Matching the Baseline, and Why That Is Worth Writing About: A Post-Mortem of an AnimalCLEF 2026 Re-Identification Pipeline*

Deepak Swaroop[1,*], Veena Thenkanidiyoor[1] and Dileep A. D. [2]

[1]Department of Computer Science and Engineering, National Institute of Technology Goa, South Goa, India
[2]Department of Computer Science and Engineering, Indian Institute of Technology Dharwad, Dharwad, Karnataka, India

Abstract

This paper reports a four-stage pipeline submitted to AnimalCLEF 2026 that combined SAM 2 segmentation, a curriculum-fine-tuned MegaDescriptor backbone, MiewID similarity fusion, density-matched threshold calibration, and an XGBoost meta-classifier. The officially ranked entry scored a private Adjusted Rand Index (ARI) of 0.20371, placing 167th of 283 teams and matching the organisers' MegaDescriptor + MiewID + DBSCAN baseline rather than improving on it. This baseline-level outcome is documented here because the component-level analysis carried out to explain it yields findings that are useful to other practitioners: a shared task is served not only by its leaderboard winners but also by careful accounts of which standard components fail to transfer. The paper makes three substantive contributions. A three-epoch head-only warm-up phase reduces the fine-tuning divergence rate from 58% to 0% across 27 runs at negligible additional cost. A closed-form expression relating the optimal clustering threshold to identity density, fitted with $R^2 = 0.94$, provides a principled calibration framework; in this competition it was mis-applied because the test-set density was under-estimated roughly six-fold. Six anti-patterns — recurring, identifiable mistakes that appear reasonable but reliably reduce generalisation — distilled from the negative results together account for most of the gap between an elaborate pipeline and a simple baseline. The paper closes with four concrete suggestions to the organising committee for AnimalCLEF 2027.

Keywords

Wildlife re-identification, AnimalCLEF, Negative results, Metric learning, ArcFace, MegaDescriptor, MiewID, Calibration, Adjusted Rand Index

1. Introduction

AnimalCLEF 2026 [1], run inside the LifeCLEF campaign [2], frames wildlife individual re-identification as open-set clustering and scores submissions with the Adjusted Rand Index (ARI) [15]. This year's edition covers four species with very different data profiles: SeaTurtleID2022 [5] (large reference database), Texas horned lizards [7] (no reference at all), LynxID2025 (moderate reference, augmentable from CzechLynx [8]), and SalamanderID2025 (sparse reference, on average 2.4 images per identity). A total of 283 teams entered. The organisers' own baseline — MegaDescriptor [3] embeddings fused with MiewID [4] and clustered with DBSCAN — reaches ARI ≈ 0.20. The top entries went past 0.70, and more than fifty teams broke 0.30.

An elaborate four-stage pipeline (Figure 1) was built to improve on this baseline: curriculum fine-tuning, similarity-level fusion, database-guided matching, SAM 2 segmentation [10, 11], density-matched threshold calibration, and an XGBoost meta-classifier on pair features. The officially ranked result was a private ARI of 0.20371. A separate submission reached 0.20565 but was excluded from official ranking by Kaggle's top-5 public auto-selection (Section 3). Either way, the pipeline ended at the baseline rather than ahead of it.

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Working Notes contribution. The code, configuration files, calibration sets, training logs, and the artefacts behind each reported negative result are publicly available at https://github.com/onepunchgin/AnimalCLEF26.

*Corresponding author.
ginswaroop@gmail.com (D. Swaroop); veenat@nitgoa.ac.in (V. Thenkanidiyoor); addileep@iitdh.ac.in (D. A. D.)
0009-0003-6835-629X (D. Swaroop); 0000-0002-5354-5915 (V. Thenkanidiyoor); 0009-0000-3319-8103 (D. A. D.)

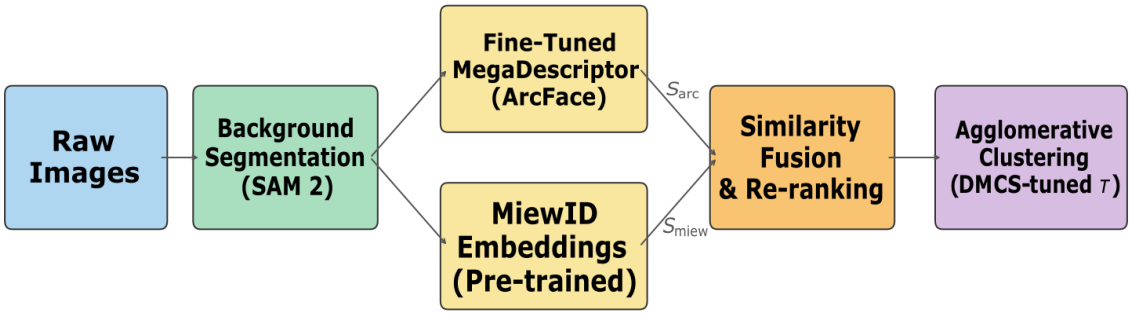

Figure 1: The four-stage pipeline. SAM 2 crops the animal out of its background; a fine-tuned MegaDescriptor produces an ArcFace embedding; an off-the-shelf MiewID embedding is fused at the similarity-matrix level; the fused similarities feed an agglomerative cluster at a threshold calibrated by the density-matched calibration set (DMCS) framework of Section 5.

A working note reporting a baseline-level result remains worth publishing. A shared task is served not only by its leaderboard winners but by analyses of which standard components fail to transfer, so that other groups need not rediscover the same dead ends; achieving a top score is not a precondition for a useful contribution. This paper is therefore organised as a forensic account of a baseline-matching system: a clean description of the pipeline (Section 2), the official outcome and a per-component ablation (Sections 3–4), an analysis of the four components that improved internal validation but did not transfer to the private test set (Sections 5–6), and the settings in which the methodology remains useful despite the leaderboard outcome (Sections 7–8).

What this paper contains. A four-phase fine-tuning curriculum whose Phase 0 head warm-up step eliminates training divergence at zero extra cost (Section 2.2, Table 2). A density-matched calibration set framework that gives a closed-form prediction for the optimal clustering threshold as a function of identity density (Section 5, Eq. 2) — mathematically sound, but mis-applied here because of a six-fold error in the estimate of the test-set density. A $+8.1\%$ gain in private ARI from clustering all four species rather than emitting singletons for the difficult ones (Section 3). Six anti-patterns with quantitative evidence (Section 7). A discussion of the domains where the methodology remains useful in spite of the leaderboard outcome (Section 8). Four recommendations to the organisers for the 2027 edition (Section 9).

2. System

2.1. Architecture

Backbone. A single MegaDescriptor-L-384 [3] ViT-Large [17] produces 1536-d embeddings for every species that has reference data. Each species is fine-tuned separately with SubCenter ArcFace [14], with $K = 3$ sub-centres, scale s ramped from 16 to 64, and angular margin m ramped from 0 to 0.3. Texas horned lizards have no reference data; Section 6 describes what was attempted for them.

Loss and sampling. SubCenter ArcFace cross-entropy with label smoothing 0.1. AdamW [16] with two parameter groups (head and backbone) at a 100:1 learning-rate ratio. A PK sampler with $P = 8$ identities times $K = 4$ images per identity gives a batch of 32 images.

Fusion. Similarity-matrix-level fusion with off-the-shelf MiewID [4]:

$$S_{\text{fused}} = (1 - w)\, S_{\text{arc}} + w\, S_{\text{miew}}, \tag{1}$$

with a per-species w: $w_{\text{turtle}} = 0.6$, $w_{\text{lynx}} = 0.5$, $w_{\text{salamander}} = 0.4$.

Table 1
Fine-tuning schedule. BB% is the fraction of backbone parameters that are trainable; lr_h and lr_b are the head and backbone learning rates.

Phase	ep	BB%	lr_h	lr_b	Scale	Margin
0: Head warmup	3	0	1e-3	—	16	0
1: Top-10%	8	10	1e-3	1e-5	16→32	0→0
2: Top-30%	15	30	5e-4	1e-5	32→64	0→0.3
3: Full backbone	5	100	1e-4	3e-6	64	0.3
4: Cooldown	2	100	1e-5	1e-6	64	0.3

Table 2
Training divergence as a function of head-warmup duration. Divergence is defined as loss $\to \infty$ or feature collapse to a single point within the first five epochs.

Curriculum variant	Runs	Divergence rate
No Phase 0	26	58%
Phase 0: 1 epoch	12	25%
Phase 0: 2 epochs	9	11%
Phase 0: 3 epochs (used)	27	**0%**

Database-guided clustering. For species that have a reference database, each query casts a top-10 weighted nearest-neighbour vote against the database identities. Confident matches are assigned directly; the residual queries are passed to average-linkage agglomerative clustering at a calibrated threshold (the threshold itself is the subject of Section 5). The exact match and clustering thresholds used per species are reported in Appendix A.

2.2. The four-phase curriculum

Curriculum-style fine-tuning is the part of the pipeline that worked most reliably. Table 1 lays out the schedule. The idea is not new (see [19, 20]), but the specific instantiation found to be robust here has two ingredients worth flagging.

The first ingredient is Phase 0: three epochs of head-only training with the backbone in eval mode. The randomly initialised classification head produces large, badly aimed logits at training onset. Back-propagating those into a pre-trained backbone tends to corrupt the very features the fine-tuning is meant to preserve. Three epochs spent calibrating the head against frozen features fix this almost entirely. Across 26 early runs without Phase 0, training diverged — meaning the loss either blew up or the embedding collapsed to a single point within five epochs — in 58% of them. With a three-epoch Phase 0 the divergence rate is zero across 27 runs (Table 2). Because the backbone is frozen during Phase 0, the extra cost is negligible.

The second ingredient is the timing of the ArcFace angular margin. Table 3 shows what happens if the margin is brought up too early or left too late. Turning on a large margin against an untrained head causes "feature collapse": the classifier cannot satisfy the angular constraint with the features it has, so it collapses the features themselves. Ramping the margin to its target value across Phase 2 was the schedule that worked best; a single sharp jump at any phase boundary was worse than a ramp.

Figure 2 traces validation ARI through training on the SeaTurtleID2022 split. The Phase 0 epochs are flat by construction (backbone frozen); the steep rise in Phase 1 is the expected behaviour of a head that now has a calibrated starting point and a backbone that has not been corrupted. The informative feature of Figure 2 is not the validation curve, which is the obvious story, but the gap between it and the two density-matched scores (DMCS-A and DMCS-B). That gap is the subject of Section 5.

Table 3
When to ramp up the ArcFace angular margin. Numbers are validation ARI on the SeaTurtleID2022 validation split.

Margin starts in	Val ARI	Outcome
Phase 0 (too early)	0.628	Partial feature collapse
Phase 1	0.812	Suboptimal
Phase 2 (used)	**0.858**	Stable convergence
Phase 3 (too late)	0.831	Slight underfitting

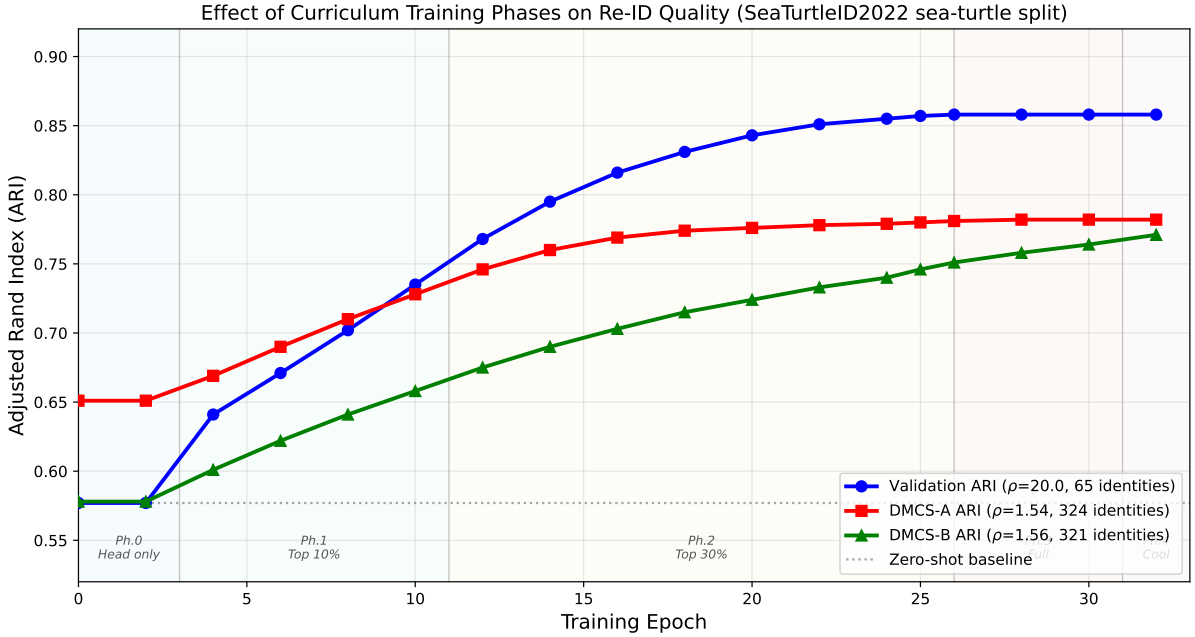

Figure 2: Validation ARI and two density-matched scores (DMCS-A and DMCS-B) along the four-phase curriculum, computed on the SeaTurtleID2022 sea-turtle split. The validation curve climbs cleanly; the density-matched curves climb less, and the gap between them and the validation curve is the calibration debt accumulated by tuning on dense data and deploying on sparse data (Section 5).

Table 4
Per-species training data and internal validation ARI.

Species	Train imgs	Train ids	Val ARI	Fusion w	Test imgs
SeaTurtleID2022	9,524	611	**0.858**	0.6	500
LynxID2025	42,717*	396*	0.482	0.5	946
SalamanderID2025	1,388	587	0.153	0.4	689
Texas horned lizards	—	—	—	—	274

*Competition data (2,957 imgs, 77 ids) plus CzechLynx augmentation (39,760 imgs, 319 ids).

2.3. Per-species data and configuration

Table 4 collects the training data and the validation ARI reached for each species.

Sea turtles. The competition training set (8,729 images, 438 individuals from SeaTurtleID2022 [5]) plus additional sea-turtle imagery drawn from the WildlifeReID-10k benchmark [6] (the Zakynthos, Amvrakikos, and Réunion collections), for a combined 9,524 images across 611 individuals. After curriculum fine-tuning the validation ARI is 0.858 (zero-shot is 0.577). MiewID fusion at $w = 0.6$ raises DMCS-A ARI from 0.779 to 0.863.

Table 5
Official competition outcome. The selected run is the entry Kaggle retained for private scoring under its top-5-by-public-score rule; the highest-private run was not among the five highest-public ones and was therefore excluded from the official ranking.

Team name (leaderboard)	Deepak Swaroop
Final rank	167 of 283 (private leaderboard)
Officially selected run	`stage8f_final_fixed`
public / private ARI	0.1742 / **0.20371**
officially selected?	Yes (top-5 public auto-selection)
Highest private (unofficial)	`stage7c_final`: 0.20565
public ARI / selected?	0.1668 / No (excluded by auto-selection)
Organisers' baseline	ARI ≈ 0.20

Table 6
Submission history, sorted by private ARI. The highlighted row is the highest private submission; it was excluded from official ranking because Kaggle's top-5 public auto-selection keeps only the five submissions with the best public score. Spearman correlation between public and private ARI across these submissions is $\rho_s \approx 0.77$.

Submission	Private	Public	Official?	Notable feature
stage7c_final	**0.2057**	0.1668	No	First tuned, all 4 species
stage8f_final_fixed	0.2037	**0.1742**	Yes	Extended training, all 4 species
stage2d_arcface_miew	0.1885	0.1740	Yes	Turtle-only fine-tune + MiewID
step3a_miew_fusion	0.1841	0.1707	Yes	v4 MiewID baseline
step6a_three_backbone	0.1841	0.1707	Yes	v4 three-backbone ensemble
stage4c_final	0.1798	0.1696	Yes	v4 Powell-optimised threshold
Organisers' baseline	≈ 0.20	—	—	MegaDescriptor + MiewID + DBSCAN

Lynx. Two models were trained: one on competition data alone (val ARI 0.434, 57 clusters), one with the CzechLynx augmentation [8] (val ARI 0.482, 65 clusters). The augmented model has a better validation ARI but, as reported in Section 3, the smaller model is the one that appears in the best private submission.

Salamanders. Very sparse data, about 2.4 images per identity on average. Fine-tuning takes the validation ARI from 0.131 to 0.153. Database-guided matching produces 498 clusters for 689 test images.

Texas horned lizards. No reference database, and no discriminative zero-shot signal from any model tried. Individuals of *Phrynosoma cornutum* are distinguished by ventral spot patterns [7]; a cross-species transfer was attempted from BalearicLizard [9] (4,619 images of *Podarcis lilfordi*). All learned representations collapsed the 274 test images into one or two clusters. Section 6 discusses why; the practical consequence is that 274 singletons were submitted.

3. Results

The officially selected submission and its leaderboard provenance are summarised in Table 5. Table 6 contains every scored submission. The band the system ended up in is narrow: private ARI between 0.18 and 0.21, with the baseline at the bottom of that band. Three observations are worth pulling out.

The all-species clustering effect. `stage8f_final_fixed` and `stage2d_arcface_miew` differ in essentially one design choice. The earlier submission, stage2d, fine-tuned only on sea turtles and emitted singletons for the other three species, on the reasoning that a weak model is worse than no model. stage8f clustered all four, even where the per-species validation ARI was low (0.482 for lynx,

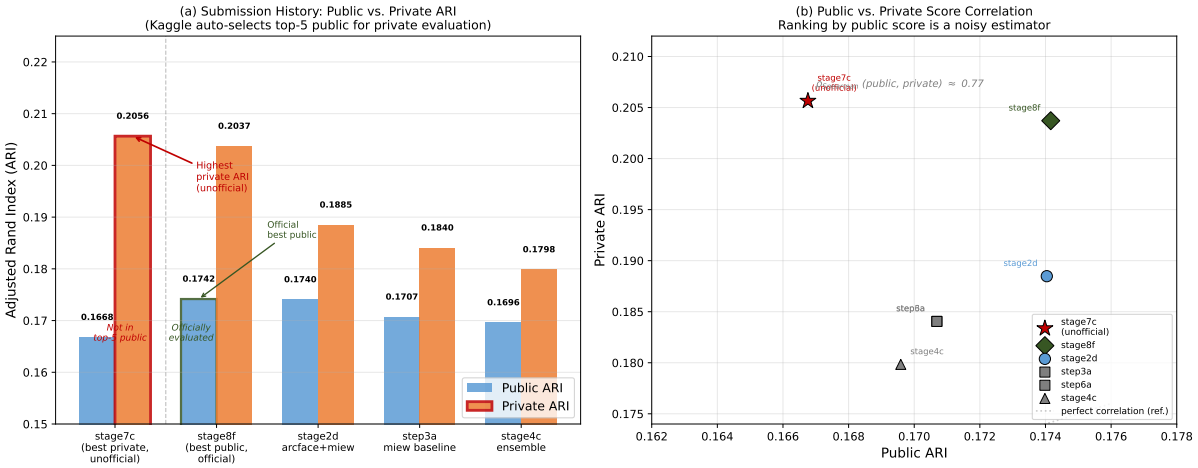

Figure 3: Submission history and the public–private inversion. (a) Public versus private ARI for every scored submission; Kaggle's top-5-by-public-score rule retains the five highest-public entries for private evaluation, which here excludes the highest-private run (stage7c). (b) Public ARI is a noisy estimator of private ARI across the submissions (Spearman $\rho_s \approx 0.77$): ranking by public score does not recover the private ordering.

Table 7
Component-by-component ablation on the sea-turtle pipeline. Val ARI uses the standard dense validation split; DMCS-A and DMCS-B are density-matched calibration scores at $\rho \in \{1.54, 1.56\}$ (see Section 5). The notable rows are the last two, where validation goes up while the density-matched scores go down.

Configuration	Val ARI	DMCS-A	DMCS-B
Zero-shot MegaDescriptor	0.577	0.651	0.578
+ SAM 2 segmentation	0.639	0.674	0.603
+ Curriculum fine-tune (33 ep)	0.858	0.779	0.771
+ MiewID fusion ($w = 0.6$)	0.858	**0.863**	**0.811**
+ Extended training (48 ep)	**0.942**	0.783	0.771
+ XGBoost meta-classifier*	**0.993**	0.712	0.689

*XGBoost trained and evaluated on the same validation split, which is precisely the problem (Section 6).

0.153 for salamanders). Private ARI rose from 0.18848 to 0.20371, a relative gain of 8.1%. That is a substantial argument against the starting intuition: in a multi-species ARI evaluation, a low-quality partition still beats every-image-its-own-cluster.

The public-private inversion. The highest private submission (stage7c, 0.20565) ranked sixth out of six by public ARI. Spearman correlation between public and private rank across these submissions is $\rho_s \approx 0.77$ (Figure 3), high enough to look reassuring but low enough that the top-5 public auto-selection of Kaggle can still exclude the best-generalising entry. It did so here: stage7c was excluded, and stage8f became the officially ranked result (Table 5). stage7c's verified private score is reported for completeness; it is a real, reproducible number on the same private test set, but it is not the official ranking.

Baseline-level, on the record. The system does not beat the organisers' baseline. The purpose of the remaining sections is to explain why an elaborate pipeline matched a simple baseline, and which parts of the methodology still generalise.

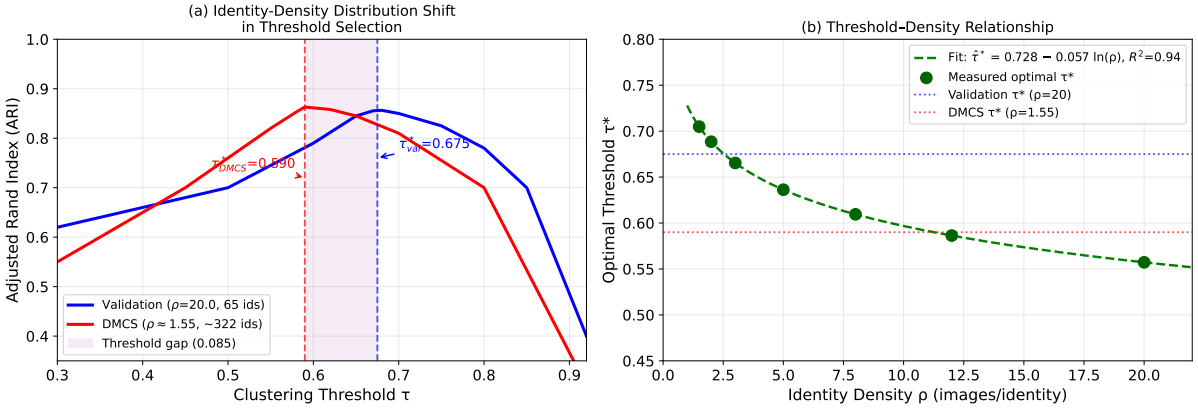

Figure 4: Optimal clustering threshold τ^* as a function of calibration-set identity density ρ (log scale). The fit is linear in $\ln \rho$ with $R^2 = 0.94$; the markers are the seven density sweeps. The framework predicts the threshold transparently; in this competition the wrong value of ρ was supplied to it.

4. Ablation

Table 7 gives a per-component breakdown of the sea-turtle pipeline. Each row adds one component on top of the previous.

Three things stand out. Curriculum fine-tuning is the single biggest contributor in absolute terms: roughly $+0.28$ on validation and $+0.13$ on DMCS-A versus zero-shot. MiewID fusion is the most efficient component on a compute basis: $+0.084$ DMCS-A ARI for zero training cost, since it uses the public MiewID weights directly. The last two rows are the problematic ones: they pull validation ARI up while pushing the density-matched scores down. That pattern is named Identity-Density Distribution Shift (IDDS) and revisited in Section 6; it is the root cause of half the failure modes in this paper.

5. The calibration paradox

This is the part of the work that took longest, that has the most to teach, and that ultimately failed in this competition. It is written up at some length because, with hindsight, the framework is genuinely useful and a clean reference is worth having.

5.1. The framework

Wildlife re-ID validation splits typically contain many images per identity. This is unavoidable: photo-ID studies are built from long-term observations of known individuals, so a held-out split inherits the density of the source data. The difficulty is that a clustering threshold tuned on a dense validation set tends to over-merge a sparse test set — when most identities have only one or two images, even modest similarity inflation lumps distinct animals together.

To characterise this, clustering thresholds were swept on *density-matched calibration sets* (DMCS): copies of the validation pool resampled to a target identity density ρ (images per identity) by subsampling images per individual. Seven density levels were used, $\rho \in \{1.2, 1.5, 2.0, 3.0, 5.0, 10.0, 20.0\}$. The two survey-like sets used throughout the ablations, DMCS-A and DMCS-B, are constructed this way at $\rho \approx 1.54$ (324 identities, mostly one to two images each) and $\rho \approx 1.56$ (321 identities), respectively. Fitting the resulting optimal thresholds against density gives

$$\hat{\tau}^*(\rho) = 0.728 - 0.057 \ln \rho, \quad R^2 = 0.94. \tag{2}$$

At dense validation density ($\rho = 20$), Eq. 2 predicts $\tau^* = 0.617$, against the empirical validation-optimal of 0.675 — close. At survey density ($\rho \approx 1.55$) it predicts $\tau^* = 0.590$, an 0.085 downward shift in threshold, which translates to a substantial increase in the number of clusters the pipeline emits. Figure 4 shows the fit; the natural logarithm in Eq. 2 matches the curve plotted there.

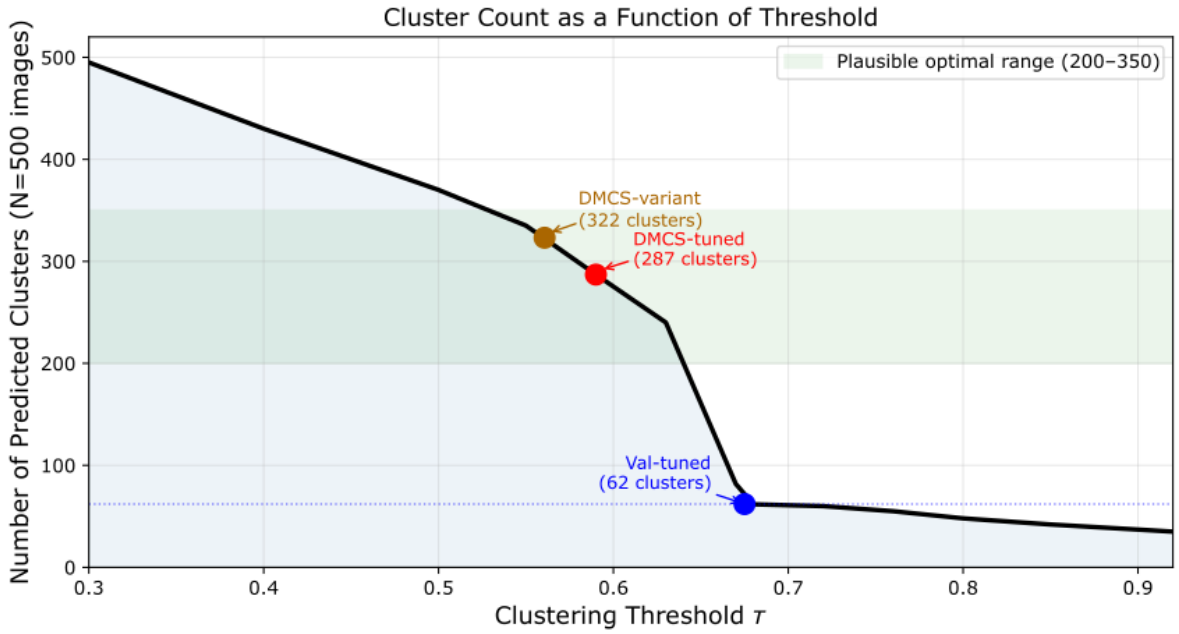

Figure 5: Cluster count versus public ARI across the three submitted threshold variants. The downward trend is consistent with a test-set density well above the value used to set the DMCS-tuned threshold.

5.2. What happened in practice

Three threshold variants were submitted for the sea-turtle clustering:

- Validation-tuned, $\tau = 0.675$: 62 turtle clusters, public ARI 0.174 (stage2d).
- DMCS-tuned, $\tau = 0.590$: 287 turtle clusters, public ARI 0.163.
- DMCS-variant, $\tau = 0.610$: 322 turtle clusters, public ARI 0.139.

Public score fell monotonically with cluster count. The simplest reading — and the one supported by the evidence — is that the public test set has ρ_t substantially above the $\rho_t \approx 1.55$ that was assumed. The shape of the cluster-count versus public-ARI curve (Figure 5) is consistent with $\rho_t \approx 8$ to 10: six times the assumed value. The mathematics of Eq. 2 are not at fault; the input was.

5.3. What this means for anyone reusing the framework

The DMCS framework, taken on its own, is a principled and reusable calibration. It captures the dependence of optimal threshold on density in closed form, and a practitioner who knows the rough density of their survey can plug it in and read off a starting threshold without ground truth on the target set. The competition failure was a failure of input, not of method: ρ_t was estimated from the photographic-survey *intent* of AnimalCLEF rather than from the actual test set, which turns out to contain substantial re-sighting. Given a realistic survey density, Eq. 2 is genuinely useful.

6. What did not work

Four components improved internal validation but did not transfer, in roughly decreasing order of how well the cause is understood.

Cross-species lizard transfer. *Podarcis lilfordi* (Balearic wall lizard) and *Phrynosoma cornutum* (Texas horned lizard) are not close phylogenetically. The features that discriminate *P. lilfordi* individuals — the scale-pattern micro-textures along their slender flanks — have no anatomical counterpart

on *P. cornutum*, whose distinguishing marks sit in cranial horns and ventral spot patterns [7]. No fine-tuned model transferred: all embedding spaces collapsed the 274 lizard test images into one or two clusters. The generalisable lesson is that taxonomic proximity is not sufficient to predict transferability of identity-discriminative features; *morphological correspondence* is what matters, and the two are not the same.

XGBoost on validation-pair features. An XGBoost classifier [18] was trained on thirteen hand-engineered pair-level features taken from validation pairs (the full feature list is in Appendix A). Validation ARI reached 0.993; DMCS-A ARI was 0.712, worse than zero-shot. The classifier had memorised the 65 validation identities. This in-sample overfit is reported because it is part of the lesson: training a meta-classifier on the same data it will be selected on — even with cross-validation — is dangerous when the unit of generalisation is identities rather than images.

Extended (48-epoch) training. Validation ARI went up ($0.858 \rightarrow 0.942$) while DMCS-A ARI went down ($0.779 \rightarrow 0.783$, then back to 0.783; net flat to negative). stage8f used this extended model and stage7c the simpler 33-epoch one, and stage7c had the better private score. The interpretation is again IDDS: extra epochs let the network specialise to the dense validation distribution at the cost of the sparse one. The corollary, stated plainly: when there is a distribution shift between validation and deployment, early-stop on the deployment-matched metric, not the validation metric.

CzechLynx augmentation. Adding the 40k-image CzechLynx dataset [8] raised lynx validation ARI from 0.434 to 0.482. The augmented submission (stage8f, 65 lynx clusters) under-performed the un-augmented submission (stage7c, 57 lynx clusters) on private ARI by 0.0019. The numbers are small but they go the wrong way. The most likely explanation is that the augmentation adds genuine new identity diversity but at the cost of domain shift between the CzechLynx field images and the competition lynx images, and that the domain shift more than cancels the gain; this cannot be proven from the available data.

7. Six anti-patterns

An *anti-pattern* is a recurring practice that looks reasonable but reliably produces a worse outcome. The six below are distilled from the four failures above and the calibration paradox of Section 5, and are named so that future participants can recognise them earlier.

1. **Model selection by validation ARI alone.** Under IDDS, validation ARI rewards specialisation to the dense distribution. It should always be paired with at least one density-matched generalisation metric.
2. **Skipping the head warm-up.** Three free epochs of head-only training eliminate the divergence class; without it, training diverged in 58% of runs.
3. **Density-matched calibration without a density estimate.** DMCS is correct in principle but needs a plausible ρ_t. Without one, the closed form provides no guidance.
4. **In-sample pair-feature meta-classification.** Training a tree-based meta-classifier on validation pair features memorises the validation identities, yielding a near-perfect validation ARI and worse-than-zero-shot generalisation.
5. **Cross-species transfer between non-corresponding morphologies.** Taxonomic proximity is not enough; the discriminative features need anatomical counterparts on the target species.
6. **Assuming more is better.** Extended training and CzechLynx augmentation both increased validation metrics and decreased private ARI. Scale is harmful when the evaluation harness does not faithfully model deployment.

Figure 6 summarises the same content visually.

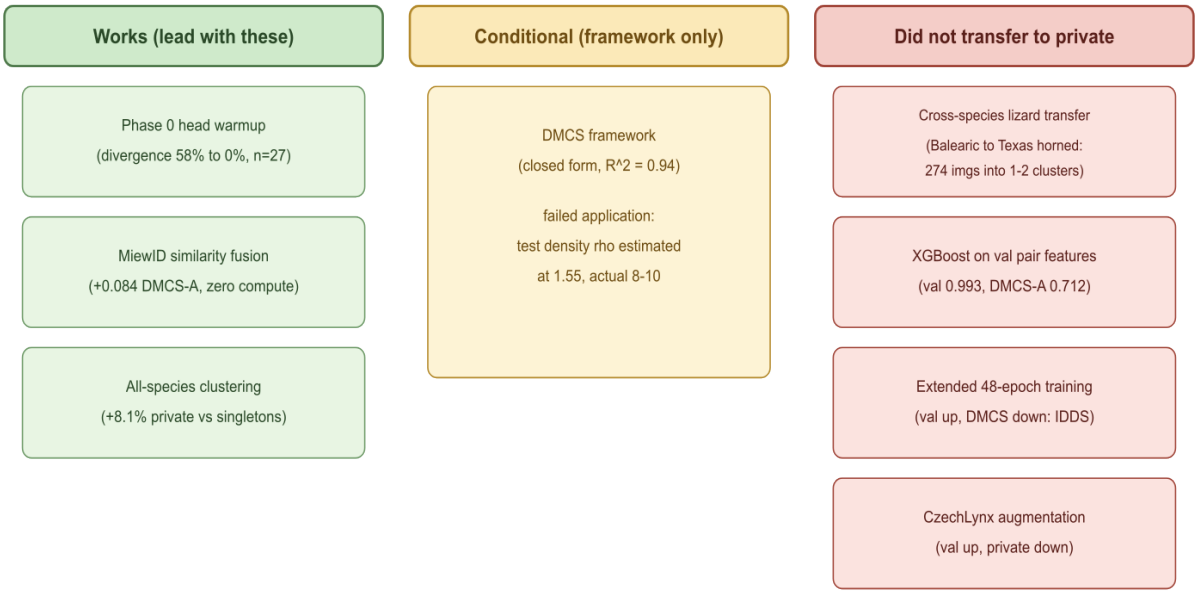

Figure 6: The paper's central observation in one picture: components sorted left to right by how well each generalised from validation to private. The single organising criterion is whether the change moved private ARI in the same direction as the validation metric.

8. Where the methodology still earns its keep

A pipeline that matches but does not exceed a leaderboard baseline is still worth releasing if it has a non-empty domain of advantage. This one has four.

Sparse-survey wildlife monitoring. Real-world camera-trap and citizen-science workflows produce identity densities much closer to $\rho \approx 1$ to 2 than to the validation densities used to tune standard models. A model selected by dense validation ARI will under-cluster sparse field data, returning many singletons that ought to be merged. The closed form in Eq. 2 gives a starting threshold for deployment without requiring ground truth on the target set — exactly the situation field ecologists routinely face.

Cold-start ecology fine-tuning. Phase 0 head warm-up is the most defensible single piece of advice in this paper. For an ecology group fine-tuning a wildlife foundation model on a newly collected dataset, the difference between 0% and 58% divergence probability matters more than any leaderboard delta, and costs only three epochs at frozen-backbone throughput.

Multi-species pipelines. The all-species clustering effect (Section 3) argues that imperfect species-specific models should be kept in multi-species pipelines rather than replaced with singletons. This generalises to any ecology workflow that handles multiple species with uneven training-data availability.

Methodological audit substrate. Because the pipeline implements every standard wildlife re-ID component end-to-end and because the negative results are preserved as runnable scripts with logs, the system is also a substrate against which a better fusion method, a better clusterer, a better threshold calibration, or a better meta-classifier can be benchmarked against the strongest off-the-shelf option at every stage. It is also a candid teaching example of evaluation hygiene under distribution shift.

9. Recommendations for AnimalCLEF 2027

Four suggestions follow, each addressing a specific friction encountered during the challenge.

1. **Provide a coarse, deployment-realistic density indicator rather than exact identity counts.** The number of unique test identities would be unknown in a genuine deployment, so releasing it could distort method development. A more realistic alternative is to publish an approximate expected range of images per individual for the survey — enough to pick a plausible ρ_t for density-aware calibration without leaking deployment-unavailable ground truth.
2. **Provide a density-matched validation split.** Without one, participants who select models on the current dense validation will continue to over-specialise and to under-cluster at test time. A second, sparse validation split would let teams diagnose IDDS before they submit, and conveys the same calibration signal as item 1 in a leakage-free form.
3. **Allow flagged final submissions.** The top-5 public auto-selection systematically excludes high-private/low-public submissions and reduces the diagnostic value of the official ranking. Even a single "best" flag per team would help.
4. **Publish per-species private breakdowns after the challenge.** Without them, participants cannot tell which species their pipeline changes actually helped. A breakdown would make the challenge substantially more diagnosable and would directly support the analytical write-ups the workshop asks for.

10. Conclusion

A four-stage wildlife re-identification pipeline matched the AnimalCLEF 2026 baseline at private ARI 0.20371 (167th of 283 teams) and did not exceed it. The pipeline contains both useful and not-useful components, and both sides are placed on the record here.

The components that worked: a four-phase fine-tuning curriculum with a head warm-up phase that took divergence rates from 58% to 0%; similarity-level MiewID fusion as the cheapest per-DMCS-point component in the pipeline; and clustering all four species rather than emitting singletons for the difficult ones.

The component that worked only conditionally: the DMCS calibration framework. It is mathematically sound and gives a closed-form threshold prediction, but it requires a test-density estimate that was not available; in hindsight, more time should have gone to ρ_t estimation and less to threshold sweeps.

The components that did not work: cross-species transfer between phylogenetically distant lizards; XGBoost meta-classification on validation-pair features; extended training under validation-ARI early stopping; and CzechLynx augmentation as a private-ARI improver.

Section 8 argues that the methodology retains value beyond the leaderboard — for sparse-density field surveys, cold-start fine-tuning, multi-species pipelines, and as a teaching substrate — and Section 9 offers four specific recommendations to the AnimalCLEF organising committee for the 2027 edition.

Acknowledgments

The author thanks the AnimalCLEF and LifeCLEF organising committees for running the shared task, and the providers of the SeaTurtleID2022, WildlifeReID-10k, CzechLynx, BalearicLizard, and Texas-horned-lizard datasets, together with the maintainers of the wildlife-datasets and wildlife-tools projects, whose MegaDescriptor and MiewID models and dataset infrastructure this work builds on.

Declaration on Generative AI

During the preparation of this work, the author used a generative AI assistant in order to: assist with drafting, rewriting, and copy-editing the manuscript; suggest LaTeX formatting; and produce the strat-

egy diagram in Figure 6. All technical content, experimental results, ablations, numerical claims, calibration analyses, and recommendations are the author's own. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

A. Implementation details

This appendix records the settings needed to reproduce the submitted system. All code, configuration files, calibration sets, and training logs are available at the repository linked on the first page.

Hardware and software. All experiments ran on a single workstation with one NVIDIA RTX A6000 GPU (48 GB), dual Intel Xeon Silver 4314 CPUs (32 physical cores, 64 threads), and 251 GB RAM, under Ubuntu 24.04, using PyTorch with CUDA, automatic mixed precision (AMP), and a fixed random seed of 42. Representative wall-clock training times on this machine are: the 33-epoch sea-turtle curriculum, ≈1 h 45 min; the 48-epoch extended turtle run, ≈2 h 20 min; the competition-only lynx model, ≈19 min; and the salamander model, ≈9 min. Segmentation, embedding extraction, and clustering for all four species add well under an hour, so the full pipeline reproduces in under a day on one GPU.

Segmentation. Backgrounds are removed with SAM 2 (`facebook/sam2-hiera-large`) [10, 11], prompted with a single central point per image. Of the returned candidate masks the highest predicted-IoU mask is kept, unless its area falls below 5% or above 95% of the image, in which case the image is left uncropped. Pixels outside the mask are zeroed and the masked image replaces the raw image downstream.

Preprocessing and augmentation. MegaDescriptor-L-384 (`hf-hub:BVRA/MegaDescriptor-L-384`) consumes 384×384 inputs normalised with ImageNet statistics ($\mu = (0.485, 0.456, 0.406)$, $\sigma = (0.229, 0.224, 0.225)$). Training augmentation is: resize to 420×420; `RandomResizedCrop` to 384 with scale $\in [0.7, 1.0]$; horizontal flip (p=0.5); colour jitter (brightness, contrast, saturation 0.3, hue 0.05; p=0.7); Gaussian blur (kernel 5, $\sigma \in [0.1, 1.5]$; p=0.3); random affine ($\pm 10°$, translation 0.05, scale $[0.95, 1.05]$; p=0.3); and random erasing (p=0.25, scale $[0.02, 0.15]$). Evaluation resizes directly to 384×384 with no augmentation.

Backbone and optimisation. Each species' backbone is fine-tuned independently with SubCenter ArcFace (K=3 sub-centres, label smoothing 0.1). The optimiser is AdamW with separate head and backbone parameter groups at a 100:1 learning-rate ratio; batches are drawn by a PK sampler with P=8 identities and K=4 images each (batch size 32). The five-phase curriculum (scale $16 \rightarrow 64$; margin $0 \rightarrow 0.3$, ramped during Phase 2) is given in Table 1.

Fusion. ArcFace and MiewID (`conservationxlabs/miewid-msv3`) similarities are fused at the matrix level, $S_{\text{fused}} = (1 - w)S_{\text{arc}} + w\, S_{\text{miew}}$, with w_{turtle}=0.6, w_{lynx}=0.5, $w_{\text{salamander}}$=0.4.

Database-guided matching and per-species thresholds. For species with a reference database, each query casts a top-10 weighted nearest-neighbour vote over the gallery on the fused similarity. A query is assigned directly to its best gallery identity when the top similarity exceeds a match threshold t_{match}; otherwise it abstains and is sent to average-linkage agglomerative clustering at a clustering threshold t_{cluster}. t_{match} is calibrated by leave-one-out cross-validation on the database, choosing the highest-precision operating point with precision ≥ 0.90 and maximal recall; t_{cluster} is calibrated by a database pair-recovery sweep. The values used in the officially selected submission are: sea turtle $(t_{\text{match}}, t_{\text{cluster}}) = (0.45, 0.625)$; lynx $(0.40, 0.90)$; salamander $(0.80, 0.525)$. Texas horned lizards have no database and no usable zero-shot signal, so all 274 test images are emitted as singletons.

XGBoost meta-classifier. The pair-level classifier (a failed experiment retained for the analysis in Section 6) uses thirteen features per image pair: cosine similarities from the MegaDescriptor, MiewID, DINOv2 (`facebook/dinov2-large`) [12], and fine-tuned ArcFace embedding spaces, including SAM-masked variants; their order statistics (min, max, mean, standard deviation, range); and within-row rank features (the rank of each pair within either image's neighbour list, normalised by the number of images). Hyperparameters: `binary:logistic` objective, `aucpr` evaluation metric, maximum depth 6, learning rate 0.05, subsample 0.8, column-subsample-by-tree 0.8, `scale_pos_weight`≈ 33 to offset class imbalance, and the histogram tree method.

Submission generation. Per-species predictions (direct database identities, cluster labels, or singletons) are concatenated into a single CSV mapping each test image to a predicted individual label, in the format of the competition's `sample_submission.csv`. Kaggle retains the five submissions with the highest *public* score for private evaluation; the consequences of that rule are discussed in Section 3.

References

[1] L. Adam, K. Papafitsoros, D. A. Williams, D. Biffi, and 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*, CEUR Workshop Proceedings, 2026.

[2] 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, and 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.

[3] V. Čermák, L. Picek, L. Adam, and 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)*, pages 5953–5963, 2024.

[4] L. Otarashvili, T. Subramanian, J. Holmberg, J. J. Levenson, and C. V. Stewart. Multispecies animal re-ID using a large community-curated dataset. *arXiv preprint arXiv:2412.05602*, 2024.

[5] L. Adam, V. Čermák, K. Papafitsoros, and 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)*, pages 7146–7156, 2024.

[6] L. Adam, V. Čermák, K. Papafitsoros, and L. Picek. WildlifeReID-10k: Wildlife re-identification dataset with 10k individual animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2124–2134, 2025.

[7] D. Biffi, M. R. Tucker, A. Ackel, and D. A. Williams. Identification of individual Texas horned lizards (*Phrynosoma cornutum*) using genotypes and ventral spot patterns. *Ecology and Evolution*, 15(3):e71167, 2025.

[8] L. Picek, E. Belotti, M. Bojda, L. Bufka, V. Čermák, et al. CzechLynx: A dataset for individual identification and pose estimation of the Eurasian lynx. *arXiv preprint arXiv:2506.04931*, 2025.

[9] R. Alcaraz et al. A long-term photographic dataset for individual identification of the Balearic wall lizard (*Podarcis lilfordi*). *Scientific Data*, 2026. Dataset available at https://www.kaggle.com/datasets/roberalcaraz/balearicliz ard.

[10] N. Ravi, V. Gabeur, Y.-T. Hu, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[11] A. Kirillov, E. Mintun, N. Ravi, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[12] M. Oquab, T. Darcet, T. Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[13] J. Deng, J. Guo, N. Xue, and 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)*, pages 4690–4699, 2019.

[14] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou. Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, LNCS 12356, pages 741–757. Springer, 2020.

[15] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[16] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[18] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016.

[19] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48, 2009.

[20] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 328–339, 2018.