Global Embeddings, Local Feature Matching, and Adaptive Validation Overfitting in Animal Re-Identification*

Maayan Yesharim[1], Yoav Ram[1,2,*]

[1]School of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[2]Edmond J. Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv 6997801, Israel

Abstract

We describe our submission to the AnimalCLEF 2026 competition on open-set individual animal re-identification across four species (Eurasian lynx, fire salamander, loggerhead sea turtle, and Texas horned lizard), which placed second by a margin of less than 0.5% of the winning score. Our second-place pipeline is a per-species ensemble using both global embeddings and local feature matching, followed by hierarchical agglomerative clustering at a target cluster count calibrated by public Adjusted Rand Index (ARI). A post-hoc analysis of our 154 scored submissions and the full 232-team leaderboard shows that public ARI predicts private ARI strongly across all teams ($\rho = 0.95$), but the correlation weakens at the top of the leaderboard: within the top five teams the rank correlation is statistically indistinguishable from zero, and within our own top submissions it is negative, consistent with prior work on adaptive overfitting on a held-out public split. Our public-best submission was the most overfit pipeline in our archive, and one of our unlocked submissions would have ranked first among all teams' selected final submissions on the private leaderboard. Reconstructing the per-submission cluster partitions, we found that *train ARI* (computed on labeled training images) retains significant predictive power for private ARI at the top of the leaderboard, where public ARI loses its predictive power. This finding suggests train-side metrics as a practical fallback signal when public leaderboard scores saturate. We end with four exploratory decision rules for portfolio selection in re-identification competitions.

Keywords

Animal Re-Identification, Open-Set Classification, Hierarchical Clustering, Local Feature Matching

1. Introduction

Individual animal re-identification (re-ID) is the task of recognizing the same individual across photographs taken at different times, places, and viewpoints. It underpins capture–recapture analyses, demographic monitoring, and behavioral ecology, and is increasingly delivered as a service rather than a one-off study [1, 2]. The AnimalCLEF series [3, 4, 5], part of the LifeCLEF lab [6], reframes re-ID as a fully *open-set* clustering problem: given a mixed pool of test images, predict identity labels such that images of the same individual share a label and previously-unseen individuals form new clusters. The competition is scored by Adjusted Rand Index (ARI) [7, 8] on a hidden test split.

The four AnimalCLEF 2026 datasets reward four very different approaches: a Eurasian lynx (*Lynx lynx*) camera-trap dataset for which dataset-specific finetuning of a re-ID backbone was our dominant lever; a fire salamander (*Salamandra salamandra*) dataset for which local-feature matching on segmented yellow patterns outperformed global embeddings, with a self-trained embedding rescuer added to bridge the dorsal/profile viewpoint pairs that keypoints could not match; a loggerhead sea turtle (*Caretta caretta*) underwater-photographic dataset for which a generic re-ID embedding was already near-saturating; and a Texas horned lizard (*Phrynosoma cornutum*) hand-held ventral dataset with no training labels and where neural re-ID backbones seem to perform at chance. We (Ram Lab) treated each dataset on its own terms; the final pipeline is a union of four heterogeneous subpipelines glued together by a single clustering and submission harness. Our pipeline finished second on the private leaderboard with

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Working note for the AnimalCLEF 2026 task at LifeCLEF, hosted at CVPR & CLEF.

*Corresponding author.
yoavram@tauex.tau.ac.il (Y. Ram)
https://www.yoavram.com (Y. Ram)
0000-0002-9653-4458 (Y. Ram)

ARI $= 0.70050$, 0.00343 behind first place (0.70393) and 0.00418 above third (0.69632). Our anchor public score of 0.81876 regressed by 0.118 on private — the largest drop in the top five.

We describe the per-species pipelines, the hyper-parameter calibration procedure, and the portfolio strategy used to select the final five submissions for the competition. We then present a post-hoc analysis of the public-vs-private leaderboard correlation that explains why our public lead — the highest among all 232 teams — failed to translate to private, and why our final portfolio missed the configuration that would have won the competition. We computed per-submission supervised clustering metrics on the salamander labeled training split, and we use them to characterize which train-set signals retain predictive power for private ARI inside our top-20 submissions by public ARI (the regime where public itself loses predictive power for private). We suggest four concrete decision rules for portfolio selection on competitions whose public score saturates.

2. Related Work

Wildlife re-identification. Constructing a competitive re-ID pipeline for a new species requires chaining different components: animal detectors (MegaDetector [9], Segment Anything [10]), global re-ID embedding models (MiewID [11], MegaDescriptor [12]), local-feature extractors (SIFT [13], SuperPoint [14], ALIKED [15], DISK [16], LoFTR [17]), local-feature matchers (LightGlue [18]), geometric verification (e.g., RANSAC [19]), and clustering. Recently, WildFusion [20] formalized the global+local fusion as a calibrated similarity, PoseSwin [21] introduced pose-aware metric-learning, and OSFR [22] used an auxiliary age classifier to help fine-tune an ID classifier.

Held-out-test-set overfitting. The phenomenon that aggressive optimization against a fixed held-out set (such as a Kaggle public leaderboard) inflates apparent generalization is well documented. Blum and Hardt's *Ladder* algorithm [23] provides a theoretically robust public scoring rule precisely to avoid this; Roelofs et al. [24] show that despite n submissions per team, *adaptive overfitting* on Kaggle's public splits is empirically smaller than the worst case — but, as they note, top-of-leaderboard subsets are a worst case in expectation. Recht et al. [25] report a parallel effect on standard benchmarks: ranks on the original test set partly predict but routinely shuffle on a re-collected one. Our observations on AnimalCLEF 2026 are consistent with this literature; the post-hoc analysis below reports them as evidence in a specific setting, not as a novel phenomenon.

3. Task and Data

3.1. Datasets

AnimalCLEF 2026 curates a dataset spanning four species across two continents: Eurasian lynx and fire salamander images from the Czech Republic, loggerhead sea turtle images from Zakynthos in Greece, and Texas horned lizard images from Texas in the United States (Table 1). The first three subsets return from AnimalCLEF 2025 [3]; the Texas horned lizard subset is new to AnimalCLEF 2026 and ships with no labeled training data.

Each subset is partitioned into a labeled *train* portion (which may be empty, as in the Texas case) and an unlabeled *test* portion to be clustered. The test split mixes images of training-set individuals with images of previously-unseen individuals — the open-set setting we discussed in Section 2. Submissions are clusterings of the test images, with the constraint that each cluster identifier carries a species-specific prefix (e.g. `cluster_LynxID2025_*`). Total dataset size is 15,483 images across 2,409 test images and 13,074 labeled train images covering 1,102 known individuals.

Participants may use any re-identification dataset accessible through the `wildlife-datasets` package [12]; the external sets we drew on — CzechLynx [26] for the lynx finetune and SeaTurtleIDHeads [27] and ZindiTurtleRecall [28] for the sea turtle finetune — are all available this way.

Figure 1: Representative images from the four AnimalCLEF 2026 subsets. **A. Texas horned lizard:** hand-held ventral photographs; the handler's fingers are routinely visible at the frame edge. **B. Loggerhead sea turtle:** underwater head photographs with variable viewpoints and optical conditions. **C. Eurasian lynx:** camera-trap frames showing day-RGB and night-IR captures, with the animal cropped and pasted onto a black background to highlight the body silhouette and coat pattern. **D. Fire salamander:** dorsal and profile views of the yellow-spot pattern; same individuals can appear in either viewpoint.

The four subsets pose qualitatively different visual challenges (Figure 1). Lynx images come from camera traps with varying viewpoint and lighting, including day-RGB and night-IR frames, and exhibit seasonal coat change and substantial inter-individual texture overlap. For salamander images, approximately 48% of same-individual pairs are cross-viewpoint (dorsal vs. profile). Sea turtle images are underwater photographs with variable optical conditions; the test set is the smallest of the four by image count, but the off-the-shelf re-ID signal is the strongest. Texas horned lizard images are hand-held ventral photographs in which the handler's hands routinely dominate the frame.

Table 1
AnimalCLEF 2026 datasets. "Train IDs" is the number of unique individuals in the labeled training set; the test set may contain a mix of these and previously-unseen individuals. Texas horned lizards have no training data; we use a 277-image, 104-individual external calibration set (TCU) for local validation only.

Dataset	Species	Location	Train images	Test images	Total images	Train IDs
LynxID2025	Eurasian lynx	Czech Republic	2,957	946	3,903	77
SalamanderID2025	fire salamander	Czech Republic	1,388	689	2,077	587
SeaTurtleID2022	loggerhead sea turtle	Zakynthos, Greece	8,729	500	9,229	438
TexasHornedLizards	Texas horned lizard	Texas, USA	0	274	274	—
Total	—	—	13,074	2,409	15,483	—

3.2. Evaluation

Submissions are scored by the Adjusted Rand Index (ARI) [7, 8] on the hidden 2,409-image test split, with the score split into a public portion (∼48% of the test images, visible during the competition) and a private portion (∼52%, revealed at deadline) [4]. Each team may make up to five submissions to the public leaderboard per day, and selects (locks) at most five of those for private scoring at the deadline. For a predicted partition π and the ground-truth partition $\pi^\star$ on the N test images, the Rand index is the fraction of unordered image pairs the two partitions agree on (either co-clustered in both, or separated in both),

$$\mathrm{RI}(\pi, \pi^\star) \;=\; \frac{a+b}{\binom{N}{2}}, \tag{1}$$

where a counts pairs co-clustered in both partitions and b counts pairs separated in both. The Adjusted Rand Index chance-corrects this by subtracting the expected RI under random partitions with the same cluster-size marginals, $\mathrm{ARI} = (\mathrm{RI} - \mathbb{E}[\mathrm{RI}])/(1 - \mathbb{E}[\mathrm{RI}])$ [7, 29]. ARI equals 1 for a perfect match, 0 in expectation under random partitions, and can be negative for systematic disagreement; it heavily penalizes both over-merging and over-splitting.

Throughout the competition we tracked public ARI as our primary tuning signal; the train-side proxies we tested in lieu of (or alongside) public probing are reported in Section 5.1.

4. Methods

The pipeline routes each image to one of four species-specific subpipelines and concatenates the results. One of two approaches is used for each species: a combination of global embeddings for species in which deep metric learning is effective, or local-feature matching when it is not effective (salamander combines both, with local matching as the primary signal and a global embeddings as a helper for cross-viewpoint matching). Table 3 gives the per-species component breakdown; Table 2 lists all hyperparameters.

4.1. LynxID2025: blended embeddings + HAC at calibrated K

The lynx subpipeline learns two complementary embeddings and combines them.

The primary embedding, *v5b*, is a MiewID [30] backbone fine-tuned with ArcFace [31] on the union of CzechLynx (a 269-individual, ∼40k-image camera-trap dataset [26]) and the AnimalCLEF 2026 lynx training data (77 individuals), for a total of 346 individuals; the 946 lynx test images were excluded from training. We crop with MegaDetector [9] prior to embedding and apply horizontal-flip test-time augmentation (TTA; the embedding is computed on both the original and horizontally-flipped crop and averaged per pair). The released lynx images are padded to square aspect but not tightly cropped to the animal — the median MegaDetector bounding box covers only ∼53% of the square frame, and 1,549 of the 3,817 lynx images with a detected box have the animal in less than half the frame — so cropping still removes roughly half the pixels. We use the raw MegaDetector box; the $70\% \times 90\%$ inner-crop fraction

Table 2
Consolidated hyperparameters and training settings for the configuration we submitted (sub#231, our private-best at 0.70050 on the second-place result). All HAC distance matrices are $1 - \text{cosine similarity}$. "TTA" is test-time augmentation: the embedding (or matcher) is run on the original image and on a transformed copy (horizontal flip or $180°$ rotation) and the per-pair scores are averaged. The Texas $K = 205$ value is observed (a consequence of threshold 0.40) rather than targeted. Salamander pseudo-labels for self-training are derived from connected components of pairs with body match count ≥ 13 or head match count ≥ 7. Variants we built but did not select as our final submitted score (e.g. Lynx K_{test}=80 at sub#250, our public-best, or the aggressive-seed sub#238) are listed in Table 8.

Component	Setting
Lynx v5b finetune	MiewID backbone (EfficientNet-V2), ArcFace (s=64, m=0.5), 346 IDs, MegaDetector crops 384×384, hflip TTA, $b = 64$, AdamW, backbone lr=10^{-5}, head lr=10^{-4}, MPerClassSampler (m=1 per class), validated by mAP@R, 25 epochs
Lynx DINOv2 finetune	DINOv2-ViT-B/14, ArcFace (s=64, m=0.5), 346 IDs, 518×518, 25 epochs, lr=$5{\cdot}10^{-5}$ cosine schedule
Lynx blend / clustering	$S = 0.65\, S_{\text{v5b}} + 0.35\, S_{\text{DINO}}$, HAC average linkage, distance threshold tuned to target K_{test}=87
Salamander preprocessing	Body: SAM3 prompt "salamander" for body mask; PCA-rotate head-to-left; LAB b^* Otsu; 5-iter dilation; re-apply to RGB. Head crop: SAM3 prompt "head of fire salamander" with orient-aware geometric fallback (25% dorsal / 45% profile)
Salamander local features	SuperPoint up to 800 keypoints; LightGlue match confidence>0.95; body and head pipelines run independently; $180°$ rotation TTA on profile images
Salamander seed merging	body match count ≥ 13 *or* head match count ≥ 7, greedy connected components
Salamander self-trained ConvNeXt	ConvNeXt-Small (ImageNet pretrain), ArcFace (s=64, m=0.5), 50 epochs at 384×384 on labeled train + 93 pseudo-classes (325 test images)
Salamander cross-viewpoint seed	ConvNeXt cosine ≥ 0.55 between cross-viewpoint pairs adds direct merges
Salamander clustering	HAC average linkage, distance threshold $0.97 \Rightarrow K_{\text{test}}$=559
Sea turtle external finetune	MiewID backbone, ArcFace (s=64, m=0.5), ∼14k images from SeaTurtleIDHeads + ZindiTurtleRecall, 25 epochs
Sea turtle blend / clustering	$S = 0.75\, S_{\text{MiewID}} + 0.25\, S_{\text{ft}}$; HAC *complete* linkage, threshold 0.78, transitivity seed at MiewID-similarity ≥ 0.80, K_{test}=176
Texas crop	MegaDetector inner-crop (70% height ×90% width to suppress hand contamination)
Texas matching	ALIKED keypoints + LightGlue + RANSAC (inlier threshold 3 px)
Texas clustering	HAC average linkage, distance threshold $0.40 \Rightarrow K_{\text{test}}$=205

we apply on the Texas subpipeline (Section 4.4) is not used here. On the AnimalCLEF training split as a held-out ranking evaluation (3,903 lynx images, ground-truth identity from the labeled portion), v5b achieves mAP@R $= 0.87$, compared to off-the-shelf MiewID-msv3 at mAP@R $= 0.26$; training on the combined dataset with the wider individual base and the in-domain images is the lever responsible for this gain.

We also use a secondary embedding to add information that v5b lacks. We screened candidates by how *decorrelated* they were from v5b — a candidate whose pair-wise similarity matrix is weakly correlated with v5b's ranks pairs differently and has more potential to correct v5b's mistakes when blended. We quantified this as the Pearson correlation r between the v5b and candidate similarity matrices across all test image pairs. A second *rescue* diagnostic guards against the failure mode of being decorrelated for the wrong reasons (a noisy secondary is also decorrelated): rescue is the count of hard same-individual test pairs that the candidate brings into v5b's top-K neighborhood but that v5b alone

Table 3
Per-species pipeline components for the configuration we submitted (sub#231, second place at private 0.70050). Each row is a self-contained subpipeline. The clustering algorithm is HAC throughout, with average linkage except where noted. Other submitted variants we generated, including our public-best (sub#250, Lynx K_{test}=80) and the unselected sub#238, are listed in Table 8.

Species	Embedding (primary)	Embedding (secondary)	Local matcher	Clustering / target K
Lynx	MiewID v5b (CzechLynx+AnimalCLEF ArcFace finetune)	DINOv2-ViT-B 25ep ArcFace	—	avg linkage, $K = 87$
Salamander	—	ConvNeXt-Small (self-trained on pseudo-labels)	SuperPoint + LightGlue on head & body	avg linkage, $K = 559$
Sea turtle	MiewID (off-the-shelf, hflip TTA)	MiewID (external sea-turtle finetune)	—	complete linkage, $K = 176$
Texas	—	—	ALIKED + LightGlue + RANSAC on inner-crop	avg linkage, $K = 205$

misses — positive rescue means the candidate is moving genuine same-ID pairs closer in the blend, negative rescue means it is displacing them with impostors. A good secondary therefore has both low r with v5b and positive rescue.

We evaluated four candidate secondaries, all ArcFace-finetuned (s=64, m=0.5) on the same combined dataset as v5b (CzechLynx + AnimalCLEF lynx, 346 IDs): DINOv2-ViT-B/14 [32, 33] at 25 epochs and at 40 epochs, both at 518×518; DINOv2-ViT-L/14 at 25 epochs (518×518, ∼300 M parameters); and ConvNeXt-Small/Large [34] at 40 epochs, 384×384. We picked DINOv2-ViT-B at 25 epochs; per-candidate r and rescue values and the rationale for this choice are reported in Section 5.2. We also tested and rejected MegaDescriptor [12] as a secondary (calibration overfits training), 256-pixel self-trained MiewID variants (resolution mismatch), and the Leiden CPM [35] community-detection algorithm as an alternative clustering step on the same blend graph (gave $+0.021$ train ARI but -0.032 on competition; HAC is the competition-optimal partitioner for this similarity structure).

The two embedding-derived cosine similarity matrices are blended as

$$S_{\text{Lynx}} = w \cdot S_{\text{v5b}} + (1 - w) \cdot S_{\text{DINO}}, \quad w = 0.65, \tag{2}$$

and clustered with average-linkage hierarchical agglomerative clustering (HAC) on the pair-wise distance matrix $D = 1 - S_{\text{Lynx}}$ (this $S \mapsto 1 - S$ convention is used for all per-species HAC distances in the paper). The single HAC hyperparameter we tune is the cut threshold τ on the linkage dendrogram — merges with $D < \tau$ are accepted, larger τ produces fewer/larger clusters — which together with the data determines the final cluster count K_{test}. We calibrated τ to a target K_{test} via repeated Kaggle public-score probes (∼30 submissions, varying τ while holding the other species fixed) and locked K_{test}=87, which sits on a wide public plateau spanning $K \in [85, 90]$; the full K_{test}-vs-public-ARI sweep and its private counterpart are reported in Section 5.1 (Figure 5).

4.2. SalamanderID2025: local-feature matching with a self-trained viewpoint rescuer

Fire salamanders are individually identified from their dorsal yellow-spot patterns; prior automatic photo-ID for this species has relied on non-deep-learning matchers [36]. The competition data adds two further challenges: the spot pattern is not perfectly stable across years (cross-year same-individual pairs visibly drift), and nearly half (48%) of same-individual pairs in the training set involve different viewpoints, dorsal vs. profile (based on the dataset's per-image viewpoint labels). Standard global embeddings (MiewID, DINOv2, MegaDescriptor) saturate at AUC ≈ 0.5 on cross-year cross-viewpoint pairs. The subpipeline therefore rests on three layers.

Figure 2: Salamander *body*-pattern preprocessing, shown on a representative *dorsal* training image. These are the four body-crop stages; the parallel head-crop extraction (also part of preprocessing, described in the text) is not shown. **(a)** Raw photograph. **(b)** SAM3 segmentation with the text prompt "salamander" (non-mask pixels replaced by black). **(c)**PCA-rotated body: the principal axis of the mask is aligned horizontally, with the head placed to the left. **(d)** Yellow-pattern image: LAB b^* Otsu threshold isolates yellow spots; the original RGB is preserved inside the spot mask and zeroed elsewhere. The yellow-pattern image is the input to SuperPoint+LightGlue for body-pattern matching.

Preprocessing: body and head. Each image is segmented with SAM3 [10] using the text prompt "salamander". The resulting body mask is then PCA-oriented such that the principal axis is horizontal and the head is at the left; head side is determined by comparing per-end thickness (since the head is wider than the tail). Then, the yellow-spot pattern is extracted via Otsu thresholding on the LAB b^* channel followed by morphological cleanup, dilation by five iterations, and re-application to the original RGB to preserve local edge texture for local-feature matching. Figure 2 illustrates the four stages.

We also extract a head crop for *every* image (both viewpoints) via a second SAM3 pass with the prompt "head of fire salamander". SAM3 finds a mask for roughly half the images, and the remainder fall back to a geometric crop of the leftmost fraction of the head-left–oriented body bounding box. That fraction is viewpoint-aware: we read the dataset's dorsal/profile label and crop the leftmost 25% for dorsal images and 45% for profiles, since the head occupies a larger share of the dorsal images.

Local-feature matching and cluster seeding. We run two independent all-pairs SuperPoint [14]+LightGlue [18] passes (up to 800 keypoints), one on the body yellow-pattern images and one on the head crops, each retaining a per-pair match count. Both passes run on every image regardless of viewpoint; for profile images the body pass additionally matches a $180°$-rotated copy and keeps the larger count, which handles cases where preprocessing misidentified head versus tail. The two match-count matrices, body and head, supply cluster seeds: pairs with body match count ≥ 15 or head match count ≥ 20 are merged greedily into connected components. Both matchers work well within a viewpoint but collapse across it: recall on cross-viewpoint same-individual pairs is only 2.9%; this is

the system's dominant blind spot. Figure 3a shows same-viewpoint same individual matches.

Cross-viewpoint rescue via self-training. We address the cross-viewpoint gap with a ConvNeXt-Small [34] rescuer trained jointly on training labels and on pseudo-labels harvested from the connected components of a looser baseline seed set (body match count ≥ 13 or head match count ≥ 7, below the production merge thresholds; training settings in Table 2). The pseudo-label set adds 93 test-only multi-image classes (325 test images, all from connected components of size ≥ 2) to the 587 training individuals. To validate pseudo-label quality, we (i) checked purity by comparing the connected-component partition against ground-truth labels on training pairs, and (ii) withheld a subset of labeled training images from pseudo-labeling to serve as a held-out validation set for the rescuer; both results are reported in Section 5.2. On this evidence we accepted each connected component as a single identity during training–no relabeling, no noise filtering of the pseudo-labels. At inference, any cross-viewpoint pair with ConvNeXt cosine similarity ≥ 0.55 is added to the seed set (alongside the SuperPoint thresholds above) and merged into the same connected components before HAC. Final clustering is HAC average linkage at distance threshold 0.97, calibrated to $K_{\text{test}} = 559$ via Kaggle public-score probing.

We tested and rejected several alternatives: ALIKED [15], DISK [16], and LoFTR [17] local-feature matching (precision $\sim$ 1% on cross-viewpoint salamander spots), gradient-reversal-layer cross-viewpoint finetuning (high train precision but cascading test failures), midline-coordinate transforms (subsumed by PCA viewpoint), and head-only ArcFace finetuning (overfit to training individuals). Self-training with cross-viewpoint pseudo-labels was the single positive lever beyond the SuperPoint baseline.

4.3. SeaTurtleID2022: dual-embedding blend with complete linkage

For loggerhead sea turtles, MiewID alone is already strong (same-ID mean similarity 0.687 vs. different-ID 0.426); unsurprisingly so, since the SeaTurtleID dataset [27] was part of the corpus on which MiewID was originally trained [11]. To add a complementary signal we finetune a copy of MiewID with ArcFace on the union of SeaTurtleIDHeads [27] and ZindiTurtleRecall [28] datasets ($\sim$14k external images in total), and blend the result with the off-the-shelf MiewID-msv3 [30] embedding using horizontal-flip test-time augmentation (TTA). This approach relies on an inherent visual similarity between left and right head profiles of loggerhead sea turtles, where facial colouration, pigmentation and texture are common throughout the turtle head; such characteristics enable deep embeddings to learn representations independent of view orientation [37]. The competition training images are deliberately *not* included in the finetune set: a pilot finetune on the competition-only training images overfit to its in-distribution identities (held-out mAP@R 0.61 vs. off-the-shelf 0.78) and scored worse on public leaderboard, so the final pipeline relies on external data alone for the finetune contribution. The blend is

$$S_{\text{ST}} = (1 - w_{\text{FT}}) \cdot S_{\text{MiewID}} + w_{\text{FT}} \cdot S_{\text{FT}}, \quad w_{\text{FT}} = 0.25, \tag{3}$$

followed by HAC *complete*-linkage at threshold 0.78. Before running HAC we apply a *transitivity-seed* pre-merge: any chain of pairs $A{-}B{-}C{-}\ldots$ in which each consecutive pair has MiewID-blend similarity ≥ 0.80 is collapsed into a single locked cluster, recovering within-individual chains that complete linkage would otherwise fragment (the high threshold makes this pre-merge essentially noise-free on training-side validation). Sea turtle is the only species for which complete linkage was empirically optimal in our screening; we compare it against average and single linkage on a held-out training-split clustering evaluation in Section 5.2. We attribute the asymmetry to dense within-individual neighborhoods on a small test set (500 images, an estimated $\sim$176 distinct individuals based on the per-species K_{test} we eventually targeted) where the cluster cores are reliable but the boundaries are vulnerable to chain merges.

4.4. TexasHornedLizards: ALIKED + LightGlue + RANSAC

The Texas horned lizards subset has no training labels and 274 test-only ventral hand-held photographs, so we cannot evaluate any per-subset choice on the AnimalCLEF data alone. As a stand-in for local validation we use TCU, an external labeled dataset of the same species: 277 ventral images of 104 individuals, identified by Biffi et al. [38] using genotypes and ventral spot patterns. We report clustering ARI on TCU.

All off-the-shelf global re-ID embedding models we tested are at chance on TCU: MiewID ARI $=$ 0.022, MegaDescriptor ARI $= 0.014$, DINOv2 ARI $= 0.004$ (chance level for the corresponding K_{test}). We therefore abandoned neural embeddings and used local-feature matching exclusively: ALIKED keypoints [15] with LightGlue [18] matching and RANSAC [19] geometric verification, on inner-cropped MegaDetector boxes (70% height $\times$ 90% width to suppress hand contamination at the frame edge). We observed the same pattern in our recent re-identification study of the Hula painted frog (*Latonia nigriventer*) on close-range ventral photographs of spotted skin [39]: global re-ID embeddings near chance, ALIKED+LightGlue carrying the productive signal. In both settings the discriminative information is concentrated in small high-frequency pattern features that local matchers preserve but generic embeddings average out. HAC average linkage at distance threshold 0.40 produced $K_{\text{test}} = 205$. We calibrated K via Kaggle public-score probing: $K = 184$ scored -0.011 vs. $K = 205$, $K = 210$ scored -0.025, isolating $K = 205$ as the optimum.

A counter-intuitive finding here was that SAM-based holding-hand removal *hurt* performance, possibly because it stripped geometric scaffolding around the grip that ALIKED relied on to find consistent matches between same-individual images held similarly. We also tested an ALIKED + SuperPoint + DISK multi-matcher fusion, but it did not beat ALIKED alone at $K_{\text{test}} = 205$ on this subset. A representative ALIKED+LightGlue+RANSAC match on a same-individual Texas pair is shown in Figure 3b.

4.5. Final-five portfolio selection

AnimalCLEF 2026 allowed five final submissions per team. Across the competition we made 154 scored submissions to the public leaderboard. The final-five selection was scoped to submissions that clustered tightly between 0.79 and 0.82 ARI on public ARI, as earlier submissions had been explorations of configurations we had already discarded. We selected five for diversity rather than for raw public score, as follows.

Pairwise-disagreement audit. Let $\pi^{(s)} \in \{1, \ldots, K^{(s)}\}^N$ be the test-image partition produced by submission s, where N is the number of test images. For two submissions s, s' we count the number of test-image *pairs* in which they disagree about whether the two images belong to the same cluster:

$$D(s, s') = \left|\{(i, j) : i < j,\ \mathbf{1}[\pi_i^{(s)} = \pi_j^{(s)}] \neq \mathbf{1}[\pi_i^{(s')} = \pi_j^{(s')}]\}\right|. \tag{4}$$

D is symmetric, ranges in $[0, \binom{N}{2}]$, and equals $1 - \mathrm{RI}(\pi^{(s)}, \pi^{(s')})$ up to normalization. We computed D for every pair of the high public ARI submissions and used it as the diversity signal for portfolio selection.

Selection procedure. We took the top-public-score submission as the anchor (sub#250) and added each next member by greedy maximum–minimum disagreement (the candidate with the largest $\min_{s' \in P} D(s, s')$ over the current portfolio P), restricted to candidates within 0.025 public of the anchor. The boundary 0.025 was chosen ad hoc to admit at least one "no self-train" hedge (sub#216 at public 0.80079); this boundary was the main source of our portfolio's miss, as discussed in Section 6. The final five were:

- sub#250 (public 0.81876): Lynx at $K_{\text{test}} = 80$, self-trained salamander pipeline, sea turtle at $K_{\text{test}} = 176$. Our top public submission; we made it the portfolio anchor. Lynx $K = 80$ was a narrow public-score peak, discovered with one day left.

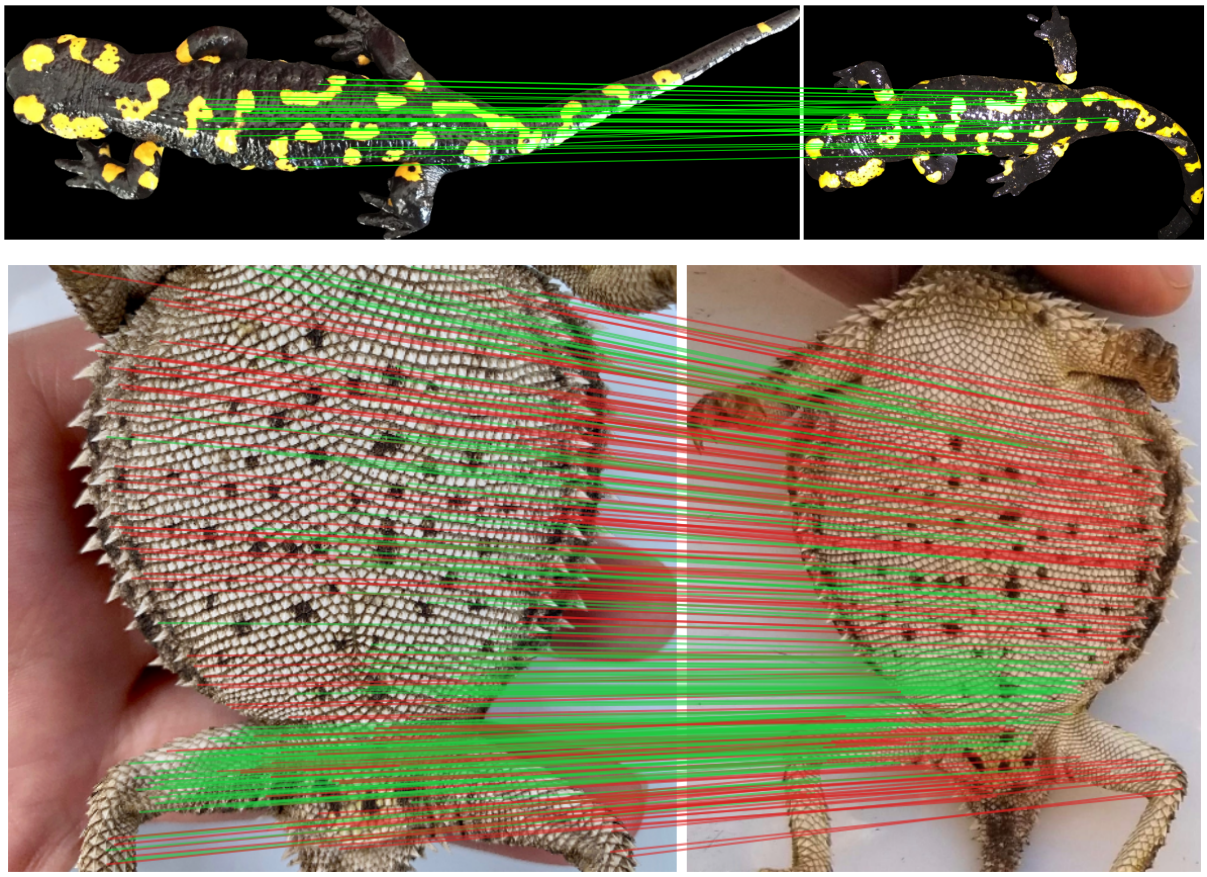

Figure 3: Representative keypoint matches from local-feature matching pipelines. **(a) Salamander** (top panel): two dorsal images of the same individual, photographed two years apart. SuperPoint+LightGlue returns 34 matches on the extracted yellow spots; the images displayed are the PCA-rotated body crops shown *before* the yellow-pattern extraction step, included for visual clarity. In the production pipeline, matching runs on the extracted yellow pattern, not the raw body crop, because SuperPoint on raw photos returns zero high-confidence matches: background keypoints dominate the top-K selection. **(b) Texas horned lizard** (bottom panel): best cross-year match in the TCU external calibration set (images 2019-HL16 and 2018-HL3), selected as the highest-inlier pair across all cross-year pairs. The dataset carries no identity labels; this is a putative same-individual match based on visual inspection. ALIKED+LightGlue+RANSAC returns 185 inliers (green) out of 331 raw matches (red = rejected by RANSAC). Many inliers fall on the near-periodic ventral scale grid rather than individually discriminative spots.

- sub#231 (public 0.81655): same configuration as sub#250 but with Lynx at $K_{\text{test}} = 87$, which sits inside a broader public-score plateau. Selected as a hedge against the narrow $K = 80$ peak being a sampling artifact.
- sub#216 (public 0.80079): Lynx at $K_{\text{test}} = 87$, salamander pipeline without the ConvNeXt self-training rescuer, sea turtle at $K_{\text{test}} = 176$. Selected as a hedge against self-training being overfit.
- sub#196 (public 0.79897): same as sub#216 but with the salamander head crop using SAM3 prompt "head of salamander" (47% success vs. 96% for the "head of fire salamander" prompt) and a uniform 25% geometric fallback rather than the orient-aware 25%/45% fallback. Selected as a hedge against the head-crop design.
- sub#199 (public 0.79040): Lynx pushed lower still to $K_{\text{test}} = 78$, salamander pipeline without self-training. Selected as a compound hedge against both the Lynx K-mode being inflated and self-training being overfit.

The portfolio was hedged against the two risks we anticipated: that self-training (a public win of $+0.016$) might be overfit, and that the Lynx K-mode at 80 might be a sampling artifact. Both hedges

activated on private and sub#231 carried our final score (Table 8).

Code availability

The code for our submission (sub#231) is available at https://github.com/yoavram-lab/AnimalCLEF-2026-sub231. Dependencies and software versions are specified in the repository's `pixi.toml` environment file. Some model checkpoints are available on HuggingFace at https://huggingface.co/yoavram-lab/AnimalCLEF2026-sub231. Training was performed on an NVIDIA A6000 GPU. Training and inference each required up to 24 hours per species. Random seed policy is documented in the source code.

5. Results

5.1. Calibration: public-score probing and train-side proxies

Within a fixed pipeline configuration, public ARI is sensitive to the inferred cluster count K_{test} (number of individuals) on each subset: Lynx $K_{\text{test}} \in [80, 87]$ spans $\sim$0.017 in public ARI; Salamander has a sharp peak at K_{test}=569 with K_{test}=576 dropping $\sim$0.013; Texas K_{test}=205 vs. K_{test}=184 and K_{test}=210 spans 0.025. The Lynx K_{test}-vs-public-ARI curve was, in our final-week probes, apparently bimodal (Figure 5): a wide plateau around K=87 (sub#231, public 0.81655) and a narrow second peak at $K \in [80, 82]$ (sub#250, public 0.81876, our public-best). The corresponding private curve, however, is unimodal (Figure 5): K=87 peaks at private 0.70050 while K=80 collapses to private 0.69193, the largest public-private gap of any of our final-phase submissions. The narrow K=80 mode was sampling noise on the public split, not a structural feature of the lynx clustering, a result we return to in the portfolio-selection discussion (Section 4.5).

We tried two train-side proxies for public ARI: *train ARI* computed by clustering the labeled train images with the same pipeline, and a *rescued* diagnostic counting hard same-individual pairs that a candidate model brings into a baseline's top-K neighborhood. Both were useful for coarse direction but appeared, in our competition-time per-K probes, to fail to track the public-score optimum at fine scale: for lynx, train ARI peaked at K_{test}=108 while public peaked at K_{test}=87; for salamander, train ARI was effectively equal across the head-only K_{test}=568/569 plateau yet a K_{test}=561 variant with 100% test-test seed precision scored -0.001 on public. We therefore used K_{test} calibration via Kaggle public-score probing as our primary tuning signal, with the train-side proxies as coarse filters; this is the calibration regime that produced the pipelines documented in Section 4 and that we evaluate in the remainder of this section.

5.2. Component-screening results

Lynx secondary screening. We screened candidate secondary embeddings by their pair-wise correlation r with primary embedding v5b and by rescue local ARI (Section 4.1): positive rescue means the candidate moves genuine same-ID pairs closer in the blend, negative rescue means it displaces them with impostors. DINOv2-ViT-B finetuned for 25 epochs had the lowest correlation ($r = 0.489$) among rescue-positive candidates (rescue $= +58$), and we picked it. The larger DINOv2-ViT-L was more decorrelated (r=0.474) but rescue-negative (-20); DINOv2-ViT-B at 40 epochs and ConvNeXt-Small were both more correlated with v5b ($r \geq 0.50$). Training mAP@R kept rising with capacity, but rescue peaked at ViT-B/25-epoch and turned negative beyond it, i.e., capacity helped single-model ranking but not the blend.

Salamander self-training rescue. The ConvNeXt-Small self-trained rescuer (Section 4.2) was finetuned from the ImageNet-pretrained backbone on connected-component pseudo-labels derived from the body and head SuperPoint+LightGlue seed set (Section 4.2). We audited the trained rescuer on labeled cross-viewpoint training pairs at cosine similarity threshold 0.5: it assigns similarity > 0.5

Table 4
AnimalCLEF 2026 final leaderboard (top-five private ARI). Public→private gap is private minus public ARI; negative values indicate overfitting to the public split.

Rank	Team	Public ARI	Private ARI	Gap
1	M.I.A	0.72124	**0.70393**	-0.01731
2	**Ram Lab (this work)**	**0.81876**	0.70050	-0.11826
3	VIPL-VSU	0.67987	0.69632	$+0.01645$
4	HSU Lab	0.75056	0.67537	-0.07519
5	DS@GT LifeCLEF	0.73275	0.67424	-0.05851

to $1{,}019$ of $1{,}041$ same-individual cross-viewpoint pairs (98% recall) and to only 26 of 4.69×10^5 different-individual cross-viewpoint pairs (0.006% false-positive rate). On a held-out subset of training images excluded from pseudo-labeling and used as pseudo-test, the trained rescuer added 54 direct cross-viewpoint merges at 97.5% precision, closing the dominant SuperPoint blind spot (2.9% cross-viewpoint recall from SuperPoint alone).

SeaTurtle linkage ablation. We compared HAC linkage choices on the labeled sea-turtle training split, clustering the per-species similarity blend at threshold 0.78 (Section 4.3). Train ARI was 0.035 with average linkage (catastrophic chain merging), < 0.01 with single linkage, and 0.85 with complete linkage.

5.3. Final standings

We finished second by 0.00343 in private ARI (Table 4). Among the top three teams we had the largest negative public-to-private gap; first place passed us by ≈ 0.003 on private despite being below us by ≈ 0.10 on public. The same five teams hold the top-5 on both leaderboards, but the order shuffles markedly: M.I.A. moves from 4th on public to 1st on private; Ram Lab from 1st to 2nd; VIPL-VSU from 5th to 3rd; HSU Lab from 2nd to 4th; and DS@GT LifeCLEF from 3rd to 5th. Thus, the public leaderboard predicts the top-5 subset but not the within-subset order. VIPL-VSU is the only top-5 team with a *positive* public-to-private gap ($+0.016$): they appear under-fit on public, the opposite failure mode from ours. The top three teams are within 0.00761 of each other on private (our deficit to first is 0.00343, our margin over third is 0.00418); the whole top-three sits inside a band where about 3–4 test images flipping cluster membership separate adjacent ranks — exactly the saturation regime we examine in Section 6.

5.4. Lever-by-lever attribution

The official AnimalCLEF 2026 organizer baseline (Starter notebook 2 [4]) used an uncalibrated MegaDescriptor-L-384 + MiewID-msv3 cosine-similarity fusion clustered with HDBSCAN per species, and scored ARI ≈ 0.203 on the public leaderboard. Our first per-species submission (a per-species hybrid of MiewID with homography-consensus matching) scored public 0.549 and private 0.595; our final pipeline scored public 0.819 and private 0.701. The within-per-species refinements described in Section 4 therefore account for $+0.27$ on public and $+0.11$ on private over our own first per-species submission.

Across the competition we made 154 scored submissions; the final 13 active days produced 51 candidates that explore the levers in Table 5. The final week produced $+0.049$ on public but only $+0.011$ on private; only the Sal head-crop upgrade (fire-salamander SAM3 prompt + orient-aware fallback) with body 180° TTA, and a $K = 555$ looser-seed-threshold variant we did not select for the final portfolio, had private gain $\geq$ public gain. Lynx K_{test} reduction and sea-turtle K_{test} reduction produced approximately half their public gain on private, self-training was approximately neutral on private, and the K=80 narrow Lynx mode was *negatively* predictive on private.

Table 5
Lever-by-lever attribution of public and private gain across submitted variants. Each row was a separate Kaggle submission that explored a specific lever; the rightmost column is our retrospective judgment. Our submitted private-best score (sub#231 at private 0.70050, second place) accumulates the first three rows. Self-train and the $K{=}80$ mode were the two largest contributors to our public lead and the two largest sources of public-private slippage. The "Sal head-crop upgrade" refers to the salamander head pipeline of Section 4.2 (SAM3 prompt "head of fire salamander" and orient-aware 25%/45% geometric fallback), replacing the earlier "head of salamander" prompt and uniform 25% fallback.

Lever	Public Δ	Private Δ	Judgment
Lynx K_{test} reduction	$+0.029$	$+0.017$	mostly real
Sea-turtle K_{test} reduction	$+0.012$	$+0.005$	half real
Sal head-crop upgrade + body 180°	$+0.002$	$+0.005$	real (private $>$ public)
Self-trained Sal cross-viewpoint	$+0.016$	$+0.0006$	mostly noise
Lynx $K = 80$ narrow mode	$+0.002$	-0.0086	pure noise
Looser Sal seed thresholds + $K = 555$	-0.0005	$+0.006$	real, *not selected*

Table 6
Spearman correlation between public and private ARI across subsets; visualized in Figure 4. Bootstrap 95% CI from 10,000 resamples; small-n subsets have wide intervals that span zero. The all-teams correlation is strong; correlations within high-public subsets are not robustly distinguishable from zero (or each other) at this sample size.

Cohort	n	Spearman ρ	95% CI
All teams	232	$+0.953$	$[+0.94, +0.97]$
Top 10 teams by public	10	$+0.842$	$[+0.32, +0.99]$
Top 5 teams by public	5	-0.100	$[-1.00, +1.00]$
Our scored submissions, all	154	$+0.948$	$[+0.92, +0.96]$
Our public ≥ 0.78 subset	22	$+0.441$	$[+0.02, +0.81]$
Our public ≥ 0.81 subset	11	-0.454	$[-0.74, +0.49]$
Our top 10 submissions	10	-0.539	$[-0.78, +0.46]$
Our top 5 submissions	5	-0.500	$[-0.50, +1.00]$

6. Lessons Learned at the Top of the Leaderboard

6.1. Public ARI predicts private — except at the top

We obtained the full public and private leaderboards (232 teams) and computed Spearman correlations across team-level and submission-level subsets (Table 6; Figure 4). Across all 232 teams, public ARI is a strong predictor of private ARI ($\rho = +0.953$). The correlation is also high within our scored submissions ($\rho = +0.948$). However, the relationship is subset-dependent: at progressively higher public-score subsets the point estimate of ρ moves through zero and becomes negative.

The pattern is positive at full population and attenuates to zero or negative at the top. This is consistent with adaptive overfitting on a held-out public split [23, 24]. It also matches prior reports that top-of-leaderboard ranks partially shuffle when the test set is re-collected [25]. This is directly visible in the left panel of Figure 4: the top-five teams are ordered differently along the public (x) and private (y) axes. We do report that, in this competition, our late-stage optimization gain was concentrated in the high-public-score submission subsets where the public-private rank correlation is plausibly zero. Concretely, on our ≥ 0.81 public ARI submissions ($n = 11$) the public-best submission ranked near the bottom by private, with $\rho = -0.454$ and a wide confidence interval of $[-0.74, +0.49]$.

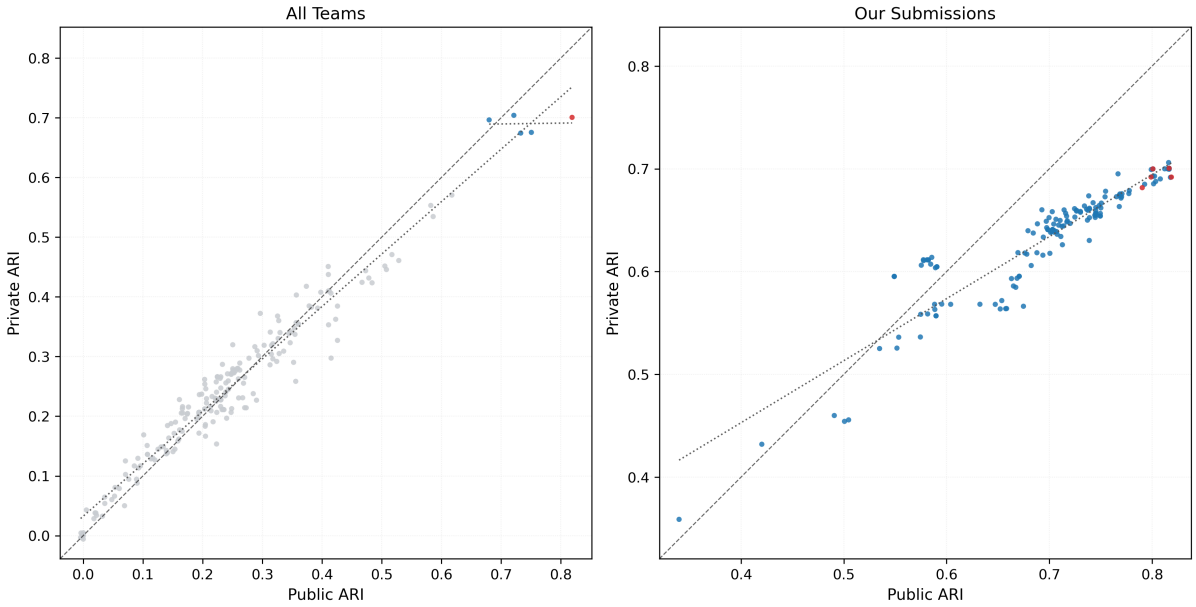

Figure 4: Public versus private leaderboard ARI on AnimalCLEF 2026, shown at two granularities: all 232 competing teams (left panel) and each of our 154 scored submissions (right panel). The dashed line in both panels is $y = x$. **Left:** Ram Lab (red) has highest public ARI and the second-largest drop on private ARI. The other top-five teams are in blue, and rest of the 226 teams in grey. The dotted line is the linear fit across all 232 teams (private $= 0.033 + 0.878 \cdot$ public) with Spearman correlation $\rho = +0.953$. The short nearly-flat dotted line at the upper right is the fit restricted to the top-five teams (private $= 0.68 + 0.014 \cdot$ public), with Spearman correlation $\rho = -0.10$ (Table 6). **Right:** Red points are the five locked portfolio picks (sub#250, sub#231, sub#216, sub#196, sub#199) and blue points are all other submissions, including sub#238 (public 0.81606, private 0.70608), our hindsight private winner that sat just outside the diversity-audit boundary and was not selected. The dotted line is the linear fit across all 154 submissions (private $= 0.211 + 0.604 \cdot$ public), with Spearman correlation $\rho = +0.948$ (Table 6).

6.2. Train-set metrics predict private ARI where public ARI stops

The public-private correlation collapse motivated a post-competition analysis. We collected per-submission private scores for our 50 top-public-ARI submissions and reconstructed the Salamander cluster partition that produced each one. For 13 of the 15 unique top-50 K_{test} values the reconstruction matches the saved competition submission CSV exactly (every test image lands in the same cluster as in the saved CSV; $n = 44$ submission rows); the remaining two match the saved CSV closely but not exactly (test-side partition ARI ≥ 0.97 but not $= 1$). We then computed, for each submission, several supervised clustering metrics on the Salamander train set (1,388 images, 587 true identities), and asked which ones track the private leaderboard score.

The headline finding: in the top-20 submissions by public ARI (public ≥ 0.799) the public-vs-private Spearman correlation collapses to $\rho = +0.41$ ($p = 0.09$, NS; Table 6), but Salamander train ARI retains a significant positive correlation with private ARI, $\rho = +0.53$ ($p = 0.02$). A joint linear model predicting private from public alone gives $R^2 = 0.23$; adding Salamander train ARI doubles R^2 to 0.47. Salamander train ARI is *complementary* to public ARI in this regime, not a replacement, and its conditional contribution is largest precisely where public's contribution stops.

Beyond train ARI we computed twelve other supervised clustering metrics on the same train set (Table 7). Two patterns stand out. First, AMI, completeness, pair-recall, and BCubed-recall slightly outperform ARI as private predictors ($\rho \approx +0.86$ vs. $+0.83$); these four all reward *identity completeness*, the signal that survives the saturation. Second, homogeneity and BCubed-precision are *negatively* correlated with private ($\rho \approx -0.68$, $p < 0.0001$), because over-segmentation is the dominant error mode in this subset: configurations using the ConvNeXt cross-viewpoint self-training rescuer (sub#231 and its $K = 559$ variants) avoided it; configurations without the rescuer ($K = 576$–579; head & body

Table 7
Spearman correlation of supervised clustering metrics (computed on the Salamander *train* set) with public and private leaderboard ARI on the top 50 submissions in our archive by public ARI plus the rank-51 entry ($n = 44$ submissions whose reconstructed Salamander partition matches the saved competition submission CSV exactly; see Section 6.2). Bold rows: metrics that beat ARI as a private predictor. Italic rows: metrics negatively correlated with private. The metrics that lead are the ones that penalize splitting a true identity across multiple predicted clusters (AMI, completeness, pair-recall, BCubed-recall); the metrics that go the wrong way are the ones that reward cluster *purity* (homogeneity, BCubed-precision), because in this subset over-segmentation is the dominant error mode and rewarding purity rewards the wrong direction. Pair-precision and mean cluster-purity are uninformative here.

Metric (on Sal train set)	ρ vs. public	ρ vs. private
ARI	$+0.92$	$+0.83$
AMI	$+0.91$	$\mathbf{+0.86}$
completeness	$+0.91$	$\mathbf{+0.86}$
V-measure	$+0.92$	$+0.83$
Fowlkes–Mallows	$+0.92$	$+0.81$
Pair-F1	$+0.92$	$+0.83$
Pair-recall	$+0.91$	$\mathbf{+0.86}$
BCubed-F1	$+0.92$	$+0.83$
BCubed-recall	$+0.91$	$\mathbf{+0.86}$
homogeneity	-0.74	-0.68
BCubed-precision	-0.74	-0.68
Pair-precision	$+0.09$ (NS)	$+0.06$ (NS)
Mean cluster purity	-0.17 (NS)	-0.17 (NS)

Table 8
Submitted variants near the top of our archive. Rows 1–5 are the five submissions we finalized as our portfolio (sub#231, the row in bold, was our submitted second-place result at private 0.70050); the bottom row is sub#238, the configuration we built and submitted to the public leaderboard but did not lock as a final entry. Hindsight rank is over our 51-submission final-phase archive. Tables 2–3 document the row in bold. "Head-crop upgrade" is the salamander head pipeline described in Section 4.2 (SAM3 with the prompt "head of fire salamander" and orient-aware 25%/45% geometric fallback); "baseline head crop" is the earlier configuration ("head of salamander" prompt and uniform 25% fallback).

Sub	Configuration	In portfolio?	Public	Private	Hindsight rank (private)
#250	Lynx $K=80$ + self-train	yes (anchor)	0.81876	0.69193	10 of 51
#231	**Lynx $K=87$ + self-train**	**yes (scoring row)**	**0.81655**	**0.70050**	**2 of 51**
#216	no self-train, head-crop upgrade	yes	0.80079	0.69990	7 of 51
#196	no self-train, baseline head crop	yes	0.79897	0.69199	9 of 51
#199	Lynx $K=78$ + no self-train	yes	0.79040	0.68165	22 of 51
#238	looser seed thresholds, $K = 555$	no (unselected)	0.81606	0.70608	1 of 51

SP+LG seeding only, so cross-viewpoint same-individual pairs remained unmerged) fell into it.

6.3. The unselected winner

Within our final five submissions, the public-best (sub#250) was *fourth* of five by private ARI. A pipeline we had submitted but not selected for the final five, sub#238, would have ranked first among all teams' selected final submissions by 0.00215 on the private leaderboard. Table 8 contrasts our final five with the unselected sub#238 and the public-best.

Sub#238 used the same salamander pipeline as sub#231 (our locked entry) but with two coupled changes: looser seed thresholds (head SP match count ≥ 15 vs sub#231's ≥ 20, and cross-viewpoint ConvNeXt similarity ≥ 0.50 vs ≥ 0.55) and a HAC threshold targeting $K = 555$ instead of $K = 559$. On public it was -0.0027 below sub#250 and looked like a slightly inferior member of a plateau. On

private it was $+0.0056$ above the best of our final five. Why we missed it: our pairwise-disagreement audit was scoped to the $K = 557$–559 self-train plateau, and sub#238 sat outside that boundary ($K = 555$, looser thresholds). The corrective lesson is twofold: the audit should have included every submission within ~ 0.01 public of the best (not only those inside the immediate K-plateau), and the tie-breaker between candidates in that window should be a train-set supervised metric rather than public ARI, as Section 6.2 shows.

Sub#238 does not stand out on either runtime feature axis: it sits mid-plateau on public ARI and inside the (tightly packed) top cluster on Salamander train ARI. In hindsight (with private ARI in hand) it has the largest positive residual in a joint public-plus-train-ARI regression: its private exceeded what the model predicts from its public and train ARI. But that residual is post-hoc, computable only after the deadline. The runtime-actionable correction is therefore to *widen the diversity-audit boundary* (Rule 4 below) so that sub#238 enters the candidate pool. We do not claim that any metric we computed would then have uniquely selected it from that pool.

6.4. Four exploratory decision rules

These are not validated rules; they are rules of thumb suggested by the findings above and consistent with prior held-out-test-set overfitting findings [23, 24, 25]. They would benefit from retrospective validation on prior AnimalCLEF or comparable competitions.

1. *Within 0.01 of the public best, distrust public.* The Spearman correlation between public and private has wide CIs in this range; in this competition the order in our top-11 was anti-correlated with private order.
2. *Within that same ± 0.01 window of the public best, fall back to a train-set supervised clustering metric to break ties.* Public ARI loses predictive power for private inside this top window, but a supervised metric on the train set retains a significant positive correlation with private there ($\rho = +0.53$, $p = 0.02$ on our top-20 subset; Section 6.2). Caveat: when the top candidates cluster tightly on train ARI (as ours did), train ARI alone will not uniquely separate them either — in that case the protective lever is the audit boundary (Rule 4), not a clever metric. Prefer AMI or completeness over ARI as the train-set signal (slightly higher correlation with private at no extra cost), and treat homogeneity and cluster-purity as a *warning sign* rather than a target, since over-segmentation fragments true identities and these metrics reward the wrong direction in this regime.
3. *Late-discovered narrow modes are sampling artifacts until proven otherwise.* The Lynx $K = 80$ peak was discovered with one day left, spanned ± 2 values, and produced the single largest public-private gap in our archive. Wider plateaus (≥ 5 values, the $K = 87$ mode) translated; the narrow one did not. Figure 5 contrasts the K-vs-public and K-vs-private profiles directly.
4. *Pairwise diversity audits should sweep, not snap.* Any submission within 0.01 of the best publc ARI should be evaluated by prediction disagreement, not by whether it falls inside a predefined plateau cluster. We would set the audit boundary at ± 0.01 on public and rank candidates by maximum-minimum disagreement against the current portfolio, with rule 2's train-set supervised metric as the tie-breaker within that window.

7. Discussion

7.1. Per-species heterogeneity

The cumulative gain over a uniform baseline came from picking the right tool per dataset, not from a single dataset-agnostic improvement: a finetuned-embedding blend for lynx, local-feature matching with a self-trained rescuer for salamander, dual-embedding with complete linkage for sea turtle, and local features only for Texas.

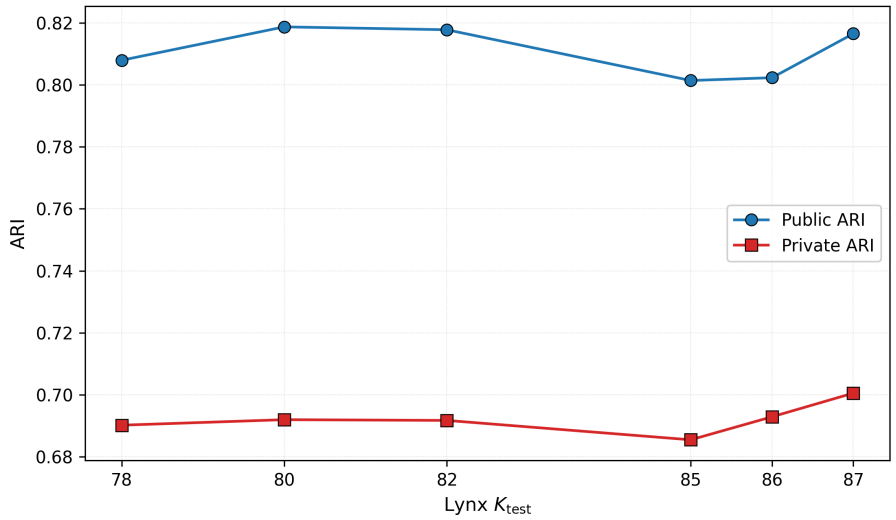

Figure 5: Effect of Lynx K_{test} on scored ARI, with all other levers held fixed (self-trained salamander + sea turtle K_{test}=176 + Texas pipeline). Blue circles trace public ARI (upper band, 0.80–0.82): bimodal, with a narrow peak at K=80 (sub#250, public 0.81876), a dip at $K \in \{85, 86\}$, and a second peak at K=87 (sub#231, public 0.81655). Red squares trace private ARI (lower band, 0.685–0.700): a single unimodal peak at K=87 (sub#231, private 0.70050) and a clear collapse at K=80 (sub#250, private 0.69193, the largest public-private gap of any of our final-phase submissions). The K=80 "second mode" on public was sampling noise on the public split, not a structural maximum — and choosing the highest-public submission as the portfolio anchor ranked it 4th of our 5 picks on private.

7.2. Self-training works as a private-neutral public lever

The self-trained ConvNeXt cross-viewpoint rescuer was the largest single salamander lever for public (+0.016) and approximately neutral for private (+0.0006). We do not interpret this as a failure: the lever closed a real cross-viewpoint gap on training pairs (Section 5.2), and the test-side partitions of subpipelines *without* the rescuer turn out to be very close to those of the self-trained pipeline (sub#216 reaches private 0.6999 without self-training, only 0.0006 below the self-trained sub#231 at 0.7005). Retaining both self-train and no-self-train submissions in the portfolio insured against the possibility that self-training was pure public-overfit; in the event, sub#231 carried us to second place and the hedges proved unnecessary but were cheap insurance.

7.3. Limitations

This paper documents one team's experience with one competition; the public-private decoupling we observed at the top of the leaderboard is a known failure mode of adaptive optimization on held-out test sets [23, 24, 25] but its strength varies across competitions. Our top-5 subset correlation is weakly powered ($n = 5$, CI spanning $[-1, +1]$); the strongest evidence is at the subset level (all 232 teams, $\rho = +0.95$) and within our 51 submissions (which show the high-public attenuation we discuss). Decisions made at competition time were based on point estimates; in retrospect we should have reported the bootstrap CIs alongside.

We also did not run a uniform-pipeline ablation across all four species at the final hyperparameters. Doing so would clarify how much each per-species choice contributes to the final ARI; we report only the headline observation that we adopted four very different subpipelines because we tried single-pipeline alternatives early and they did not match the per-species variants.

The train-set clustering metric result (Section 6.2) carries two caveats of its own. Train ARI measures clustering quality on *seen* training identities, whereas the test set also contains unseen individuals; that it nonetheless predicts private ARI is somewhat surprising and warrants further study. And the analysis is restricted to the reconstructed Salamander partition over our top-50-by-public submissions,

so a four-dataset weighted train ARI and an extension across the full 154-submission archive are both natural follow-ups.

The pairwise-disagreement audit we describe (Section 4.5) is post-hoc in two senses: we did not pre-register the disagreement boundary or the greedy selection rule, and our ± 0.025 public boundary was chosen at the time on a judgment call about which hedges to admit. The Section 6 recommendation to use ± 0.01 instead is a recommendation extracted from this run; it should be validated on independent competition data before being trusted.

8. Conclusion

We placed second at AnimalCLEF 2026 with a per-species ensemble that picks the right tool per dataset — finetuned re-ID embeddings for lynx and sea turtle, local-feature matching with self-trained cross-viewpoint rescue for salamander, and pure local-feature matching for Texas horned lizards — and that calibrates each subpipeline's K_{test} via Kaggle public-score probing. The same Kaggle public-probing strategy that helped us early hurt us at the end: as the public score saturated, additional optimization moved our predictions sideways across noise, our public-best submission ended up the most overfit of our 51 candidates, and our pairwise-disagreement audit excluded the pipeline that would have won. For AnimalCLEF 2027 we would maintain a held-out per-species private proxy throughout the run (one held-out fold per labeled species, ideally temporally-disjoint where the metadata supports it), broaden the portfolio audit boundary to ± 0.01 public around the best, and report bootstrap confidence intervals on any point estimate used to make a decision under time pressure.

Acknowledgments

We thank the AnimalCLEF organizers for organizing the competition. We thank Tsur Herman, Tal Simon, Ofir Levy, and Orr Spiegel for discussions. This work was partially supported by the Safra Center for Bioinformatics and the Center for AI & Data Science, Tel Aviv University (YR).

Declaration on Generative AI

During the preparation of this manuscript, the authors used Claude (Anthropic) for grammar and spelling review and to draft sections of the text from internal reports and submission logs. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Y. Berger-Wolf, D. I. Rubenstein, C. V. Stewart, J. A. Holmberg, J. Parham, S. Menon, J. Crall, J. Van Oast, E. Kiciman, L. Joppa, Wildbook: Crowdsourcing, computer vision, and data science for conservation, arXiv preprint arXiv:1710.08880 (2017).

[2] D. Blount, S. Gero, J. Van Oast, J. Parham, C. Kingen, B. Scheiner, T. Stere, M. Fisher, G. Minton, C. Khan, et al., Flukebook: an open-source AI platform for cetacean photo identification, Mammalian Biology 102 (2022) 1005–1023.

[3] V. Čermák, L. Picek, L. Adam, K. Papafitsoros, AnimalCLEF25 @ CVPR-FGVC & LifeCLEF, https://kaggle.com/competitions/animal-clef-2025, 2025. Kaggle competition.

[4] L. Adam, L. Picek, K. Papafitsoros, D. Williams, D. Biffi, AnimalCLEF26 @ CVPR & CLEF, https://kaggle.com/competitions/animal-clef-2026, 2026. Kaggle competition.

[5] L. Adam, K. Papafitsoros, D. A. Williams, D. Biffi, L. Picek, Overview of AnimalCLEF 2026: Discovery and re-identification of individual animals, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[6] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: AI challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.
[7] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.
[8] W. M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association 66 (1971) 846–850.
[9] S. Beery, D. Morris, S. Yang, M. Simon, A. Norouzzadeh, N. Joshi, Efficient pipeline for automating species id in new camera trap projects, 2019. doi:10.3897/biss.3.37222.
[10] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, et al., SAM 3: Segment anything with concepts, 2025. URL: https://arxiv.org/abs/2511.16719. arXiv:2511.16719.
[11] L. Otarashvili, T. Subramanian, J. Holmberg, J. J. Levenson, C. V. Stewart, Multispecies animal re-ID using a large community-curated dataset, 2024. doi:10.48550/arXiv.2412.05602. arXiv:2412.05602.
[12] V. Čermák, L. Picek, L. Adam, K. Papafitsoros, WildlifeDatasets: An open-source toolkit for animal re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5953–5963.
[13] D. T. Bolger, T. A. Morrison, B. Vance, D. Lee, H. Farid, A computer-assisted system for photographic mark–recapture analysis, Methods in Ecology and Evolution 3 (2012) 813–822. doi:10.1111/j.2041-210X.2012.00212.x.
[14] D. DeTone, T. Malisiewicz, A. Rabinovich, SuperPoint: Self-supervised interest point detection and description, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 224–236.
[15] X. Zhao, X. Wu, W. Chen, P. C. Y. Chen, Q. Xu, Z. Li, ALIKED: A lighter keypoint and descriptor extraction network via deformable transformation, IEEE Transactions on Instrumentation and Measurement 72 (2023) 1–16.
[16] M. J. Tyszkiewicz, P. Fua, E. Trulls, DISK: Learning local features with policy gradient, Advances in Neural Information Processing Systems (NeurIPS) 33 (2020) 14254–14265.
[17] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, LoFTR: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 8922–8931.
[18] P. Lindenberger, P.-E. Sarlin, M. Pollefeys, LightGlue: Local feature matching at light speed, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17627–17638.
[19] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (1981) 381–395.
[20] V. Cermák, L. Picek, L. Adam, L. Neumann, J. Matas, WildFusion: Individual animal identification with calibrated similarity fusion, in: European Conference on Computer Vision, Springer, 2024, pp. 18–36.
[21] B. Rosenberg, M. Zhou, N. Wolf, M. W. Mathis, B. P. Harris, A. Mathis, Individual identification of brown bears using pose-aware metric learning, Current Biology 36 (2026) 645–659.e14. doi:10.1016/j.cub.2025.12.022.
[22] X. Guo, Y. Chen, Y. Guan, H. Wang, T. Wang, J. Ge, L. Bao, Individual identification of wild raptors using a deep learning approach: A case study of the white-tailed eagle, Ecological Informatics 91 (2025) 103379. doi:10.1016/j.ecoinf.2025.103379.
[23] A. Blum, M. Hardt, The Ladder: A reliable leaderboard for machine learning competitions, in: Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, pp. 1006–1014.
[24] R. Roelofs, V. Shankar, B. Recht, S. Fridovich-Keil, M. Hardt, J. Miller, L. Schmidt, A meta-analysis

of overfitting in machine learning, Advances in Neural Information Processing Systems 32 (2019).
[25] B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do ImageNet classifiers generalize to ImageNet?, in: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 5389–5400.
[26] L. Picek, J. Straka, M. Jiřík, E. Belotti, M. Duľa, J. Krausová, M. Bojda, V. Čermák, L. Bufka, R. Dvořák, L. Hrdý, V. Kocourek, J. Labuda, L. Toman, V. Trulík, M. Váňa, M. Kutal, CzechLynx: A dataset for individual identification and pose estimation of the Eurasian lynx, Scientific Data 13 (2026) 511. doi:10.1038/s41597-026-06853-9.
[27] L. Adam, V. Čermák, K. Papafitsoros, L. Picek, SeaTurtleID2022: A long-span dataset for reliable sea turtle re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7146–7156.
[28] Zindi, Turtle recall: Conservation challenge, https://zindi.africa/competitions/turtle-recall-conservation-challenge, 2022. Zindi competition.
[29] D. Steinley, Properties of the Hubert-Arabie adjusted Rand index, Psychological Methods 9 (2004) 386–396. doi:10.1037/1082-989X.9.3.386.
[30] Conservation X Labs, miewid-msv3: Multi-species vertebrate identification model, Hugging Face, https://huggingface.co/conservationxlabs/miewid-msv3, 2024.
[31] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4690–4699.
[32] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., DINOv2: Learning robust visual features without supervision, Transactions on Machine Learning Research (2024). URL: https://openreview.net/forum?id=a68SUt6zFt.
[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[34] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.
[35] V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: Guaranteeing well-connected communities, Scientific Reports 9 (2019) 5233.
[36] L. Schulte, C. Faul, P. Oswald, et al., Performance of different automatic photographic identification software for larvae and adults of the European fire salamander, PLOS ONE 19 (2024) e0298285. doi:10.1371/journal.pone.0298285.
[37] L. Adam, K. Papafitsoros, C. Jean, A. F. Rees, V. Čermák, Exploiting facial side similarities to improve ai-driven sea turtle photo-identification systems, Ecological Informatics 89 (2025) 103158.
[38] D. Biffi, M. R. Tucker, A. Ackel, D. A. Williams, Identification of individual Texas horned lizards (*Phrynosoma cornutum*) using genotypes and ventral spot patterns, Ecology and Evolution 15 (2025) e71167. doi:10.1002/ece3.71167.
[39] M. Yesharim, R. G. B. Perl, U. Roll, S. Gafny, E. Geffen, Y. Ram, Near-perfect photo-ID of the Hula painted frog with zero-shot deep local-feature matching, 2026. doi:10.48550/arXiv.2601.08798. arXiv:2601.08798.