0.258–0.702	90.80	91.41	3	2
All	0.069–0.702	54.19	60.94	36	3

Table 19: **Logged computational cost** on the test queries (single A100).

Method / stage	Total	Per query
Qwen3-VL-Reranker-8B	5.2 h	38 s
Qwen3-VL-8B	**5.1 h**	**37 s**
Qwen3-VL-32B	8.8 h	64 s
InternVL3.5-8B	9.7 h	71 s
Ours: face expert	0.87 h	6.4 s
Ours: number expert	0.51 h	3.7 s
Ours: stitching (CPU)	1.06 h	7.8 s
Ours: total (face, number, stitching)	**2.44 h**	**17.9 s**

Table 20: Notation used in the evidence aggregation framework.

Symbol	Meaning
q	Query player observation.
$\mathcal{C}$, $K = \lvert\mathcal{C}\rvert$	Shared player candidate set and its size.
z_F $(= e)$, z_A, z_N	Expert representations of the face, appearance, and number cues (Section 4.2).
P_F, P_A, P_N	Face, appearance, and jersey-number distributions over $\mathcal{C}$.
P_{AN}	Appearance–number interaction distribution over $\mathcal{C}$.
P_S, P_{cons}	Structured distribution and face–structured consensus distribution.
q_m	Quality of cue m.
ρ_m, $\widetilde{\rho}_m$	Query-specific reliability of cue m and its conflict-adjusted value.
λ_m	Fixed cue weight.
$H(P)$, $\rho_H(P)$	Shannon entropy and entropy reliability $1 - H(P)/\log K$.
D_{JS}, κ	Normalized Jensen–Shannon divergence and peak-support compatibility.
$\boldsymbol{\alpha}$	Dirichlet concentration over $\mathcal{C}$.
$r_{\text{Dir}}(c)$, $r_{\text{cons}}(c)$	Rank of c under the posterior and under P_{cons}.

D Method Details

This section specifies the notation, candidate space, cue extraction, reliability, conflict, consensus, and fusion of our method, together with its main constants (Table 21). The reference resources it uses (face gallery, kit references, and rosters) are described in Appendix A.4.

D.1 Problem Formulation and Notation

For a query broadcast observation q, the goal is to rank canonical player identities in a candidate set $\mathcal{C} = \{c_1, \ldots, c_K\}$ with $K = |\mathcal{C}|$. The method combines three experts $m \in \{F, A, N\}$ for face, appearance (kit), and jersey-number evidence. Two structured channels are derived from them: AN, the explicit appearance–number roster interaction, and cons, the face–structured consensus. Table 20 lists the notation, and Table 21 lists all constants and how each was set.

D.2 Candidate Space Construction

The roster database (75 team–season squad lists, 1,905 unique players) is normalized into canonical player identities, and all roster players enter the candidate set. The face expert additionally contributes its top-100 gallery identities. Face names are aligned to roster names by exact normalized-

name matching, followed by fuzzy matching with a similarity threshold of 0.86; a face identity that matches no roster name is kept as a separate face-only candidate:

$$\mathcal{C} = \mathcal{C}_{\text{roster}} \cup \mathcal{C}_{\text{face-only}}, \qquad \bar{K} = 1{,}983 \text{ on FA-489.} \tag{3}$$

This prevents the rosters from becoming a hard closed-set constraint. No match, team, or league metadata of the query is used.

D.3 Face Expert

D.3.1 Multi-View Face Representation

The face expert uses the InsightFace `buffalo_l` models: the SCRFD detector (Guo et al., 2021), run at a detection threshold of 0.25 and input size 1024×1024 on both the player crop and the face crop, and an ArcFace (Deng et al., 2019) ResNet-50 (He et al., 2016) recognizer. For each detected face, it forms three recognition embeddings,

$$e_{\text{raw}}, \qquad e_{\text{mask}}, \qquad e_{\text{flip}}, \tag{4}$$

from the aligned face, a soft-masked view, and the horizontally flipped face. For the soft-masked view, the aligned face (resized to 224×224) is encoded by DINOv2 (Oquab et al., 2024); each patch token is compared by cosine similarity with the mean token of the central patches (horizontal range $[0.28, 0.72]$, vertical range $[0.20, 0.76]$ of the grid), the similarity map is rescaled between its 5th and 95th percentiles, multiplied with an elliptical face prior ($s^{0.8}\,\pi^{0.9}$), smoothed with a Gaussian ($\sigma = 5$ px), and raised to the power 0.8 to give a soft mask a; the view is $a \cdot I + (1-a) \cdot \text{blur}_{\sigma=10}(I)$. Their nominal weights are $w_{\text{raw}} = 0.55$, $w_{\text{mask}} = 0.30$, and $w_{\text{flip}} = 0.15$, and the auxiliary branches are gated by their consistency with the raw embedding:

$$g_{\text{mask}} = \text{clip}\left(\frac{\cos(e_{\text{raw}}, e_{\text{mask}}) - 0.20}{0.70}, 0.05, 1\right), \qquad g_{\text{flip}} = \text{clip}\left(\frac{\cos(e_{\text{raw}}, e_{\text{flip}}) - 0.10}{0.80}, 0.10, 1\right). \tag{5}$$

With $\bar{w}$ the re-normalized products $w\,g$ (and $g_{\text{raw}} = 1$), the face representation is $e = \text{L2Norm}(\bar{w}_{\text{raw}} e_{\text{raw}} + \bar{w}_{\text{mask}} e_{\text{mask}} + \bar{w}_{\text{flip}} e_{\text{flip}})$, so masked or flipped views cannot dominate when they become inconsistent with the original identity embedding. When faces are found in both crops, their embeddings are pooled according to face quality and cross-crop consistency.

D.3.2 Face Quality

The face quality combines five terms in $[0, 1]$:

$$q_F = \text{clip}\left(0.25\, q_{\text{det}} + 0.20\, q_{\text{size}} + 0.18\, q_{\text{blur}} + 0.15\, q_{\text{pose}} + 0.22\, q_{\text{cons}},\ 0.05,\ 1\right), \tag{6}$$

where $q_{\text{det}} = \text{clip}((s_{\text{det}} - 0.20)/0.70)$ for detection score s_{det}, $q_{\text{size}} = \text{clip}((\ell_{\min} - 28)/72)$ for the shorter face-box side $\ell_{\min}$ in pixels, $q_{\text{blur}} = \text{clip}\left((\log(1+b) - \log 13)/(\log 181 - \log 13)\right)$ for the Laplacian variance b of the aligned face, $q_{\text{pose}} = \exp(-|\text{yaw}|/45 - |\text{pitch}|/60 - |\text{roll}|/75)$ with angles in degrees, and $q_{\text{cons}} = \text{clip}\left((0.7\cos(e_{\text{raw}}, e_{\text{mask}}) + 0.3\cos(e_{\text{raw}}, e_{\text{flip}}) - 0.20)/0.75\right)$; clip without bounds clips to $[0, 1]$.

D.3.3 Face Retrieval

Each gallery identity is scored by the maximum cosine similarity between e and the embeddings of its gallery images. The top-100 identities are retained, and their scores are converted to P_F with a softmax of temperature $T_F = 0.06$; all other candidates receive zero face mass.

D.4 Appearance Expert

The appearance expert uses DINOv2-small (Oquab et al., 2024) together with complementary color and pattern descriptors and ranks all 208 kit references. Six component similarities are used: optimal-transport matching of color distributions (Color-OT), HSV histogram similarity, optimal-transport matching of DINOv2 patch features (DINO-OT), DINOv2 global-feature similarity, and

pattern and layout similarities. The component weights adapt to the mean quality $\bar{q}_A \in [0,1]$ of the query's upper-body views. Before normalization,

$$w_{\text{color}} = 0.34 + 0.12\bar{q}_A, \qquad w_{\text{DINO-OT}} = 0.26 + 0.08(1-\bar{q}_A), \qquad w_{\text{pattern}} = 0.10, \tag{7}$$

$$w_{\text{HSV}} = 0.16 + 0.05\bar{q}_A, \qquad w_{\text{DINO-global}} = 0.14 + 0.05(1-\bar{q}_A), \qquad w_{\text{layout}} = 0.08, \tag{8}$$

and the six weights are then normalized to sum to one. For each component, only the strongest two query views are aggregated, and each component score s is mixed with its within-component rank score r as $s' = 0.70s + 0.30r$. The kit-label distribution combines arithmetic and geometric ensembles of the components, $P_A^{\text{label}} = 0.48P_{\text{arith}} + 0.52P_{\text{geom}}$, is divided by $n_\ell^{0.5}$ for a label with n_ℓ reference images, and is re-normalized. Kit labels are then marginalized over kit type (home, away, third) to a team–season distribution $P_A(t,s)$.

D.5 Number Expert

The number expert applies EasyOCR (Kittinaradorn, 2020), restricted to digits with minimum confidence 0.05, to several number-oriented crops of the query and to deterministic preprocessing variants of each crop. A reading n in view v receives the score $s_{v,n} = c_{\text{OCR}}\, w_{v,n}\, (0.45 + 0.55 q_v^N)$, where c_{OCR} is the OCR confidence, $w_{v,n} \in (0,1]$ is a fixed prior weight for the preprocessing variant, box size, and reading length (its values are part of the released code), and $q_v^N \in [0,1]$ is the image quality of number crop v. Preprocessing variants of the same physical crop are not treated as independent evidence: only the strongest score of a number within each crop is kept. Scores are then summed across the V_n independent crops that support n and rewarded for repeated support, $s_n \leftarrow s_n[1 + 0.34(V_n - 1)]$, and converted to a jersey-number distribution with temperature $T_N = 0.30$; an *unknown* state absorbs the remaining mass when no stable number is read.

D.6 Projection into a Shared Player Space

Appearance. For roster entry (c_i, t, s) of player c_i, team t, and season s,

$$\text{score}_A(c_i, t, s) = \frac{P_A(t,s)}{|\mathcal{R}_{t,s}|}, \tag{9}$$

where $|\mathcal{R}_{t,s}|$ is the roster size. If no kit reference exists for the exact season, team-only evidence is weakened, $P_A(t,s) = 0.20 P_A(t)$. A player listed in several roster entries keeps only the strongest one, $P_A(c_i) \propto \max_{(t,s)\in\mathcal{R}_i} \text{score}_A(c_i,t,s)$, and the vector is normalized over $\mathcal{C}$.

Number. The mass of number n is divided equally among all roster players who wear n in any roster, again keeping each player's strongest roster entry; this projection uses no appearance information.

Appearance–number interaction. $P_{AN}(c_i) \propto \max_{(t,s)\in\mathcal{R}_i} P_A(t,s)\, P_N^{\text{num}}(n_{i,t,s})/d_{t,s,n}$, where $n_{i,t,s}$ is the jersey number of c_i in roster (t,s) and $d_{t,s,n}$ is the number of players sharing that number in that roster; it supports the player wearing the predicted number in the predicted team–season.

D.7 Reliability

For a player-space distribution P, the entropy reliability is $\rho_H(P) = 1 - H(P)/\log K$. The cue reliabilities are

$$\rho_F = q_F\, \rho_H(P_F), \tag{10}$$

$$\rho_A = \Big(1 - \frac{H(P_A^{\text{team}})}{\log|\mathcal{T}|}\Big)\,(0.5 + 0.5\, q_{\text{map}}), \tag{11}$$

$$\rho_N = \min(q_N^0,\, \bar{q}_N)\ \text{clip}\,\big(0.55 + 0.25\,(1-u_N)\,\rho_H(P_N^{\text{num}}) + 0.20\,\delta_N\big), \tag{12}$$

where P_A^{team} is the appearance distribution over the set $\mathcal{T}$ of team–season labels, q_{map} is the fraction of appearance mass that maps to exact team–season rosters, P_N^{num} is the distribution over the 100 jersey numbers, u_N is the mass of the *unknown* state, q_N^0 is the OCR-level number quality, δ_N is the margin between the two most probable numbers, and $\bar{q}_N$ is a cap set by independent-crop support:

0 if no number is read; 0.10 if the preprocessing stage found no digits or no crop supports the top reading; 0.35 if exactly one crop supports it or no per-crop readings are available; and, if at least two crops support it, 0.90/0.75, 0.60/0.50, 0.40/0.30, or 0.35/0.20 for crop agreement ratios of at least 0.80, 0.67, 0.50, or below 0.50, where the first value applies when the top reading also matches the reading of the preprocessing stage. A sharp number distribution alone therefore cannot produce a high number reliability.

Soft roster gate. Before fusion, the structured confidence $c_S = \max(\rho_A, \rho_N)$ controls a soft gate on the face distribution. If $c_S \geq 0.28$, the face probability of every candidate outside all rosters is multiplied by $1 - c_S(1 - 0.72) \geq 0.72$ and P_F is re-normalized; ρ_F is computed on the gated P_F.

D.8 CONFLICT

For two distributions P and Q over $\mathcal{C}$, we use the normalized Jensen–Shannon divergence $D_{\mathrm{JS}}(P, Q) = \left[\frac{1}{2} D_{\mathrm{KL}}(P\|M) + \frac{1}{2} D_{\mathrm{KL}}(Q\|M)\right] / \log 2 \in [0, 1]$ with $M = \frac{1}{2}(P + Q)$, and the peak-support compatibility

$$\kappa(P, Q) = \max_i \sqrt{\frac{P(c_i)}{\max_j P(c_j)} \cdot \frac{Q(c_i)}{\max_j Q(c_j)}}, \tag{13}$$

which is 1 when some candidate is strongly supported by both. The pairwise conflict is

$$\gamma_{mn} = D_{\mathrm{JS}}(P_m, P_n) \left(1 - \kappa(P_m, P_n)\right), \tag{14}$$

so broad but compatible distributions, such as a kit plateau over a roster, are not treated as conflicting. Each cue is discounted only by cues that are more reliable by more than a margin of 0.03:

$$\bar{\gamma}_m = \frac{\sum_{n:\, \rho_n > \rho_m + 0.03} \rho_n \gamma_{mn}}{\sum_{n:\, \rho_n > \rho_m + 0.03} \rho_n}, \qquad \widetilde{\rho}_m = \max\left(0.35\, \rho_m,\ \rho_m \exp(-\beta\, \bar{\gamma}_m)\right), \quad \beta = 1.5, \tag{15}$$

with $\bar{\gamma}_m = 0$ if no such cue exists. The most reliable cue is therefore never discounted, and the floor $0.35\rho_m$ keeps every valid cue as complementary evidence. The interaction reliability is

$$\rho_{AN} = \sqrt{\widetilde{\rho}_A \widetilde{\rho}_N}\ \kappa(P_A, P_N)\ \left(0.5 + 0.5\, \rho_H(P_{AN})\right). \tag{16}$$

D.9 CONSENSUS

The structured branch combines appearance and appearance–number evidence,

$$P_S \propto 0.35\, P_A + 0.65\, P_{AN}, \tag{17}$$

where both inputs are normalized first. The face–structured consensus and its reliability are

$$P_{\mathrm{cons}}(c_i) \propto \sqrt{P_F(c_i)\, P_S(c_i)}, \qquad \rho_{\mathrm{cons}} = \sqrt{\widetilde{\rho}_F\ \max(\widetilde{\rho}_A, \rho_{AN})}\ \kappa(P_F, P_S). \tag{18}$$

D.10 DIRICHLET FUSION

All channels are accumulated as Dirichlet evidence over $\mathcal{C}$:

$$\boldsymbol{\alpha} = \alpha_0 \mathbf{1} + \sum_{m \in \{F, A, N\}} K \lambda_m \widetilde{\rho}_m P_m + K \lambda_{AN} \rho_{AN} P_{AN} + K \lambda_{\mathrm{cons}} \rho_{\mathrm{cons}} P_{\mathrm{cons}}, \tag{19}$$

with $\alpha_0 = 1$, $(\lambda_F, \lambda_A, \lambda_N, \lambda_{AN}) = (4, 2, 0.4, 0.6)$, and $\lambda_{\mathrm{cons}} = 0.35$. The posterior is $s_{\mathrm{Dir}}(c_i) = \alpha_i / \sum_j \alpha_j$. The fixed-weight variant of Appendix C.2 sets all reliabilities to one and omits the consensus term. Two uncertainty diagnostics are computed but not used for ranking: evidence insufficiency $u_{\mathrm{ev}} = K\alpha_0 / \sum_j \alpha_j$ and conflict u_{conf}, the average of γ_{mn} weighted by $\rho_m \rho_n$.

Table 21: Constants of our method. λ is selected by the validation sweep in Table 17; all other constants were fixed during development using only validation data and were never changed after inspecting test results.

Component	Constants	Set by
Candidate space	top-100 face candidates; fuzzy name threshold 0.86	development
Face views	$(w_{\text{raw}}, w_{\text{mask}}, w_{\text{flip}}) = (0.55, 0.30, 0.15)$; gates in Eq. (5); mask constants above	development
Face retrieval	$T_F = 0.06$	development
Appearance	component weights above; $s' = 0.7s + 0.3r$; ensemble $0.48/0.52$; count exponent 0.5; team-only factor 0.20	development
Number	$c_{\min} = 0.05$; support bonus 0.34; $T_N = 0.30$; caps $\bar{q}_N$	development
Roster gate	activation 0.28; floor 0.72	development
Conflict	$\beta = 1.5$; floor ratio 0.35; margin 0.03	development
Fusion	$\alpha_0 = 1$; $\lambda = (4, 2, 0.4, 0.6)$; $\lambda_{\text{cons}} = 0.35$; structured mix $0.35/0.65$	development
Final ranking	$k_{\text{RRF}} = 60$ (Cormack et al., 2009); $\lambda = 0.4$	k: standard; λ: validation sweep

D.11 Final Ranking

Let $r_{\text{Dir}}(c)$ be the rank of c by decreasing s_{Dir} (ties broken by candidate identifier) and $r_{\text{cons}}(c)$ its rank by decreasing P_{cons} among candidates with $P_{\text{cons}}(c) > 0$. Reciprocal rank fusion (Cormack et al., 2009) gives

$$s_{\text{RRF}}(c) = \frac{1}{k_{\text{RRF}} + r_{\text{Dir}}(c)} + \mathbf{1}_{\{P_{\text{cons}}(c)>0\}} \frac{1}{k_{\text{RRF}} + r_{\text{cons}}(c)}, \qquad k_{\text{RRF}} = 60, \tag{20}$$

and the final score is

$$s_{\text{final}}(c) = \lambda \frac{s_{\text{Dir}}(c)}{\max_{c'} s_{\text{Dir}}(c')} + (1 - \lambda) \frac{s_{\text{RRF}}(c)}{\max_{c'} s_{\text{RRF}}(c')}, \qquad \lambda = 0.4, \tag{21}$$

with λ selected on the validation split (Table 17). Candidates are sorted by decreasing s_{final}, with ties broken by r_{Dir}. This rule produces all reported results of our method.

E Qualitative Analysis

We trace one success and one failure through the stages of our method: the base posterior, the face–structured consensus P_{cons}, reciprocal rank fusion (RRF), and the final score (Eq. (21)). Figures 5 and 6 show the cue crops, and Table 22 lists the logged values. In the success case (Wes Morgan), two face-only distractors outrank the ground truth in the posterior but belong to no roster and receive no structured support, so fusion promotes the ground truth from rank 3 to rank 1, whereas all four VLMs fail. In the failure case (Angel Di Maria), the structured branch ranks the ground truth first and the posterior ranks it second behind Yoann Andreu; the exactly reversed ranks tie in RRF, so the posterior decides and the ground truth stays at rank 2, while the three Qwen-based VLMs rank it first. In both cases the shorts number is visible in the player crop, but the number crop lands on the torso.



Figure 5: **Wes Morgan's cue crops**: localized player, face crop, upper-body crop, and jersey-number crop.



Figure 6: **Angel Di Maria's cue crops**: localized player, face crop, upper-body crop, and jersey-number crop.

Table 22: **Stage-by-stage trace of a success and a failure case.** GT: ground truth.

Stage	**Success: Wes Morgan**	**Failure: Angel Di Maria**
Posterior	GT rank 3, behind Tashan Oakley-Boothe and Habib Diallo, two face-only candidates in no roster	GT rank 2 ($P = 0.0346$) behind Yoann Andreu ($P = 0.0490$)
Kit	Leicester 2016–17 kit: $P_{\text{kit}} = 6.95 \times 10^{-4}$ for the GT, 0 for both distractors	Paris Saint-Germain kit: $P_{\text{kit}} = 1.39 \times 10^{-3}$ for the GT vs. 2.8×10^{-4}
Number	wrong reading, reliability 0.085; shorts number 5 missed	no stable reading (reliability 0); shorts number 11 missed
Reliability	$\rho_F = 0.199$; conflict lowers kit $0.054 \rightarrow 0.033$ and number $0.085 \rightarrow 0.030$	$\rho_F = 0.246$; conflict lowers kit $0.068 \rightarrow 0.044$
Consensus	GT $P_{\text{cons}} = 0.0754$ (rank 2); distractors 0	GT $P_{\text{cons}} = 0.1621$ (rank 1); Andreu 0.0881 (rank 2)
RRF	GT rank 1	exact tie with Andreu (0.032522)
Final	**GT rank 1**	GT rank 2 ($S_{\text{final}} = 0.8825$ vs. 1.000)
VLMs	all fail: Qwen3-VL-32B and 8B (Jeffrey Schlupp; GT ranks 15 and 6), Reranker (N'Golo Kanté; 20), InternVL3.5 (Yaya Sanogo; 3)	Qwen3-VL-32B, 8B, and Reranker rank the GT first; InternVL3.5 ranks Marco Verratti first (GT rank 2)