METAFAUNA: A MULTI-SPECIES ANIMAL RE-IDENTIFICATION DATASET WITH STRUCTURED METADATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Animal re-identification (ReID) aims to recognize the same individual across images, supporting non-invasive wildlife monitoring and ecological studies. In camera-trap imagery, however, identity-relevant visual evidence varies with observation conditions: viewpoint and occlusion affect which cues are visible, while illumination, blur, and scale affect how clearly those cues are captured. Most existing animal ReID datasets provide only images and identity labels, requiring models to distinguish identity-related differences from variation associated with these observation conditions using visual input alone. Recent work has begun incorporating auxiliary metadata, such as temperature, capture time, and orientation, into animal ReID. However, most of these attributes describe the temporal or environmental context of capture rather than the observation conditions of each image. To address this gap, we construct MetaFauna, a multi-species animal ReID dataset comprising 44,317 images of 2,500 individuals across eight species. Each image is paired with structured metadata describing its observation conditions, grouped into visual evidence visibility and capture conditions, and, where available, its environmental context. To study whether explicit descriptions of observation conditions improve animal ReID, we further introduce MetaRouter, a reference framework that conditions visual feature extraction on the metadata of each image by weighting low-rank adaptation (LoRA) experts within a frozen visual backbone. Across all eight species, MetaRouter outperforms visual-only and metadata-aware baselines. Controlled comparisons further show that the gains depend on using the metadata paired with each image to guide routing.

1 INTRODUCTION

Animal re-identification (ReID) aims to recognize the same individual across images collected at different times and locations, supporting individual-level wildlife monitoring and ecological studies of population dynamics, movement, and behavior (Schneider et al., 2019; Papafitsoros et al., 2021; Schofield et al., 2022). Camera traps provide a scalable and non-invasive way of observing animals in natural habitats, making them an important source of imagery for automated ReID (Beery, 2023; Xu et al., 2024). Reliable ReID in these images relies on fine-grained identity-relevant visual cues, such as coat patterns and markings (Wu et al., 2026). Images of the same individual are often captured across separate encounters, so the conditions under which the individual is observed vary across images. These observation conditions affect the available identity evidence in two ways: the visibility of identity cues varies with conditions such as viewpoint and occlusion, whereas the clarity of visible cues varies with conditions such as illumination and blur (Perneel et al., 2025).

However, this variation is not explicitly represented in standard animal ReID supervision. Most datasets provide images and identity labels, but little or no information about the observation conditions associated with each image (Čermák et al., 2024; Adam et al., 2025; Naik et al., 2024). Models trained on these data must therefore distinguish identity-related differences from visual variation associated with observation conditions using the image alone (Schall et al., 2026; Hou et al., 2025). Metadata associated with each image could make some of these conditions explicit.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

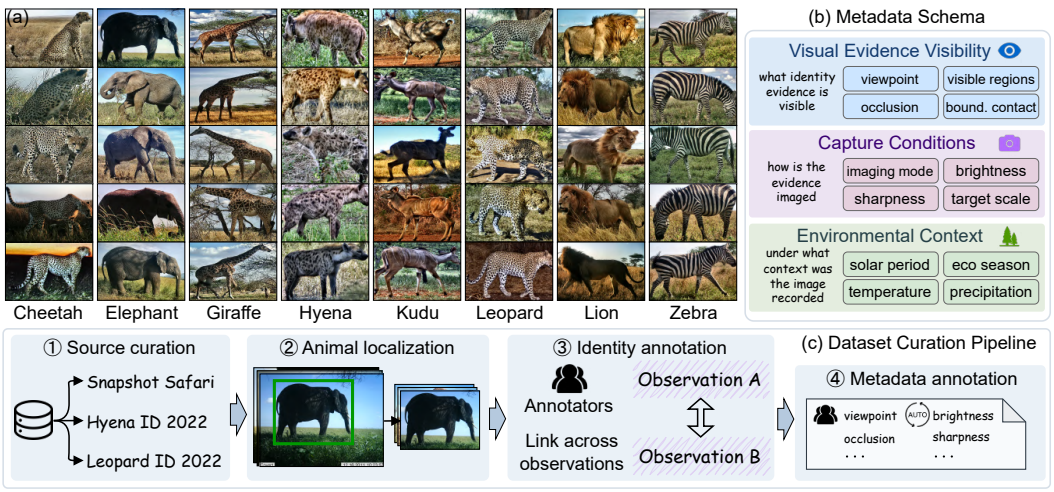


Figure 1: **Overview of MetaFauna.** (a) MetaFauna covers eight animal species. (b) Each image is paired with structured metadata describing visual evidence visibility, capture conditions, and environmental context. (c) The dataset is constructed through source curation, animal localization, cross-observation identity annotation, and metadata annotation.

Recent work has begun to incorporate metadata into animal ReID (Sørdalen et al., 2026). MetaWild pairs images from six species with temperature, circadian rhythm, and face orientation, and introduces the Meta-Feature Adapter to integrate these attributes with visual features (Li et al., 2025). More recently, Tur et al. (2026) condition parameter-efficient vision–language adaptation on capture time and body size during training. These studies show that metadata can improve animal ReID, but most of the attributes they consider describe the temporal or environmental context of capture, or properties of the animal, rather than the observation conditions of each image. Existing metadata, however, does not directly record which identity-relevant cues are visible in each image or how clearly those cues are captured. This leaves open the question of whether explicit descriptions of these observation conditions improve animal ReID.

To address this gap, we construct **MetaFauna** (Figure 1), an animal ReID dataset containing 44,317 images of 2,500 individuals across eight species. Each image is paired with structured metadata describing its observation conditions, organized into visual evidence visibility and capture conditions, together with environmental context where available. First, *visual evidence visibility* describes which identity-relevant body regions and visual cues are observable, including viewpoint, visible body regions, occlusion, and boundary contact. Second, *capture conditions* characterize how the animal and its identity cues are imaged, including imaging mode, brightness, sharpness, and target scale. Third, *environmental context* records the temporal and ecological conditions associated with image capture, including the solar period, ecological season, air temperature, and precipitation. To study whether explicit descriptions of observation conditions improve animal ReID, we propose **MetaRouter**, a reference framework for metadata-conditioned visual adaptation. MetaRouter encodes metadata for each image and uses it to weight low-rank adaptation (LoRA) (Hu et al., 2022) experts within a frozen visual backbone, thereby providing a parameter-efficient mechanism for metadata-conditioned feature extraction.

In summary, this paper makes three main contributions. First, we construct MetaFauna, an animal ReID dataset that pairs each image with an individual identity and structured metadata describing its observation conditions, enabling the study of whether explicit descriptions of these conditions improve animal ReID. Second, we propose MetaRouter, a reference framework for conditioning visual feature extraction on the metadata associated with each image. MetaRouter uses metadata to weight LoRA experts within a frozen visual backbone, allowing feature adaptation to vary with the metadata of each image. Third, experiments across all eight species show that MetaRouter consistently outperforms visual-only and metadata-aware baselines. Controlled ablations further show that the benefit comes from conditioning on the metadata paired with each image.

2 Related Work

Animal ReID Models aim to learn visual representations that distinguish individuals despite changes in pose, viewpoint, occlusion, and background. Species-specific approaches have exploited distinctive markings and pose-aligned body parts. The ATRW study, for example, introduced pose-part modeling to align deformable tiger markings across images (Li et al., 2020). More recent work has focused on representations that transfer across species and datasets. UniReID combines category-specific visual prompts with semantic knowledge to generalize to unseen wildlife categories (Jiao et al., 2023), while OpenAnimals evaluates person ReID designs on multiple animal benchmarks and introduces ARBase as a general animal ReID baseline (Hou et al., 2025). Other approaches incorporate additional visual or semantic guidance: AdaFreq emphasizes frequency-aware visual features, PAW-ViT learns anatomical part representations through semantic segmentation distillation, and CARE uses image-conditioned textual descriptions and semantic feature alignment to guide identity representation learning (Li et al., 2024; Neto et al., 2026; Wu et al., 2026). These methods address observation variation primarily through information extracted from the image itself. Our work instead studies whether metadata that makes observation conditions explicit can help models better interpret identity-relevant visual evidence across varying conditions.

Animal ReID Datasets range from species-specific collections to multi-species resources assembled for standardized training and evaluation. WildlifeDatasets provides common tools for accessing and evaluating public animal ReID datasets, while WildlifeReID-10k resamples existing datasets into a benchmark containing more than 140k images of 10k individuals from 33 species (Čermák et al., 2024; Adam et al., 2025). Recent work has begun pairing identity annotations with metadata. MetaWild contains 20,890 images from six species with temperature, circadian rhythm, and face orientation (Li et al., 2025). Tur et al.'s study applies continuous metadata conditioning to the longitudinal Melops dataset, which provides capture-level metadata, including body length, capture location, and date (Tur et al., 2026). MetaFauna complements these resources by providing structured image-level annotations that more directly describe the visibility and capture of identity-relevant visual evidence, as well as temporal and environmental context.

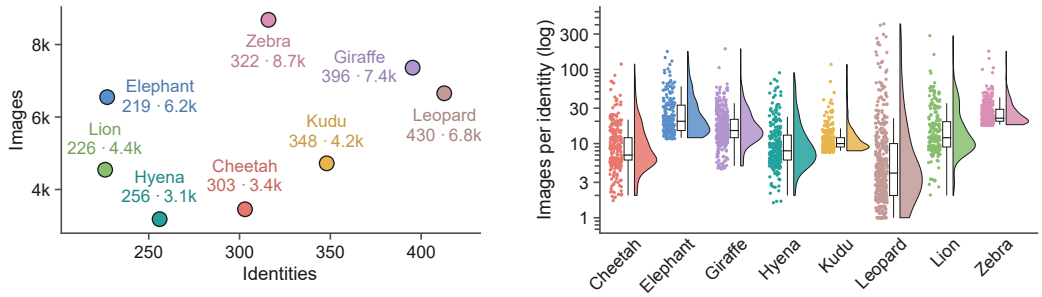
Parameter-Efficient Adaptation and Expert Routing. MetaRouter builds on two related lines of work: parameter-efficient adaptation and expert routing. LoRA keeps pretrained weights frozen and represents task-specific updates with low-rank matrices (Hu et al., 2022). Expert-routing architectures use a learned router to assign tokens or inputs to specialized modules; V-MoE, for example, routes visual tokens to a subset of feed-forward experts within a Vision Transformer (Riquelme et al., 2021). Recent work combines these ideas by treating low-rank updates as experts. MoLE learns layer-wise gates to compose separately trained LoRA modules, while Omni-SMoLA uses soft routing over low-rank experts in multimodal models (Wu et al., 2024b;a). Expert mixtures have also been applied to ReID: META adaptively aggregates domain-specific experts using normalization statistics, while MoSCE-ReID clusters semantic attributes and uses group-specific LoRA experts (Xu et al., 2022; Ren et al., 2025). MetaRouter adopts this general low-rank expert formulation but uses structured metadata paired with each image as the routing input, linking expert weighting directly to the observation conditions recorded for that image.

3 The MetaFauna Dataset

MetaFauna pairs animal images and identity annotations with structured metadata describing which parts of the animal are available for identification and how they are captured, together with temporal and environmental context where available. The metadata is designed to make image-level observation conditions explicit while retaining the natural variation present in camera-trap imagery. We first describe the dataset sources, composition, and organization, and then detail the construction of animal instances, identities, and structured metadata.

3.1 Dataset Composition

MetaFauna is constructed from three publicly available wildlife datasets. Most images are sourced from Snapshot Safari (Swanson et al., 2015; Pardo et al., 2021), a large-scale camera-trap collection available through the Labeled Information Library of Alexandria: Biology and Conserva-



(a) Images and identities per species.

(b) Distribution of images per identity.

Figure 2: **MetaFauna dataset statistics.** MetaFauna contains 44,317 images of 2,500 identities across eight species. (a) Image and identity count for each species. (b) Distribution of images per identity for each species, shown on a logarithmic scale.

tion (LILA).[1] For the selected Snapshot Safari images, we retain the source media identifiers and available sequence-level information while constructing individual identity annotations as part of MetaFauna. MetaFauna additionally incorporates Hyena ID 2022 and Leopard ID 2022, for which we retain the manually confirmed individual identities provided by the source releases (Botswana Predator Conservation Trust, 2022a;b). The resulting dataset contains 44,317 images of 2,500 individuals across eight species. Figure 2 summarizes the number of images and identities for each species, together with the distribution of images per identity.

To organize the three source datasets in a consistent format, MetaFauna adopts a data structure adapted from Camtrap DP (Bubnicki et al., 2024). Each released ReID sample is linked to its source image and, where available, to the sequence or capture event to which that image belongs. Individual identity and structured metadata are then associated with the corresponding sample, while the original source identifiers are retained to maintain traceability to the source data. The original images remain under the licenses and citation requirements of Snapshot Safari, Hyena ID 2022, and Leopard ID 2022, respectively. Annotations created for MetaFauna, including the individual identities constructed for Snapshot Safari and the structured metadata added across all three sources, are released under the CDLA-Permissive-1.0 license.[2]

3.2 Dataset Construction

MetaFauna is constructed in three stages: we first localize target animal instances, then establish individual identities across observations, and finally annotate or derive metadata for each image. Because variation in observation conditions is central to the dataset, challenging images are not filtered solely on the basis of occlusion, blur, unusual viewpoints, boundary truncation, or small target scale; they are retained whenever the target instance and identity can be reliably established. Detailed construction and quality-control procedures are provided in Appendix A.

Image Selection and Instance Curation. For each source, we first select images or sequences containing the target species and obtain instance-level animal crops for subsequent identity annotation. Where reliable instance annotations are available, we use the source-provided bounding boxes or crops; otherwise, animal instances are localized using a fixed pretrained MegaDetector model (Beery et al., 2019). Each detected animal is treated as a separate candidate instance, so a source image may contribute multiple ReID samples when individual animals can be localized and unambiguously assigned to identities. Instances with ambiguous localization or identity assignment are excluded. Images containing occlusion, truncation, blur, or small targets are retained when the target animal can still be reliably localized. Bounding boxes are expanded by a fixed margin and clipped to the image boundary before cropping. Detector settings, crop construction, and instance-level quality control are described in Appendix A.

[1] https://lila.science/datasets

[2] https://cdla.dev/permissive-1-0/

Identity Annotation. Snapshot Safari organizes camera-trap images into short sequences captured by the same camera over a brief time interval. This temporal continuity provides a natural basis for linking repeated detections of the same animal across adjacent frames, while the grouped frames can also provide complementary visual evidence for subsequent identity matching. We therefore first group crops corresponding to the same animal within each sequence. For each group, we use the temporal midpoint image as its representative frame and extract a visual embedding using DINOv3 (Siméoni et al., 2025). Candidate matches are ranked within the same species by embedding similarity to prioritize likely matches for manual review. Each candidate pair is independently reviewed by two annotators, who assign *same individual*, *different individuals*, or *unresolved* based on the visual evidence across the corresponding sequences. If both annotators select *different*, the candidate pair is rejected. If both select *same*, a third annotator independently reviews the pair, and an identity association is accepted only when all three agree. Any disagreements or unresolved decisions are also reviewed by the third annotator. Confirmed same-individual associations are then combined to form individual identities. Hyena ID 2022 and Leopard ID 2022 already provide manually confirmed identities, so their source identity labels are retained after verification.

Metadata Construction. MetaFauna organizes its metadata into three groups: visual evidence visibility (G1), capture conditions (G2), and environmental context (G3). G1 and G2 provide explicit image-level descriptions of the observation conditions central to our study, whereas G3 records additional temporal and environmental context associated with image capture. Fields requiring semantic interpretation of the animal are manually annotated, while fields with direct operational definitions are computed using fixed procedures. Environmental attributes are derived from capture records or matched external data and are left missing when reliable spatial or temporal matching is unavailable.

G1 records the viewpoint, visible body regions, external occlusion, and boundary contact, which are annotated from the original frame and the verified target box. G2 describes imaging mode, brightness, sharpness, and target scale; imaging mode is obtained from source or camera information, while the remaining fields are computed from the target crop or bounding box. G3 records solar period, ecological season, air temperature, and precipitation using capture timestamps, location information, and external environmental records. Detailed field definitions, label spaces, annotation rules, and computation procedures are provided in Appendix A.2.

4 THE METAROUTER FRAMEWORK

MetaRouter is a reference framework for studying whether structured metadata that makes observation conditions explicit can improve animal ReID. It uses the metadata associated with each image to condition parameter-efficient adaptation within a pretrained visual backbone. We first describe the structured metadata representation, followed by metadata-conditioned LoRA adaptation and the training and inference procedures. Figure 3 provides an overview of MetaRouter.

Structured Metadata Representation. MetaFauna contains heterogeneous metadata fields with different representations and availability, which MetaRouter encodes individually before aggregating them into a fixed-dimensional metadata representation. For an image x_i, we denote its metadata by $m_i = \{(f, v_i^f) \mid f \in \mathcal{F}_i\}$, where $\mathcal{F}_i$ is the set of available metadata fields and v_i^f is the value of field f. Since the meaning and representation of a value depend on the field to which it belongs, we encode each field-value pair as:

$$t_i^f = E_{\text{field}}(f) + \phi_f(v_i^f), \tag{1}$$

where $E_{\text{field}}(f)$ identifies the metadata field and $\phi_f(\cdot)$ maps its value into the common token space. Categorical values use learned embeddings, continuous values are normalized and linearly projected, and vector-valued fields are projected directly.

To combine the available fields into a single metadata representation, we prepend a learnable `[META]` token and process the resulting tokens with a two-layer Transformer encoder (Vaswani et al., 2017). Let t_{meta} denote the `[META]` token and Enc_m the metadata encoder. Since the metadata fields have no fixed sequential order and their identities are explicitly represented by $E_{\text{field}}(f)$, the available field tokens are processed without positional embeddings and can be provided in any order. For an ordering $(f_1, \ldots, f_{|\mathcal{F}_i|})$ of the available fields, we construct the input token sequence T_i^m and obtain

$$T_i^m = \left[t_{\text{meta}}, t_i^{f_1}, \ldots, t_i^{f_{|\mathcal{F}_i|}}\right], \qquad \widehat{T}_i^m = \text{Enc}_m(T_i^m), \qquad h_i^m = \widehat{T}_{i,\text{meta}}^m, \tag{2}$$

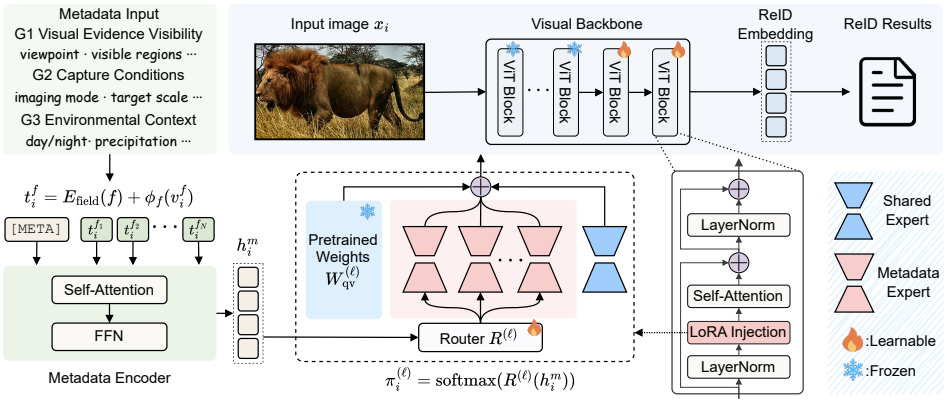


Figure 3: **Overview of MetaRouter.** For each image, the available metadata is encoded into a metadata representation that is used to predict the mixture weights of LoRA experts within the frozen visual backbone. The resulting metadata-conditioned low-rank updates allow visual feature extraction to vary with the metadata associated with each observation.

where $\widehat{T}_i^m$ denotes the encoded token sequence and h_i^m is the output corresponding to the [META] token. Because only available fields are encoded, the same mechanism accommodates different metadata subsets across images, with h_i^m serving as the input to the routing mechanism.

Metadata-Conditioned LoRA Adaptation. Standard LoRA applies the same low-rank update to every image. To condition visual adaptation on metadata while retaining a metadata-independent component, MetaRouter combines a shared LoRA update with a bank of K routed LoRA experts whose mixture weights are determined by the metadata representation h_i^m. We apply these updates to the query and value projections of the last L_a Transformer blocks in the pretrained DINOv3 visual backbone (Siméoni et al., 2025), while keeping the backbone weights frozen. For projection $u \in \{q, v\}$ at adapted layer ℓ, each update follows the standard rank-r LoRA parameterization $\Delta W_{e,u}^{(\ell)} = (\eta/r) A_{e,u}^{(\ell)} B_{e,u}^{(\ell)}$, where $A_{e,u}^{(\ell)}$ and $B_{e,u}^{(\ell)}$ are the learnable low-rank factors, $e \in \{0, \ldots, K\}$ indexes the updates, r is the LoRA rank, and η is the scaling factor.

The metadata representation h_i^m determines the mixture of experts used at each adapted layer. Let $\rho_i = \mathbb{I}[|\mathcal{F}_i| > 0]$ indicate whether at least one metadata field is available. When $\rho_i = 1$, a layer-specific router $R^{(\ell)}$ computes the expert weights from h_i^m. For image x_i,

$$\pi_i^{(\ell)} = \text{softmax}\Big(R^{(\ell)}(h_i^m)\Big), \quad \Delta W_{i,u}^{(\ell)} = \Delta W_{0,u}^{(\ell)} + \sum_{k=1}^{K} \pi_{i,k}^{(\ell)} \Delta W_{k,u}^{(\ell)}. \tag{3}$$

Here, $\pi_{i,k}^{(\ell)}$ is the routing weight assigned to expert k. The shared update is always applied, while the routed mixture is active only when metadata is available; if $\mathcal{F}_i = \emptyset$, $\rho_i = 0$, and MetaRouter reduces to the shared LoRA path.

Training Objective and Inference. We train MetaRouter using standard ReID supervision, with $\mathcal{L}_{\text{ReID}} = \mathcal{L}_{\text{id}} + \lambda_{\text{tri}} \mathcal{L}_{\text{tri}}$, where $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{tri}}$ denote the identity-classification and batch-hard triplet losses, respectively. The metadata encoder, layer-specific routers, shared and routed LoRA parameters, and ReID head are optimized jointly, while the pretrained DINOv3 backbone remains frozen.

To improve robustness to incomplete metadata, we apply metadata dropout during training. For each image, let $\widetilde{\mathcal{F}}_i \subseteq \mathcal{F}_i$ denote the metadata fields retained after dropout. We randomly remove individual fields and, with a separate probability, the entire metadata record. All metadata encoding and routing are then performed using $\widetilde{\mathcal{F}}_i$, with $\rho_i = \mathbb{I}[|\widetilde{\mathcal{F}}_i| > 0]$ indicating whether any metadata remains after dropout. Full-record dropout, therefore, exposes the model to the shared-LoRA-only path, while field-level dropout exposes it to different subsets of the available metadata.

At inference, metadata dropout is disabled and each query and gallery image is encoded independently using its available metadata. The resulting visual features are ℓ_2-normalized and ranked by

cosine similarity. Metadata, therefore, conditions image-level feature extraction but does not directly enter the query–gallery similarity.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We conduct all experiments on MetaFauna using fixed identity-disjoint train/test partitions for each species, with test images divided into disjoint query and gallery sets and the same splits used across all methods. Models are trained independently for each species and evaluated using mAP (Wu et al., 2017) and Rank-1 accuracy (Moon & Phillips, 2001) over five random seeds; hyperparameters are selected on fixed identity-disjoint validation subsets constructed from the training partitions. We compare MetaRouter with CLIP-ReID (Li et al., 2023), UniReID (Jiao et al., 2023), AdaFreq (Li et al., 2024), ARBase (Hou et al., 2025), CARE (Wu et al., 2026), and MFA-enhanced variants of CLIP-ReID and UniReID (Li et al., 2025). We additionally evaluate a single model jointly trained across all eight species. Detailed split statistics, validation and model-selection procedures, metric definitions, training settings, and the multi-species training protocol are provided in Appendix C.

Table 1: **Main results on MetaFauna.** Each entry reports mAP/Rank-1 (%). The best result for each species and metric is shown in bold.

Method	Cheetah	Elephant	Giraffe	Hyena	Kudu	Leopard	Lion	Zebra
Visual-only ReID								
CLIP-ReID	41.0/43.9	68.9/81.1	73.9/83.8	46.8/74.4	70.6/81.5	43.4/80.2	64.1/70.8	68.0/84.2
UniReID	43.0/43.4	65.7/79.8	73.2/82.2	43.7/73.6	70.0/76.7	39.1/75.8	65.5/72.1	65.8/83.2
AdaFreq	47.8/48.1	69.8/80.0	77.4/85.4	56.2/80.2	77.7/83.4	45.0/81.5	70.3/77.3	70.5/83.6
ARBase	48.9/47.5	69.9/79.6	75.2/83.9	53.6/80.0	75.0/82.0	46.5/82.7	69.7/74.2	61.6/76.1
CARE	40.5/40.0	69.4/78.2	74.4/83.4	51.5/76.9	71.7/79.3	42.9/76.8	68.7/74.5	67.9/80.8
Metadata-aware ReID								
CLIP-ReID$_{+\mathrm{MFA}}$	44.0/46.4	69.7/81.0	75.7/84.8	51.1/77.1	72.0/82.3	47.0/82.1	65.8/73.0	69.3/84.8
UniReID$_{+\mathrm{MFA}}$	45.3/45.5	66.9/80.2	74.6/83.4	47.4/75.7	71.0/77.5	42.3/77.4	67.1/74.5	66.9/83.6
MetaRouter	**54.2/53.0**	**72.0/83.0**	**79.2/86.4**	**61.5/82.6**	**80.0/84.5**	**51.6/84.0**	**72.6/79.2**	**73.6/85.5**

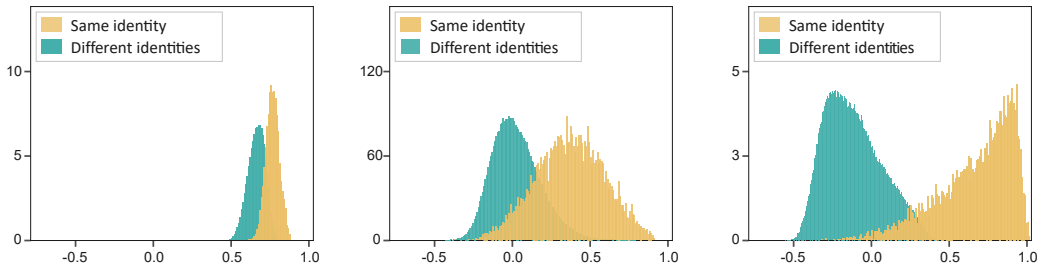
5.2 MAIN RESULTS

Overall Performance. Table 1 compares visual-only and metadata-aware ReID methods on MetaFauna. Adding MFA generally improves its corresponding visual-only baseline, although the magnitude of the gain varies considerably across species and between mAP and Rank-1. MetaRouter achieves the best mAP and Rank-1 performance on all eight species, improving mAP over the strongest visual-only method by 1.8–5.3 points. The largest gains occur on Cheetah, Hyena, and Leopard, which also exhibit the lowest performance among the strongest visual-only baselines, while smaller but consistent improvements are observed on species with stronger visual-only retrieval. This pattern suggests that structured metadata is particularly useful when visual evidence alone provides a less reliable identity representation, a relationship we examine more directly in the analyses below.

Similarity Distribution. Figure 4 provides a feature-space view of the retrieval differences on Hyena, one of the more challenging MetaFauna species. The CLIP-ReID+MFA and UniReID+MFA retain substantial overlap between same-identity and different-identity pairs, whereas MetaRouter produces a clearer separation between the two distributions. This is consistent with the larger mAP gain observed on Hyena in Table 1, suggesting that metadata-conditioned adaptation improves not only the top-ranked prediction but the separation of identity matches throughout the retrieval space.

5.3 ABLATION AND ANALYSIS

We use controlled experiments to examine whether MetaRouter's gains arise from additional adaptation capacity, input-dependent routing, or metadata paired with each image, and further analyze



(a) CLIP-ReID+MFA (b) UniReID+MFA (c) MetaRouter

Figure 4: Similarity distributions of (a) CLIP-ReID+MFA, (b) UniReID+MFA, and (c) our MetaRouter on Hyena. The horizontal axis represents cosine similarity, and the vertical axis represents probability density.

which metadata groups contribute to retrieval and how the model behaves when metadata is incomplete. Detailed control definitions and evaluation protocols are provided in Appendix D.

Table 2: **Ablation of metadata conditioning and expert routing.** Meta, Route, and Experts indicate metadata input, input-dependent routing, and the LoRA expert bank, respectively. A superscript $*$ denotes shuffled metadata records. Entries report mAP (%).

Variant	Meta	Route	Experts	Cheetah	Elephant	Giraffe	Hyena	Kudu	Leopard	Lion	Zebra
Shared LoRA	–	–	–	49.1	70.5	77.1	55.3	76.9	46.4	69.9	69.2
Static Expert Mix	–	–	✓	49.8	70.4	77.0	55.9	77.5	47.2	69.7	69.8
Feature Fusion	✓	–	✓	51.7	71.1	78.1	58.9	78.3	49.3	71.0	71.8
Visual Router	–	✓	✓	50.7	71.0	78.4	58.0	78.8	48.5	71.4	72.0
Shuffled MetaRouter	✓*	✓	✓	49.3	70.8	76.6	54.8	77.9	46.5	70.2	69.3
MetaRouter	✓	✓	✓	**54.2**	**72.0**	**79.2**	**61.5**	**80.0**	**51.6**	**72.6**	**73.6**

Where does the improvement come from? Table 2 separates the effects of expert capacity, input-dependent routing, and metadata conditioning. Static Expert Mix remains close to Shared LoRA, with small gains for some species and slight decreases for others, indicating that simply adding LoRA experts is insufficient. Feature Fusion and Visual Router both improve over Static Expert Mix, but their relative performance varies across species, and the usefulness of explicit metadata integration and visually inferred routing depends on the species and observation conditions. With shuffled metadata, performance returns to roughly the level of Static Expert Mix, whereas MetaRouter achieves the strongest result on all eight species. The gains over the partial controls are largest on Cheetah, Hyena, and Leopard, indicating that the benefit comes from using the metadata paired with each image to guide visual adaptation.

Table 3: **Metadata-group ablation on MetaFauna.** G1 denotes visual evidence visibility, G2 capture conditions, and G3 environmental context. Entries report mAP (%).

Metadata input	G1	G2	G3	Cheetah	Elephant	Giraffe	Hyena	Kudu	Leopard	Lion	Zebra
None	–	–	–	49.8	70.4	77.0	55.9	77.5	47.2	69.7	69.8
G1	✓	–	–	52.6	70.8	77.9	59.6	78.1	50.3	70.8	72.4
G2	–	✓	–	51.3	71.2	78.4	58.1	79.1	48.9	71.6	71.5
G3	–	–	✓	50.4	70.6	77.6	N/A	78.0	N/A	70.3	70.7
G1+G2	✓	✓	–	53.7	71.7	78.9	**61.5**	79.5	**51.6**	72.1	73.0
All	✓	✓	✓	**54.2**	**72.0**	**79.2**	**61.5**	**80.0**	**51.6**	**72.6**	**73.6**

Which metadata is useful? To examine which metadata groups contribute to retrieval, Table 3 evaluates G1, G2, and G3 individually and in combination. Both G1 and G2 improve retrieval across all eight species, but their relative contributions vary: G1 yields larger gains for Cheetah, Hyena, Leopard, and Zebra, whereas G2 is more effective for Giraffe, Kudu, and Lion. The difference is

particularly pronounced on Hyena and Leopard, where visibility metadata accounts for a large part of the improvement. Combining G1 and G2 improves over either group alone for every species, suggesting that visual evidence visibility and capture conditions provide complementary information for retrieval. This indicates that describing which body regions are visible and how visual evidence is captured provides complementary information, especially when either description alone is insufficient. G3 produces smaller improvements where available and provides further gains when added to G1+G2. Overall, the results identify visual evidence visibility and capture conditions as the primary sources of improvement, while environmental context provides a smaller additional signal.

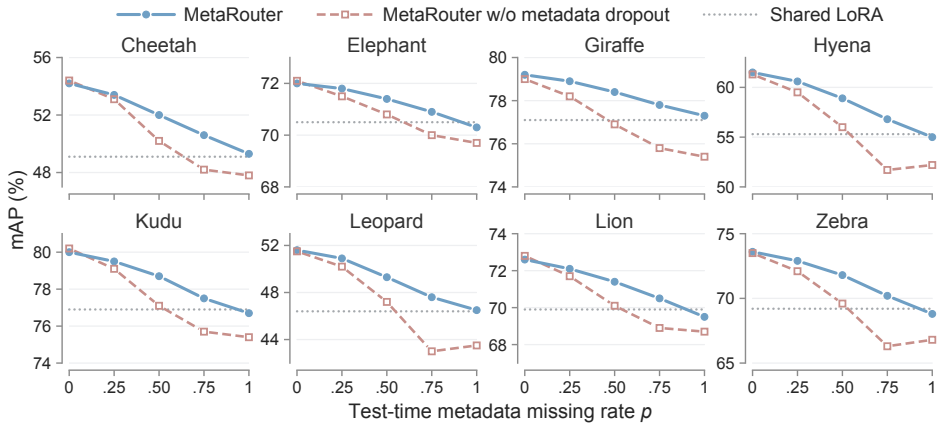


Figure 5: **Robustness to incomplete metadata.** Each originally observed metadata field is independently masked at inference with probability p, where $p = 0$ preserves the original metadata, and $p = 1$ removes all observed metadata. We compare MetaRouter with a no-dropout variant; the dotted line denotes Shared LoRA.

How robust is MetaRouter when metadata is incomplete? To evaluate MetaRouter under incomplete metadata, we progressively remove observed metadata fields during inference. At $p = 0$, metadata dropout has little effect on mAP, while the no-dropout variant degrades more rapidly as metadata is removed. At $p = 0.5$, MetaRouter loses 1.6 mAP on average from its $p = 0$ performance, compared with 3.4 points without dropout. Cheetah, Hyena, Leopard, and Zebra show the largest drops, consistent with their larger gains from metadata in the complete setting. At $p = 0.75$, MetaRouter remains above Shared LoRA on all eight species, whereas the no-dropout variant falls below it on all eight. On Hyena, Leopard, and Zebra, the no-dropout variant performs slightly worse at $p = 0.75$ than at $p = 1$, showing that severely incomplete metadata can be more harmful than an empty record when the router has not been trained on sparse field combinations. When $p = 1$, MetaRouter switches to its shared-only path and stays within 0.4 mAP of Shared LoRA across all species, while the no-dropout variant remains 0.8–3.1 points lower. Metadata dropout therefore preserves the benefit of available metadata while making degradation under missingness more gradual and maintaining a reliable metadata-free fallback. Additional ablations in Appendix D.4 separate the effects of field-level and full-record dropout and evaluate systematic failures in metadata sources.

6 Conclusion

In this paper, we studied whether making observation conditions explicit through structured metadata can improve animal ReID. We introduced MetaFauna, which pairs animal images and identity annotations with structured metadata describing visual evidence visibility, capture conditions, and environmental context, together with MetaRouter, a reference framework that uses this metadata to condition LoRA-based visual adaptation. Across all eight species, MetaRouter outperforms visual-only and metadata-aware baselines; controlled and robustness analyses show that the gains depend on the metadata paired with each image, with visual evidence visibility and capture conditions accounting for most of the observed improvement, while performance degrades gradually under missing metadata and retains a metadata-free fallback. We hope MetaFauna provides a useful basis for studying when and how structured metadata can improve animal ReID, and supports broader applications of AI in wildlife research, ecology, and biodiversity conservation.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments and to assist with implementing parts of the proposed method and experimental pipeline. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, assist with translation, clean or reformat datasets, or interpret experimental results; tasks concerning mathematical claims and proofs and qualitative or thematic data analysis are not applicable to this work. Additionally, generative AI tools were used to edit parts of the manuscript to improve readability. All AI-assisted methodological feedback and implementation suggestions were reviewed by the authors, and any AI-assisted code used in the experiments was manually inspected and tested before use. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MetaFauna is constructed from previously released wildlife and camera-trap datasets and does not involve new field data collection, animal handling, intervention, or human-subject research. MetaFauna does not redistribute the source images; instead, the release provides the annotations introduced in this work together with source identifiers that allow users to obtain the corresponding images from Snapshot Safari, Hyena ID 2022, and Leopard ID 2022 under their respective licenses and citation requirements. Annotations created for MetaFauna are released under the CDLA-Permissive-1.0 license. MetaFauna is intended to support non-invasive animal ReID, wildlife monitoring, and related ecological and conservation research. To limit the release of sensitive wildlife location information, fine-grained camera-site coordinates used during metadata construction are not included in MetaFauna annotations; only the derived environmental fields are released.

REPRODUCIBILITY STATEMENT

The procedures needed to reproduce MetaFauna and the MetaRouter experiments are documented throughout the paper and appendix. MetaFauna construction, metadata specification, and annotation procedures are described in Section 3 and Appendix A; the MetaRouter formulation, training and evaluation protocols, and implementation details are provided in Sections 4 and 5.1 and Appendix C; and the controlled analyses are detailed in Section 5.3 and Appendix D. An anonymized implementation is available at `https://anonymous.4open.science/r/MetaFauna`. Upon acceptance, we will publicly release the MetaFauna annotations, source-image mappings, and evaluation splits, together with documentation for reconstructing the dataset from the original source datasets.

REFERENCES

Lukas Adam, Vojtech Cermak, Kostas Papafitsoros, and Lukas 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*, pp. 2124–2134, 2025.

T Michael Anderson, Staci A Hepler, Ricardo M Holdo, Jason E Donaldson, Robert J Erhardt, J Grant C Hopcraft, Matthew C Hutchinson, Sarah E Huebner, Thomas A Morrison, Jeffry Muday, et al. Interplay of competition and facilitation in grazing succession by migrant serengeti herbivores. *Science*, 383(6684):782–788, 2024.

Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review, 2019.

Sara Meghan Beery. *Where the Wild Things Are: Computer Vision for Global-Scale Biodiversity Monitoring*. California Institute of Technology, 2023.

Botswana Predator Conservation Trust. Hyena id 2022. Labeled Information Library of Alexandria: Biology and Conservation (LILA BC), 2022a. URL `https://lila.science/datasets/hyena-id-2022/`.

Botswana Predator Conservation Trust. Leopard id 2022. Labeled Information Library of Alexandria: Biology and Conservation (LILA BC), 2022b. URL `https://lila.science/datasets/leopard-id-2022/`.

Jakub W Bubnicki, Ben Norton, Steven J Baskauf, Tom Bruce, Francesca Cagnacci, Jim Casaer, Marcin Churski, Joris PGM Cromsigt, Simone Dal Farra, Christian Fiderer, et al. Camtrap dp: an open standard for the fair exchange and archiving of camera trap data. *Remote Sensing in Ecology and Conservation*, 10(3):283–295, 2024.

Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5953–5963, 2024.

International Electrotechnical Commission et al. Multimedia systems and equipment-color measurement and management-part 2-1. *Color Management-Default RGB Color Space-sRGB*, 1999.

Saihui Hou, Panjian Huang, Zengbin Wang, Yuan Liu, Zeyu Li, Man Zhang, and Yongzhen Huang. Openanimals: Revisiting person re-identification for animals towards better generalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14369–14379. IEEE, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Bingliang Jiao, Lingqiao Liu, Liying Gao, Ruiqi Wu, Guosheng Lin, Peng Wang, and Yanning Zhang. Toward re-identifying any animal. *Advances in Neural Information Processing Systems*, 36:40042–40053, 2023.

Örjan Johansson, Charudutt Mishra, Guillaume Chapron, Gustaf Samelius, Purevjav Lkhagvajav, Tom McCarthy, and Matthew Low. Seasonal variation in daily activity patterns of snow leopards and their prey. *Scientific Reports*, 12(1):21681, 2022.

Chenyue Li, Shuoyi Chen, and Mang Ye. Adaptive high-frequency transformer for diverse wildlife re-identification. In *European Conference on Computer Vision*, pp. 296–313. Springer, 2024.

Shuyuan Li, Jianguo Li, Hanlin Tang, Rui Qian, and Weiyao Lin. ATRW: A benchmark for amur tiger re-identification in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2590–2598, 2020. doi: 10.1145/3394171.3413569.

Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1405–1413, 2023.

Yuzhuo Li, Di Zhao, Tingrui Qiao, Yihao Wu, Bo Pang, and Yun Sing Koh. Metawild: A multimodal dataset for animal re-identification with environmental metadata. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13009–13015, 2025.

Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, 2019.

Josephine Mahony, Ellen Dyer, and Richard Washington. The precipitation patterns and atmospheric dynamics of the serengeti national park. *International Journal of Climatology*, 41:E2051–E2072, 2021.

Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001. doi: 10.1068/p2896.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13:4349–4383, 2021. doi: 10.5194/essd-13-4349-2021.

Hemal Naik, Junran Yang, Dipin Das, Margaret C Crofoot, Akanksha Rathore, and Vivek H Sridhar. Bucktales: A multi-uav dataset for multi-object tracking and re-identification of wild antelopes. *Advances in Neural Information Processing Systems*, 37:81992–82009, 2024.

Eugênio Dias Ribeiro Neto, Marc Chaumont, Gérard Subsol, Michel de Garine-Wichatitsky, and Hélène Guis. An improved architecture for part-based animal re-identification through semantic segmentation distillation. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5280–5289. IEEE, 2026.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19001–19012. IEEE, 2022.

Kostas Papafitsoros, Aliki Panagopoulou, and Gail Schofield. Social media reveals consistently disproportionate tourism pressure on a threatened marine vertebrate. *Animal Conservation*, 24 (4):568–579, 2021.

Lain E Pardo, Sara Bombaci, Sarah E Huebner, Michael J Somers, Herve Fritz, Colleen Downs, Abby Guthmann, Robyn S Hetem, Mark Keith, Aliza le Roux, et al. Snapshot safari: A large-scale collaborative to monitor africa's remarkable biodiversity. *South African Journal of Science*, 117(1-2):1–4, 2021.

Maarten Perneel, Ines Adriaens, Jan Verwaeren, and Ben Aernouts. Dynamic multi-behaviour, orientation-invariant re-identification of holstein-friesian cattle. *Sensors*, 25(10):2971, 2025.

Said Pertuz, Domenec Puig, and Miguel Angel Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013.

Kai Ren, Chuanping Hu, Hao Xi, Yongqiang Li, Jinhao Fan, and Lihua Liu. Mosce-reid: Mixture of semantic clustering experts for person re-identification. *Neurocomputing*, 626:129587, 2025.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Maximilian Schall, Felix Leonard Knöfel, Noah Elias König, Jan Jonas Kubeler, Maximilian von Klinski, Joan Wilhelm Linnemann, Xiaoshi Liu, Iven Jelle Schlegelmilch, Ole Woyciniuk, Alexandra Schild, et al. Gorillawatch: An automated system for in-the-wild gorilla re-identification and population monitoring. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8364–8375. IEEE, 2026.

Stefan Schneider, Graham W Taylor, Stefan Linquist, and Stefan C Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470, 2019.

Gail Schofield, Kostas Papafitsoros, Chloe Chapman, Akanksha Shah, Lucy Westover, Liam CD Dickson, and Kostas A Katselidis. More aggressive sea turtles win fights over foraging resources independent of body size and years of presence. *Animal Behaviour*, 190:209–219, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

TK Sørdalen, Ketil Malde, C Sauvaitre, AB Skiftesvik, Cigdem Beyan, T Larsen, and KT Halvorsen. A wild fish image dataset for individual re-identification and phenotyping. *Scientific Data*, 13(1): 708, 2026.

Avirath Sundaresan, Jason Parham, Jonathan Crall, Rosemary Warungu, Timothy Muthami, Jackson Miliko, Margaret Mwangi, Jason Holmberg, Tanya Berger-Wolf, Daniel Rubenstein, et al. Adapting the re-id challenge for static sensors. *IET Computer Vision*, 19(1):e70027, 2025.

Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2(1):150026, 2015.

Anil Osman Tur, Tonje Knutsen Sørdalen, Kim Tallaksen Halvorsen, and Cigdem Beyan. Parameter-efficient vision–language adaptation with continuous metadata conditioning for animal re-identification. *Expert Systems with Applications*, pp. 133618, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14205–14215. IEEE, 2024a.

Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024b.

Yihao Wu, Di Zhao, Yuzhuo Li, Matthew Alajas, Alistair S Glen, Jingfeng Zhang, Gillian Dobbie, Daniel Wilson, and Yun Sing Koh. Overcoming fine-grained visual challenges in animal re-identification via semantic feature alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 371–381, 2026.

Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *European Conference on Computer Vision*, pp. 372–388. Springer, 2022.

Pengfei Xu, Yuanyuan Zhang, Minghao Ji, Songtao Guo, Zhanyong Tang, Xiang Wang, Jing Guo, Junjie Zhang, and Ziyu Guan. Advanced intelligent monitoring technologies for animals: A survey. *Neurocomputing*, 585:127640, 2024.

APPENDIX

This appendix provides supplementary dataset details, methodological formulation, experimental settings, and additional analyses. It is organized as follows:

- Appendix A provides additional details on dataset construction and quality control, followed by the complete metadata specification and representative image–metadata examples.
- Appendix B provides the formal problem formulation and the complete MetaRouter training and inference algorithm.
- Appendices C–E provide the evaluation and implementation details, additional ablations and metadata analyses, and supplementary quantitative and qualitative results.

A ADDITIONAL METAFAUNA DATASET DETAILS

A.1 DATASET CONSTRUCTION DETAILS

A.1.1 IMAGE SELECTION AND INSTANCE CURATION

Where reliable instance-level annotations are provided by the source dataset, we use the corresponding bounding boxes or animal crops directly. For source images without such annotations, we localize animal instances using a fixed pretrained MegaDetectorV6 model (Beery et al., 2019). We use the same detector checkpoint and confidence threshold across species and keep the detector weights fixed throughout dataset construction.

Each detected animal is treated as an independent candidate instance. Multi-animal images are therefore not discarded by default: a source image may contribute multiple ReID samples when the individual animals can be localized separately and subsequently assigned to identities without ambiguity. Each candidate crop included in the released dataset is manually reviewed against the source frame to confirm the target species, verify that the box corresponds to a single animal instance, and correct inaccurate localization when necessary. Candidates for which the target instance cannot be established reliably are excluded.

For a bounding box of width w and height h, we expand the box by $0.08w$ on the left and right and by $0.08h$ on the top and bottom, clipping the resulting coordinates to the source image boundary. The unpadded box is retained separately for image-level measurements, such as target scale. Occlusion, truncation, blur, illumination variation, unusual viewpoints, and small target scale are not themselves exclusion criteria when the animal instance remains identifiable.

Exclusion criteria. Figure A.1 illustrates representative cases excluded during dataset construction. Multi-animal images are retained when individual target instances can be localized and assigned separately and unambiguously; candidate instances are excluded when this separation cannot be reliably established. We also exclude frames without a valid target animal and observations in which visual degradation leaves insufficient evidence for reliable identity annotation. Moderate occlusion, truncation, blur, illumination variation, unusual viewpoints, and small target scale are otherwise retained.

A.1.2 IDENTITY ANNOTATION AND VERIFICATION

For the Snapshot Safari, we first group crops corresponding to the same animal within each source sequence. Annotators inspect the complete sequence when establishing these within-sequence groups, so that adjacent frames can provide complementary views of the animal. When multiple animals are present, their crops are grouped separately across frames. Cases in which the correspondence between animals across adjacent frames cannot be established reliably are kept separate.

For each resulting group, we select the temporal midpoint image as its representative frame and extract a DINOv3 visual embedding (Siméoni et al., 2025). Candidate groups are ranked by embedding similarity within the same species and presented in this order for manual identity comparison. The embedding similarity is used only to prioritize candidate comparisons and is not used as evi-

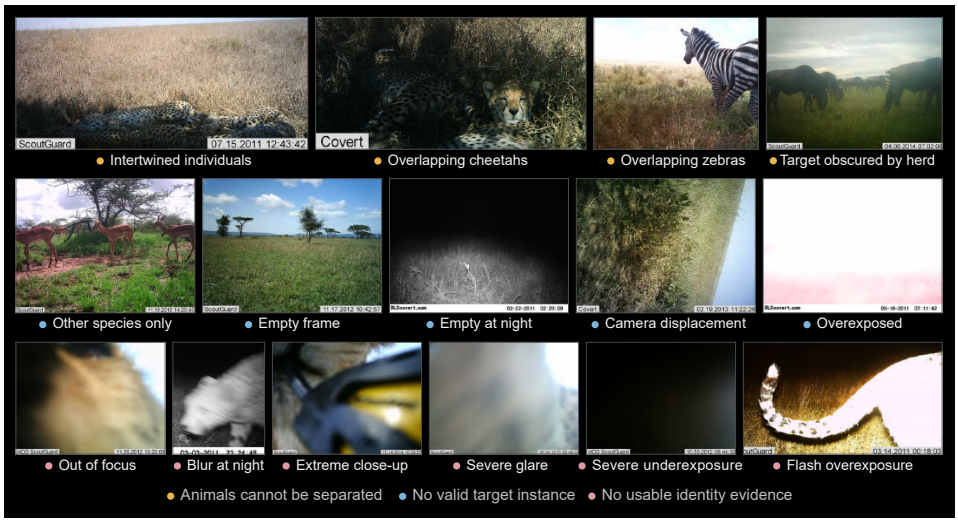


Figure A.1: **Representative exclusion cases during MetaFauna construction.** Examples illustrate ambiguous target instances, frames without a valid target animal, and observations with insufficient visual evidence for reliable identity annotation. Challenging observation conditions are otherwise retained when the target can be reliably localized and identified.

dence for assigning an identity label. During annotation, reviewers inspect the complete sequences associated with both candidate groups rather than only the representative frames used for retrieval.

Each candidate pair is independently reviewed by two annotators and assigned one of three labels: *same individual*, *different individuals*, or *unresolved*. Annotators compare persistent visual cues that can distinguish individuals within the corresponding species, including coat or stripe patterns, local markings, facial features, horn morphology, scars, and other stable appearance characteristics when available. A comparison is marked as *unresolved* when the available visual evidence is insufficient to support either a same-individual or a different-individual decision.

If both annotators select *different*, the candidate groups remain separate. If both select *same*, a third annotator independently reviews the pair, and the association is accepted only when all three annotators agree on *same*. Any disagreement between the first two annotators, as well as any comparison containing an *unresolved* decision, is also reviewed by the third annotator using the complete sequences. Cases that remain unresolved after review do not create an identity association.

Confirmed same-individual associations are combined to form individual identities. Before release, all sequence groups assigned to the same identity are reviewed together to check that the identity remains visually consistent across the associated sequences. Ambiguous associations identified during this review are returned for re-evaluation rather than propagated to the final identity labels.

Hyena ID 2022 and Leopard ID 2022 provide manually established identity labels in their source releases. For these subsets, MetaFauna retains the source identities and verifies the correspondence between the released samples and their source identity records.

A.2 Metadata Specification

Table A.1 summarizes the metadata fields in MetaFauna, including their representations, definitions, and construction sources. The fields are organized into three groups: visual evidence visibility (G1), capture conditions (G2), and environmental context (G3). The following sections detail the annotation rules, computation procedures, and external-data matching used to construct these fields.

A.2.1 Visual Evidence Visibility (G1)

G1 describes the visual evidence available for identifying the target animal through viewpoint, visible regions, external occlusion, and boundary contact. Annotations are made from the original

Table A.1: Summary of the metadata fields in MetaFauna.

Field	Representation	Definition	Construction
		Visual Evidence Visibility (G1)	
Viewpoint	Categorical	Camera-relative viewing orientation of the animal.	Manual annotation
Visible regions	Multi-label	Body regions visible in the image.	Manual annotation
External occlusion	Binary: present / absent	Whether external objects or animals occlude the target.	Manual annotation
Boundary contact	Boolean: left / top / right / bottom	Frame boundaries contacted by the target animal.	Geometry + manual
		Capture Conditions (G2)	
Imaging mode	Categorical	Image acquisition mode associated with the capture.	Source / camera metadata
Brightness	Continuous	Brightness within the manually verified animal box.	Image-derived
Sharpness	Continuous	Sharpness within the manually verified animal box.	Image-derived
BBox area ratio	Continuous: $[0, 1]$	Verified box area relative to the original frame area.	BBox-derived
		Environmental Context (G3)	
Solar period	Categorical: dawn / day / dusk / night	Astronomical light period at image capture.	Timestamp + location
Ecological season	Categorical	Ecological season associated with the study area and capture date.	Study area + date
Temperature	Continuous	Temperature associated with the study area and capture date.	External data
Precipitation	Continuous	Precipitation accumulated over a fixed pre-capture period.	External data

source frame with the verified target bounding box indicating the target instance. These fields are annotated independently, allowing partial visibility and their source to be represented separately.

Viewpoint. We retain the four-category granularity used for face direction in MetaWild (Li et al., 2025), but define viewpoint from the orientation of the animal body rather than the head alone. The categories are *front*, *rear*, *left*, and *right*, where left and right refer to the anatomical flank exposed to the camera. An animal may turn its head without changing which body surface is visible, whereas body orientation determines which flank markings and anatomical regions are available for comparison (Perneel et al., 2025; Sundaresan et al., 2025). For oblique views, annotators assign the nearest category when one body aspect is clearly predominant. The value is recorded as *unknown* when the image has been assessed but no viewpoint can be assigned reliably from the visible body structure, for example because of severe occlusion or truncation, extreme foreshortening, an ambiguous pose, or insufficient visual detail. Figure A.2 shows representative annotations for all five recorded values across the eight species.

Visible regions. Visible regions are represented as a multi-label annotation over six coarse anatomical regions: *head*, *neck*, *torso*, *forelimbs*, *hindlimbs*, and *tail*. These regions provide a coarse grouping of anatomical landmarks commonly used in cross-species animal pose annotation (Ng et al., 2022). A region is marked as visible when a recognizable portion can be identified in the image; full exposure is not required. For *forelimbs* and *hindlimbs*, visibility of at least one corresponding limb is sufficient.

External occlusion. External occlusion is recorded as *present* or *absent*. It is marked as present when another animal or an external scene element lies in front of the target and obscures a recognizable part of its body. Self-occlusion due to the animal's pose is not included, and visibility loss at the image boundary is represented separately as boundary contact.

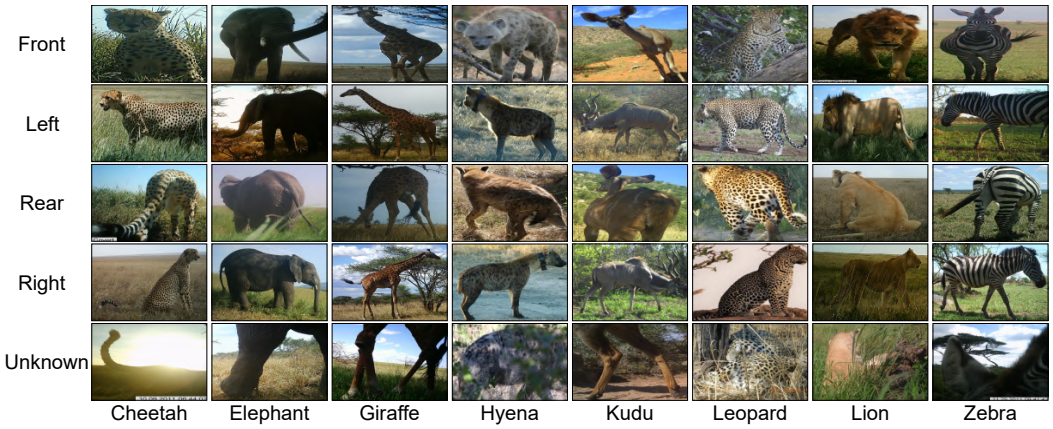


Figure A.2: **Viewpoint annotation in MetaFauna.** Examples of the viewpoint labels across the eight MetaFauna species. *Left* and *right* refer to the anatomical flank predominantly exposed to the camera, while oblique views are assigned to the closest category when one body aspect is clearly predominant. *Unknown* denotes an image that has been assessed but for which body orientation cannot be assigned reliably from the visible body structure.

Boundary contact. Boundary contact records whether the target animal reaches each boundary of the original source frame. It is represented by four Boolean values corresponding to the left, top, right, and bottom boundaries. Let

$$b_i = (x_i^{\min}, y_i^{\min}, x_i^{\max}, y_i^{\max}) \tag{4}$$

denote the verified, unpadded target bounding box in a frame of width W_i and height H_i. Using a fixed relative tolerance ϵ, candidate boundary contacts are computed as

$$\widetilde{c}_i^L = \mathbf{1}\{x_i^{\min}/W_i \le \epsilon\}, \qquad \widetilde{c}_i^R = \mathbf{1}\{(W_i - x_i^{\max})/W_i \le \epsilon\}, \tag{5}$$

$$\widetilde{c}_i^T = \mathbf{1}\{y_i^{\min}/H_i \le \epsilon\}, \qquad \widetilde{c}_i^B = \mathbf{1}\{(H_i - y_i^{\max})/H_i \le \epsilon\}. \tag{6}$$

These geometric flags are used only to initialize the annotation. Annotators inspect the original frame and confirm or correct each value according to whether the animal itself reaches the corresponding image boundary. The final annotation is stored as

$$\mathbf{c}_i = [c_i^L, c_i^T, c_i^R, c_i^B] \in \{0,1\}^4. \tag{7}$$

Boundary contact indicates possible truncation at the source-frame boundary, but does not by itself specify whether a particular body region is truncated.

A.2.2 Capture Conditions (G2)

G2 describes how the target animal is recorded in the image through imaging mode, brightness, sharpness, and target scale. Imaging mode is obtained from source or camera information, where available, and otherwise determined from the source frame, while the remaining fields are computed from the image and the verified, unpadded target bounding box using fixed procedures.

Imaging mode. Imaging mode describes the effective image modality available for ReID and is recorded as *RGB*, *grayscale*, *near-infrared*, or *unknown*. When source records or documented camera information identify near-infrared acquisition, the image is labeled *near-infrared*. Otherwise, we distinguish RGB and grayscale images from the source frame.

Rendering mode is derived from the full source frame, since it describes an image-level property and the color distribution of the target itself may be naturally low in chromatic variation. Let F_i denote the source frame of image i, with channel values R_p, G_p, and B_p normalized to $[0,1]$. We compute the mean channel deviation

$$c_i = \frac{1}{|F_i|} \sum_{p \in F_i} [\max(R_p, G_p, B_p) - \min(R_p, G_p, B_p)]. \tag{8}$$

A threshold τ_c is determined once from a manually reviewed calibration subset and then fixed for the remaining images. Images below the threshold are labeled *grayscale*, while the remaining images are labeled ***RGB***; samples close to the threshold are manually checked. When the imaging mode cannot be established reliably, it is recorded as *unknown*.

Brightness. Brightness is computed within the manually verified animal bounding box so that the field reflects the recorded brightness of the target rather than the surrounding scene. We use mean relative luminance because it provides a fixed, channel-aware measure of recorded brightness for sRGB images (Commission et al., 1999). Let Ω_i denote the pixels within the verified bounding box. For a color image, the sRGB channels are first normalized to $[0, 1]$ and linearized using

$$g(u) = \begin{cases} u/12.92, & u \le 0.04045, \\ \left(\dfrac{u + 0.055}{1.055}\right)^{2.4}, & u > 0.04045. \end{cases} \tag{9}$$

The relative luminance of pixel p is then

$$Y(p) = 0.2126\, g(R_p) + 0.7152\, g(G_p) + 0.0722\, g(B_p), \tag{10}$$

and target brightness is defined as

$$B_i = \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} Y(p). \tag{11}$$

For grayscale and near-infrared images, $Y(p)$ is given by the normalized pixel intensity.

Sharpness. Sharpness describes the amount of local image detail preserved in the target animal, since blur can suppress fine-grained markings and textures used for individual identification. It is computed from the animal crop defined by the manually verified bounding box so that the measurement is not dominated by background texture. To reduce dependence on the original crop resolution, each crop is resized to a fixed spatial scale r_s while preserving its aspect ratio and is then converted to grayscale. We use the variance of the Laplacian, a standard focus measure based on high-frequency image responses (Pertuz et al., 2013). Let $\widetilde{C}_i$ denote the standardized grayscale crop. Sharpness is defined as

$$S_i = \mathrm{Var}\left(\nabla^2 \widetilde{C}_i\right). \tag{12}$$

Larger values correspond to stronger high-frequency image detail. The same resize rule and Laplacian operator are applied to all images.

Target scale. Target scale is represented by the fraction of the original source frame occupied by the manually verified animal bounding box. For bounding box b_i and source image I_i, we define

$$A_i = \frac{\mathrm{area}(b_i)}{\mathrm{area}(I_i)}, \qquad A_i \in [0, 1]. \tag{13}$$

The resulting value is stored as `bbox_area_ratio`.

Figure A.3 provides a visual reference for the three continuous G2 fields. Percentiles are computed over all images with an observed value: 44,317 images for brightness and sharpness and 41,188 for target scale, as the Hyena source release does not provide the detection boxes required for target-scale computation. For visualization only, we manually select one released example within 1.5 percentile points of each target percentile while avoiding repeated camera deployments; the percentile distributions and summary statistics are computed from the full eligible set. Brightness and sharpness examples use the animal crops, center-cropped to a 4:3 aspect ratio with at most 7% removed from one spatial axis. Camera-stamp rows, which are excluded from the corresponding measurements, are also removed from display. Target-scale examples use the original source frames with the verified target box outlined. The three fields span substantial variation across MetaFauna. Brightness has a median of 0.19, with 95% of images below 0.44. Sharpness spans approximately a 45-fold range from P5 to P95 (103 to 4,599). Target scale has a median of 29%, while P95 reaches 100%; 7.8% of images have a target box covering at least 99% of the frame, typically corresponding to close-up observations in which the animal extends beyond the image boundary.

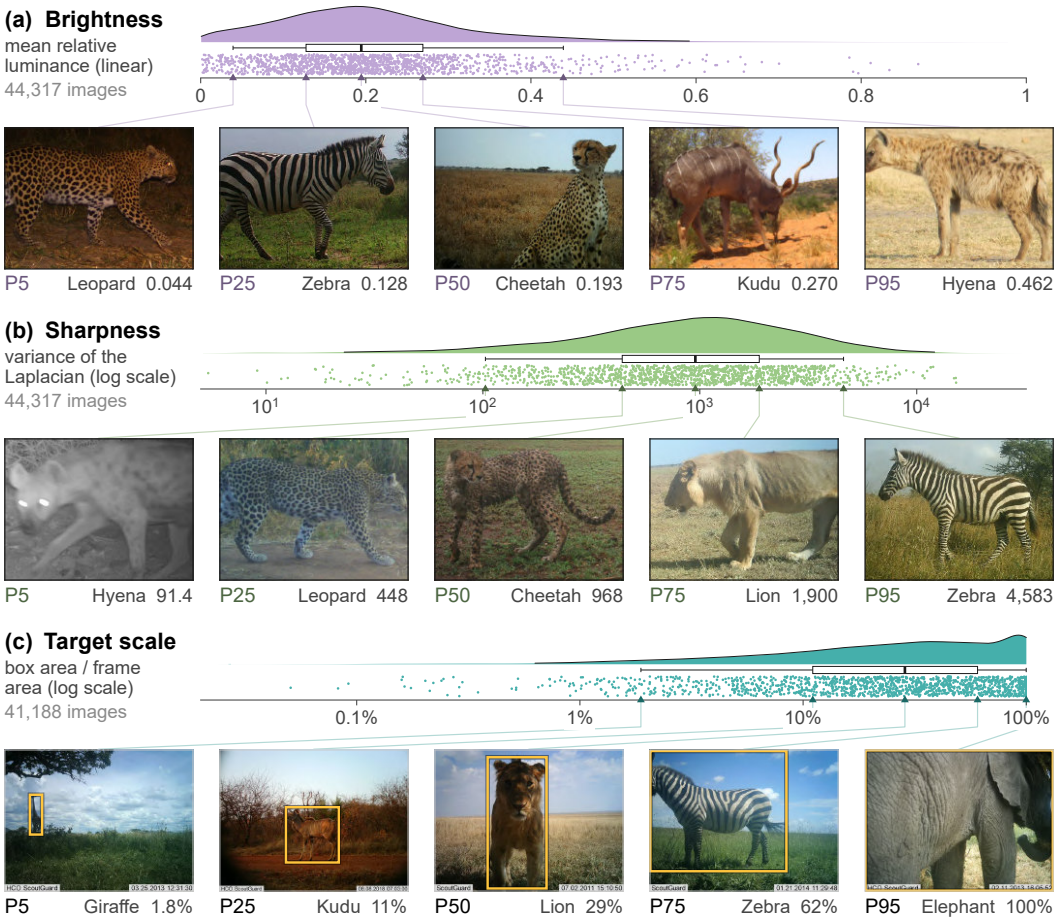


Figure A.3: **Examples across continuous G2 capture conditions.** Representative images near the 5th, 25th, 50th, 75th, and 95th percentiles of brightness, sharpness, and target scale. For each field, the accompanying distribution summarizes all images with an observed value, with box-plot whiskers marking P5 and P95. Brightness and sharpness are shown using the animal crops, whereas target scale is visualized on the source frame with the verified target box outlined.

A.2.3 Environmental Context (G3)

G3 describes the temporal and environmental conditions associated with image capture through solar period, ecological season, air temperature, and precipitation. Solar period is derived from the capture timestamp and location, ecological season from published descriptions of the corresponding study area, and temperature and precipitation from external environmental records. Fields are left missing when the source data does not provide sufficiently reliable spatial or temporal information for matching.

Solar period. Solar period records the astronomical light regime at the time of image capture using four categories: *dawn*, *day*, *dusk*, and *night*. Unlike a fixed clock-based day/night split, the boundaries are calculated based on the capture date and location, allowing changes in sunrise and sunset times to be reflected without additional manual annotation. The four-way representation also preserves the distinction between morning and evening twilight, which is commonly used when describing diel activity patterns in animal ecology (Johansson et al., 2022).

Let t_i^{cdawn}, t_i^{rise}, t_i^{set}, and t_i^{cdusk} denote civil dawn, sunrise, sunset, and civil dusk, respectively, for the location and date associated with image i. We define

$$\text{solar_period}_i = \begin{cases} \texttt{dawn}, & t_i^{\text{cdawn}} \le t_i < t_i^{\text{rise}}, \\ \texttt{day}, & t_i^{\text{rise}} \le t_i < t_i^{\text{set}}, \\ \texttt{dusk}, & t_i^{\text{set}} \le t_i < t_i^{\text{cdusk}}, \\ \texttt{night}, & \text{otherwise.} \end{cases} \tag{14}$$

Sunrise and sunset are computed using the conventional apparent-horizon definition, corresponding to a solar elevation of approximately $-0.833°$, while civil twilight extends until the center of the Sun reaches $-6°$ below the horizon.[3] Reliable camera-site coordinates are used when available; otherwise, a study-area location may be used when its spatial precision is sufficient for the solar calculation. Images without a reliable capture timestamp are assigned a missing value.

Ecological season. Ecological season records the broad seasonal regime associated with image capture using the categories *wet* and *dry*. The mapping from capture date to season is defined separately for each study area using published ecological or climatological descriptions. For Snapshot Safari, the wet season extends from November through May and the dry season from June through October (Mahony et al., 2021). We therefore define

$$\text{ecological_season}(d) = \begin{cases} \texttt{wet}, & d \in \{\text{November}, \ldots, \text{May}\}, \\ \texttt{dry}, & d \in \{\text{June}, \ldots, \text{October}\}. \end{cases} \tag{15}$$

Air temperature. Temperature is represented by the ERA5-Land 2-m air temperature estimate, matched to the location and capture time of each image. ERA5-Land provides hourly land-surface reanalysis fields on a $0.1°$ latitude–longitude grid, including 2-m air temperature (Muñoz-Sabater et al., 2021). For a camera location ℓ_i and capture timestamp t_i, timestamps are first converted to UTC, and the camera location is matched to the corresponding ERA5-Land grid cell. We then use the estimate at the nearest available hour:

$$T_i = T_{\text{ERA5-Land}}\left(\ell_i, \text{nearestHour}(t_i)\right) - 273.15, \tag{16}$$

where ERA5-Land temperature is provided in Kelvin and T_i is stored in degrees Celsius as `air_temperature_2m_c`. When the source data do not provide sufficiently reliable spatial or temporal information for matching, the field is left missing.

Antecedent precipitation. Antecedent precipitation summarizes recent rainfall preceding image capture. We use a four-day accumulation to capture short-term rainfall conditions beyond an individual precipitation event while keeping this field distinct from the broader wet/dry ecological season. Four-day temporal windows and four-day lagged rainfall variables have previously been used in camera-trap analyses of herbivore dynamics within the Serengeti ecosystem (Anderson et al., 2024).

Using the same spatial matching procedure as for air temperature, we accumulate ERA5-Land precipitation over the 96 hours preceding capture:

$$P_i^{(4d)} = 1000 \sum_{h \in (t_i - 96\,\text{h},\, t_i)} P_{\text{ERA5-Land}}(\ell_i, h), \tag{17}$$

where ERA5-Land total precipitation is expressed in metres of water equivalent and the factor 1000 converts the accumulated value to millimetres (Muñoz-Sabater et al., 2021). Only precipitation preceding the capture time is included. The resulting value is stored as `antecedent_precipitation_4d_mm`. As with temperature, the field is left missing when reliable spatial or temporal matching is not possible.

A.3 Representative Image–Metadata Examples

Figure A.4 provides representative image–metadata examples from MetaFauna. Each card shows two observations sharing the same recorded identity label and places their metadata alongside the

[3] Solar and twilight definitions follow the NOAA Global Monitoring Laboratory (`https://gml.noaa.gov/grad/solcalc/calcdetails.html`) and the U.S. Naval Observatory (`https://aa.usno.navy.mil/faq/RST_defs`).

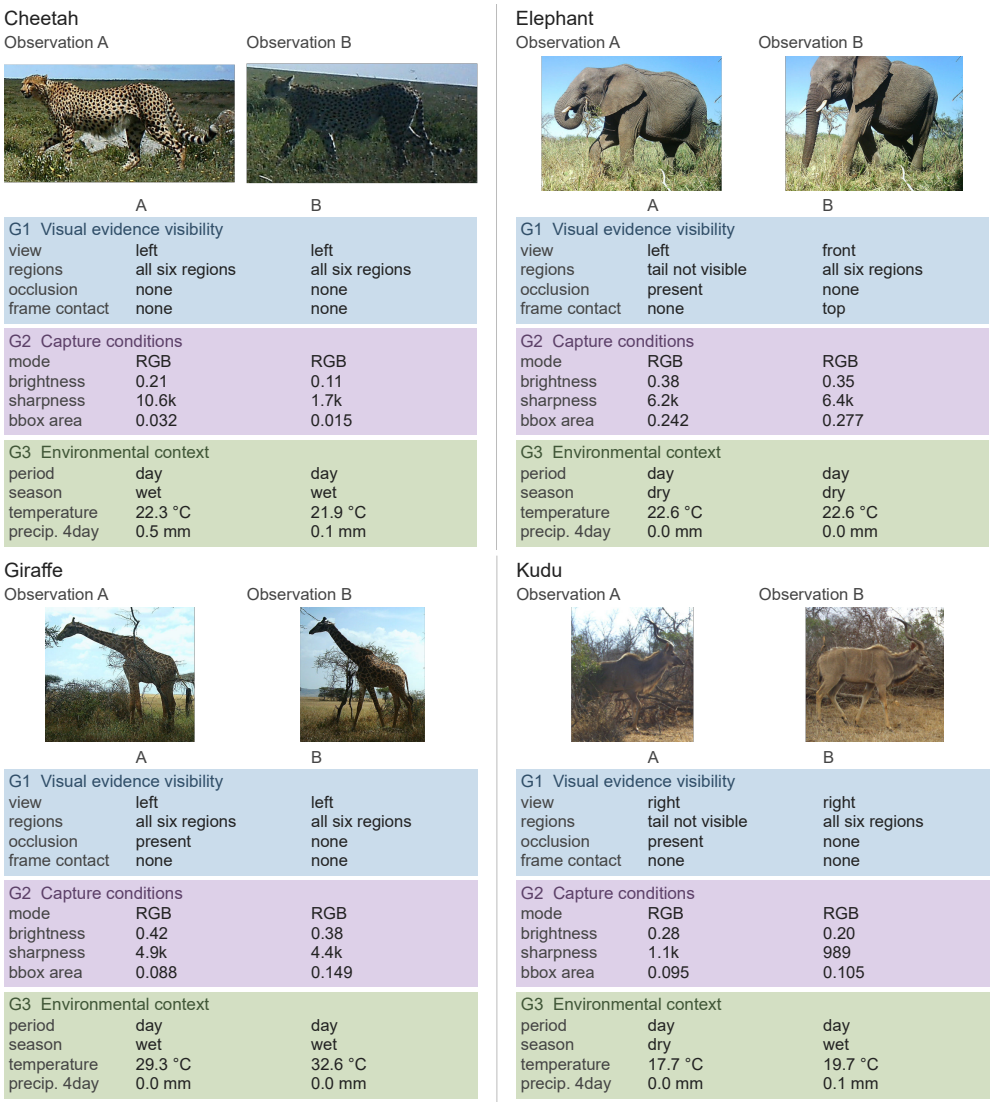


Figure A.4: **Representative image–metadata examples from MetaFauna.** For each species, observations A and B share a recorded identity label and illustrate variation in the visual evidence available for identification and in the conditions under which it is captured. Each card reports visual evidence visibility (G1), capture conditions (G2), and environmental context (G3). Hyena and Leopard use identity labels provided by their source datasets; encounter information for these pairs is unavailable. `unknown` denotes an assessed but unresolved value, whereas `missing` indicates that the field is unavailable.

corresponding images. The selected pairs illustrate how the visibility and appearance of identity-relevant cues vary with viewpoint, visible body regions, external occlusion, boundary contact, rendering mode, brightness, and sharpness. G3 additionally reports the temporal and environmental context available for each observation.

The examples are qualitative and are intended to illustrate the correspondence between images and their structured metadata rather than dataset-wide metadata distributions or the utility of individual fields. Because the displayed crops are scaled independently for visualization, their displayed sizes should not be interpreted as the target scale; `bbox_area_ratio` is defined relative to the original source frame. For Hyena and Leopard, the paired images share the identity labels provided by the source datasets, while encounter information is unavailable.

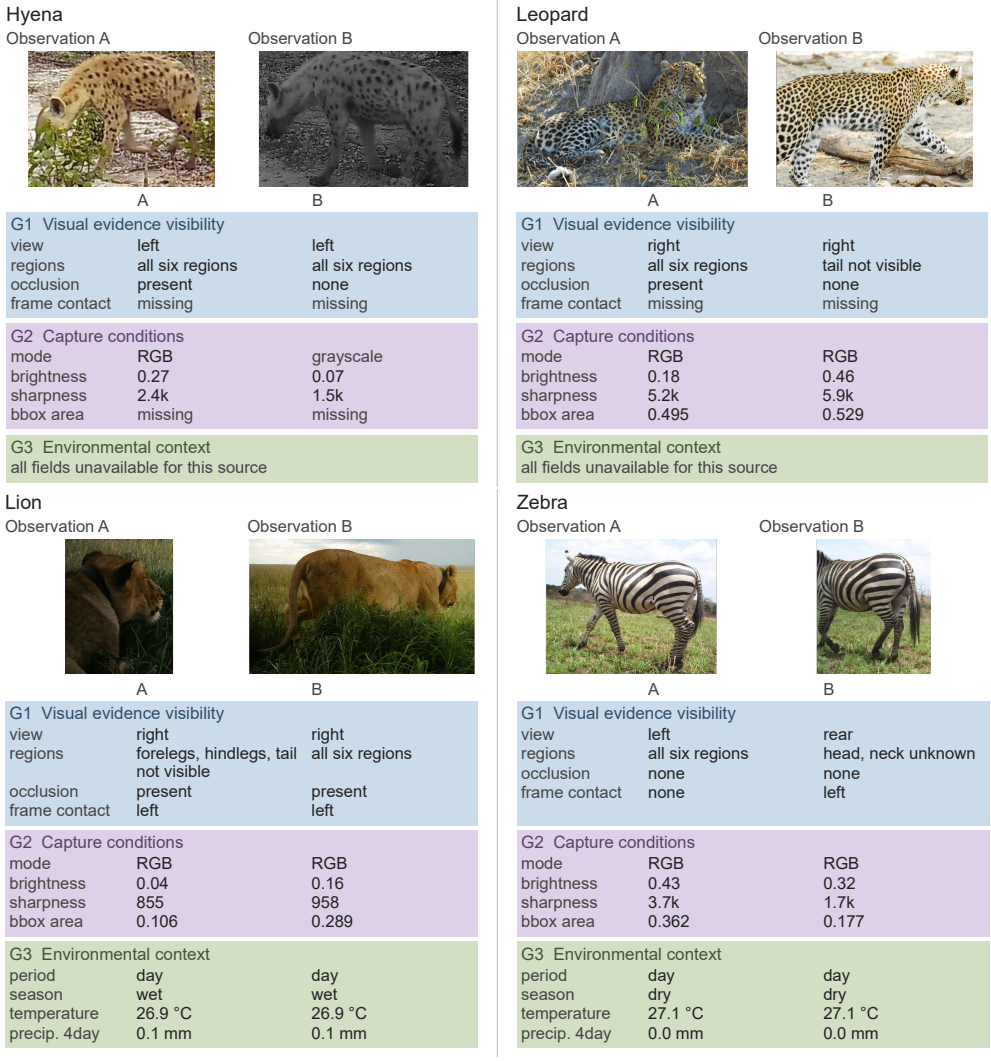


Figure A.4: Representative paired image–metadata examples from MetaFauna (continued).

B MetaRouter Formulation and Algorithm

B.1 Problem Formulation

For completeness, we provide the formal definition of the metadata-conditioned ReID setting used throughout the paper. Animal ReID aims to retrieve images of the same individual across different observations, despite variations in its visual appearance. Given the training data $\mathcal{D} = (\mathcal{X}_\mathcal{D}, \mathcal{Y}_\mathcal{D}) = \{(x_i, y_i^c)\}_{i=1}^{N_D}$, with a size of N_D, where $x_i \in \mathbb{R}^{H\times W\times C}$ represents an image sample and y_i^c is its identity label with class c, we define a gallery set $\mathcal{G} = \{(\mathcal{X}_\mathcal{G}, \mathcal{Y}_\mathcal{G})\} = \{(\beta_j, y_j^c)\}_{j=1}^{N_G}$ and a query set $\mathcal{Q} = \{(\mathcal{X}_\mathcal{Q}, \mathcal{Y}_\mathcal{Q})\} = \{(\alpha_k, y_k^c)\}_{k=1}^{N_Q}$, where $\beta_j \in \mathbb{R}^{H\times W\times C}$ and $\alpha_k \in \mathbb{R}^{H\times W\times C}$ are image samples, and N_G and N_Q denote the sample sizes of the gallery and query sets. At inference time, the identities in $\mathcal{Q}$ and $\mathcal{G}$ are unseen during training, i.e., $\mathcal{Y}_\mathcal{D} \cap \mathcal{Y}_\mathcal{G} = \emptyset$ and $\mathcal{Y}_\mathcal{D} \cap \mathcal{Y}_\mathcal{Q} = \emptyset$, while each query has at least one same-identity sample in the gallery under the evaluation protocol.

Each training image x_i, query image α_k, and gallery image β_j is associated with available metadata m_i, m_k, and m_j, respectively. Accordingly, we extend the standard visual matching function $f : \mathcal{X}_\mathcal{Q} \times \mathcal{X}_\mathcal{G} \to \mathbb{R}$ to a metadata-conditioned matching function $f : (\mathcal{X}_\mathcal{Q} \times \mathcal{M}_\mathcal{Q}) \times (\mathcal{X}_\mathcal{G} \times \mathcal{M}_\mathcal{G}) \to \mathbb{R}$, where $\mathcal{M}_\mathcal{Q} = \{m_k\}_{k=1}^{N_Q}$ and $\mathcal{M}_\mathcal{G} = \{m_j\}_{j=1}^{N_G}$ denote the metadata associated with the query and

gallery images, respectively. Evaluating f over all query–gallery pairs produces a score matrix $S \in \mathbb{R}^{N_Q \times N_G}$, where $S_{k,j} = f((\alpha_k, m_k), (\beta_j, m_j))$. A desirable ReID model should rank positive gallery samples above negative ones; that is, for a query image α_q, a positive gallery image β_p, and a negative gallery image β_r, the objective is $S_{q,p} > S_{q,r}$ when $y_q^c = y_p^c$ and $y_q^c \neq y_r^c$.

B.2 MetaRouter Algorithm

Algorithm 1 summarizes the complete training and inference procedure of MetaRouter. The procedure FORWARD encodes the metadata fields available for an image, predicts one routing distribution per adapted layer, and forms the metadata-conditioned low-rank update in equation 3. The same forward procedure is used during training and inference. During training, METADROP randomly removes individual fields or the complete metadata record as described in Section 4. When no metadata field is available, $\rho = 0$ and the image is processed using the shared low-rank update alone. The ReID head and the placement of the identity and triplet losses follow the BNNeck implementation described in Appendix C.2.

C Experiment Details

C.1 Evaluation Protocol

Train, query, and gallery splits. For each species, we partition identities into disjoint training and test sets while targeting approximately 60%, 15%, and 25% of the images for training, query, and gallery, respectively. All images of an identity are assigned to either the training or test partition, so no identity is shared between them. For the Snapshot Safari species, query–gallery assignment within each test identity is performed at the source-sequence level: all images from the same camera-trap sequence are assigned to the same partition. This prevents adjacent frames from the same source sequence from appearing in both query and gallery. Comparable sequence or encounter information is unavailable for Hyena ID 2022 and Leopard ID 2022, for which query–gallery assignment is therefore performed at the image level. Each test identity contributes at least one sample to both query and gallery. The resulting partitions are fixed across all compared methods and training seeds. Table C.1 summarizes the resulting image and identity counts for all eight species.

Table C.1: **Evaluation splits of MetaFauna.** We report the fixed train, query, and gallery splits used for each species throughout all experiments.

Species	Train		Query		Gallery		Total	
	Imgs	IDs	Imgs	IDs	Imgs	IDs	Imgs	IDs
Cheetah	2,062	180	528	123	847	123	3,437	303
Elephant	3,712	130	898	89	1,586	89	6,196	219
Giraffe	4,459	237	1,036	159	1,937	159	7,432	396
Hyena	1,855	153	425	103	849	103	3,129	256
Kudu	2,535	202	589	146	1,104	146	4,228	348
Leopard	4,084	259	1,021	171	1,701	171	6,806	430
Lion	2,632	139	569	87	1,186	87	4,387	226
Zebra	5,222	198	1,295	124	2,185	124	8,702	322
Total	26,561	1,498	6,361	1,002	11,395	1,002	44,317	2,500

Validation and model selection. For model selection, we construct one fixed identity-disjoint validation split from the training partition of each species. We select validation identities so that their images account for approximately 15% of the original training images, and assign all images of each selected identity to validation. The remaining identities are used for training, ensuring that no identity is shared between training and validation. During sampling, we preserve the distribution of images per identity as closely as possible so that the validation set is not dominated by identities with unusually few or many images. Identities that cannot populate both validation query and gallery under the corresponding source-specific splitting rule remain in the training set. Validation identities are divided into query and gallery sets using the same procedure as the test identities, and the validation partitions are fixed across all methods and training seeds.

Algorithm 1 MetaRouter: metadata-conditioned training and inference.

Require: Training data $\mathcal{D} = (\mathcal{X}_\mathcal{D}, \mathcal{Y}_\mathcal{D}) = \{(x_i, y_i^c)\}_{i=1}^{N_D}$ with associated metadata $m_i = \{(f, v_i^f) \mid f \in \mathcal{F}_i\}$; query set $\mathcal{Q}$ and gallery set $\mathcal{G}$; frozen encoder Enc_v; L_a adapted layers, K routed experts; epochs E; $p_{\mathrm{field}}, p_{\mathrm{rec}}$; trainable parameters θ

Ensure: Query–gallery score matrix $S \in \mathbb{R}^{N_Q \times N_G}$

procedure FORWARD($x, m, \mathcal{F}$)
  $\rho \leftarrow \mathbb{I}[|\mathcal{F}| > 0]$
  if $\rho = 1$ **then**
    $t^f \leftarrow E_{\mathrm{field}}(f) + \phi_f(v^f)$ for each $f \in \mathcal{F}$ ▷ typed field tokens
    $T^m \leftarrow [t_{\mathrm{meta}}, t^{f_1}, \ldots, t^{f_{|\mathcal{F}|}}], \quad h^m \leftarrow [\mathrm{Enc}_m(T^m)]_{\mathrm{meta}}$
  end if
  for $\ell = 1$ **to** L_a **do**
    if $\rho = 0$ **then**
      $\Delta W_u^{(\ell)} \leftarrow \Delta W_{0,u}^{(\ell)}$ for $u \in \{q, v\}$ ▷ shared-only fallback
    else
      $\pi^{(\ell)} \leftarrow \mathrm{softmax}(R^{(\ell)}(h^m))$ ▷ metadata-conditioned routing
      $\Delta W_u^{(\ell)} \leftarrow \Delta W_{0,u}^{(\ell)} + \sum_{k=1}^{K} \pi_k^{(\ell)} \Delta W_{k,u}^{(\ell)}$ for $u \in \{q, v\}$ ▷ Eq. 3
    end if
  end for
  $a \leftarrow \mathrm{Enc}_v(x; \{\Delta W_u^{(\ell)}\})$
  $(g^{\mathrm{pre}}, g^{\mathrm{post}}, o) \leftarrow \mathrm{Head}(a)$
  return $g^{\mathrm{pre}}, g^{\mathrm{post}}, o$
end procedure

Training
for $s = 1$ **to** E **do**
  for each identity-sampled mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    for each $(x_i, y_i^c) \in \mathcal{B}$ **do**
      $\widetilde{\mathcal{F}}_i \leftarrow$ METADROP($\mathcal{F}_i; p_{\mathrm{field}}, p_{\mathrm{rec}}$) ▷ metadata dropout
      $(g_i^{\mathrm{pre}}, g_i^{\mathrm{post}}, o_i) \leftarrow$ FORWARD($x_i, m_i, \widetilde{\mathcal{F}}_i$)
    end for
    $\mathcal{L}_{\mathrm{ReID}} \leftarrow \mathcal{L}_{\mathrm{id}}(\{o_i, y_i^c\}) + \lambda_{\mathrm{tri}} \mathcal{L}_{\mathrm{tri}}(\{g_i^{\mathrm{pre}}, y_i^c\})$
    update θ using $\nabla_\theta \mathcal{L}_{\mathrm{ReID}}$
  end for
end for

Inference
for each $x \in \mathcal{X}_\mathcal{Q} \cup \mathcal{X}_\mathcal{G}$ with metadata m and fields $\mathcal{F}$ **do**
  $(_, g_x^{\mathrm{post}}, _) \leftarrow$ FORWARD($x, m, \mathcal{F}$); $\quad z_x \leftarrow g_x^{\mathrm{post}} / \|g_x^{\mathrm{post}}\|_2$
end for
$S_{k,j} \leftarrow z_{\alpha_k}^\top z_{\beta_j}$ for $1 \le k \le N_Q$, $1 \le j \le N_G$
return S

For MetaRouter parameters shared across species, including the number of routed experts K, we select a single configuration using the mean validation mAP across the eight species. The learning rate of each compared method is selected using the same validation protocol, while architecture-specific settings otherwise follow the corresponding published configuration. After model selection, the selected configuration is fixed, and each model is retrained using the complete training partition before evaluation on the held-out test set.

Evaluation metrics. We report mean Average Precision (mAP) and Rank-k retrieval accuracy. For each query, average precision measures retrieval precision across the ranked positions of the same-identity gallery samples, and mAP averages this quantity across all queries. Rank-k measures the proportion of queries for which at least one correct gallery match appears within the top k retrieved images. All reported results are averaged over five training seeds, with the corresponding across-

seed standard deviations for the main and controlled experiments, and Rank-3, Rank-5, and Rank-10 results, are reported in Appendix E.

C.2 Implementation Details

Backbone and adaptation. MetaRouter uses DINOv3 ViT-B/16 as the visual backbone, with embedding dimension 768 and 12 Transformer blocks, while keeping all pretrained backbone parameters frozen. Images are resized to 256×256. LoRA updates are inserted into the query and value projections of the last $L_a = 4$ Transformer blocks. Each adapted projection contains one shared LoRA update and K routed expert updates, with rank $r = 8$ and scaling factor $\eta = 16$. Following the default LoRA initialization, A is initialized with Kaiming uniform initialization and B with zeros, so that each LoRA update is zero at initialization. The output layer of each router is initialized to zero, so routing begins from a uniform expert distribution and is optimized jointly from the start of training. We select K from $\{1, 2, 4, 8\}$ using the mean validation mAP across the eight species and use the selected configuration for all species-specific experiments.

Metadata encoder. All metadata fields are mapped to a 128-dimensional token space. Categorical fields use field-specific value embeddings, continuous fields are standardized using statistics computed from the corresponding training data and projected with field-specific linear layers, and vector-valued fields are projected from their multi-hot representations. The metadata encoder consists of two Transformer layers with four attention heads, feed-forward dimension 256, and dropout 0.1. No positional embeddings are used because field identity is represented explicitly by the learned field embeddings.

Training and optimization. We use a BNNeck (Luo et al., 2019) ReID head, with the batch-hard triplet loss applied to the pre-BatchNorm feature and identity classification with label smoothing applied to the post-BatchNorm feature. MetaRouter is optimized using $\mathcal{L}_{\mathrm{ReID}} = \mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{tri}}\mathcal{L}_{\mathrm{tri}}$, with $\lambda_{\mathrm{tri}} = 1$, triplet margin 0.3, and label smoothing 0.1. Each mini-batch contains 8 identities with 4 images per identity. Models are trained for 60 epochs using AdamW with weight decay 5×10^{-4} and cosine learning-rate decay. For MetaRouter, the learning rate is selected from $\{5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}\}$ using the validation protocol described in Appendix C.1. We do not use random cropping, horizontal flipping, random erasing, color jitter, blur, or other stochastic image augmentations during training, since these operations can change the viewpoint, visible regions, occlusion, boundary contact, brightness, sharpness, or target scale described by the paired metadata without updating the corresponding annotations. The same augmentation policy is applied to all methods trained on MetaFauna, while each method retains its original input resolution and normalization. Metadata dropout is applied with $p_{\mathrm{field}} = 0.15$ for individual fields and $p_{\mathrm{rec}} = 0.10$ for complete metadata records. All experiments are conducted on NVIDIA RTX PRO 6000 GPUs.

Baseline implementations. We follow the published implementations and architecture-specific settings of CLIP-ReID (Li et al., 2023), UniReID (Jiao et al., 2023), AdaFreq (Li et al., 2024), AR-Base (Hou et al., 2025), and CARE (Wu et al., 2026). For each method, target-dataset hyperparameters such as the learning rate are selected using the same validation protocol and are shared across the eight species rather than tuned separately for individual species. For the metadata-aware baselines, MFA is added to CLIP-ReID and UniReID following MetaWild (Li et al., 2025), with the same MetaFauna metadata available to MetaRouter provided to the corresponding MFA variant. All methods are evaluated using the fixed train, query, gallery, and validation partitions described above.

D Additional Ablation and Analysis

D.1 Control Definitions

The controls in Table 2 isolate the effects of expert capacity, input-dependent routing, metadata input, and the correspondence between an image and its metadata. Unless stated otherwise, all variants use the same frozen visual backbone, adapted Transformer blocks, LoRA rank, ReID head and losses, optimization settings, data splits, and training seeds as MetaRouter.

Shared LoRA removes the routed expert bank, retaining only the shared low-rank update in each adapted query and value projection. It provides the metadata-free adaptation reference used to distinguish the effect of metadata-conditioned expert routing from standard LoRA adaptation.

Table D.1: **Additional metadata-group combinations.** We report the two pairwise combinations involving environmental context (G3) that are omitted from Table 3. Results are shown only for the six species for which G3 is available. Entries report mAP (%).

Metadata input	Cheetah	Elephant	Giraffe	Kudu	Lion	Zebra	Avg.
G1+G3	53.0	70.9	78.2	78.6	71.2	72.8	70.8
G2+G3	51.9	71.3	78.7	79.4	72.0	72.0	70.9

Static Expert Mix retains the shared update and the same K routed LoRA experts as MetaRouter, but replaces the metadata-dependent routing weights with a fixed uniform mixture, $\pi_{i,k}^{(\ell)} = 1/K$, for every image and adapted layer. All LoRA parameters remain trainable. This preserves the expert-bank capacity while removing input-dependent routing and metadata conditioning.

Feature Fusion tests whether the same metadata can improve retrieval through conventional feature-level integration rather than by conditioning visual adaptation. Its visual branch uses the Static Expert Mix, while its metadata branch uses the same field encoders and metadata encoder as MetaRouter. The resulting visual and metadata representations are concatenated and projected through a lightweight fusion MLP before the ReID head. When no metadata field is available, the metadata representation provided to the fusion module is set to zero. The same field-level and full-record metadata dropout used by MetaRouter is applied during training.

Visual Router retains the shared update, routed expert bank, adapted projections, and layer-specific routing mechanism of MetaRouter, but derives the routing weights from the visual representation rather than structured metadata. At adapted layer ℓ,

$$\boldsymbol{\pi}_i^{(\ell)} = \mathrm{softmax}\left(R_{\mathrm{vis}}^{(\ell)}\left(\mathbf{h}_{i,\mathrm{cls}}^{(\ell)}\right)\right), \tag{18}$$

where $\mathbf{h}_{i,\mathrm{cls}}^{(\ell)}$ is the visual class token entering the adapted layer. No metadata values or field-availability indicators are provided to the router. The resulting expert mixture is therefore image-dependent but not explicitly conditioned on structured metadata.

Shuffled MetaRouter retains the complete MetaRouter architecture but breaks the correspondence between each image and its metadata record. Whole metadata records are permuted within the same species, source project, and metadata availability pattern, excluding records with the same identity or encounter. Training, query, and gallery partitions are permuted independently so that metadata records do not cross data splits. Training permutations are resampled across epochs, whereas fixed permutations are used for evaluation. This preserves the marginal metadata distribution, correlations among fields within each donated record, and the original missingness patterns while removing the metadata associated with the target observation.

D.2 Metadata Group Analysis

The metadata-group ablation in Table 3 changes only the metadata provided to MetaRouter while keeping the images, train/query/gallery splits, visual backbone, expert-bank configuration, optimization settings, and training seeds fixed. The *None* configuration corresponds to Static Expert Mix: it retains the shared LoRA update and the same routed expert bank, but uses no metadata and a uniform expert mixture. For each group configuration, images retain their native metadata availability. An image is therefore not removed when one or more fields within a selected group are missing; MetaRouter operates on the subset of selected fields available for that observation. Environmental context (G3) is unavailable for Hyena and Leopard. For these two species, the complete-metadata configuration therefore reduces to G1+G2, and combinations involving G3 do not define distinct experimental settings. Table D.1 reports the two additional pairwise combinations involving G3 for the six species in which environmental metadata is available.

Both G1+G3 and G2+G3 improve over their corresponding single-group configurations, but remain below G1+G2 on the G3-available species. The relative ordering between the two combinations follows the species-specific pattern in Table 3: G1+G3 is stronger when visibility metadata contributes more substantially, whereas G2+G3 is stronger for species that benefit more from capture conditions. These results reinforce the main group ablation: metadata describing which identity

evidence is visible and how it is captured accounts for most of the gain, while environmental context contributes a smaller additional signal.

D.3 Field-level Metadata Analysis

The group-level results in Table 3 show that visual evidence visibility (G1) and capture conditions (G2) contribute most of the improvement, with a smaller contribution from environmental context (G3). To examine how these group-level effects are distributed across individual metadata fields, we ask two questions: whether a given field is useful on its own, and whether it still contributes once the remaining metadata fields are available. We therefore use two retraining settings, field-only and leave-one-field-out. All variants use the same images, train/query/gallery splits, visual backbone, expert-bank configuration, optimization settings, and training seeds. Only the metadata fields provided to MetaRouter are changed. Images are not removed when a field is missing.

Field-only. For each metadata field f, we train MetaRouter using f as its only metadata input and remove all other fields from both training and evaluation records. Let M_f denote this model and $M_\emptyset$ the metadata-free reference, which corresponds to the *None* configuration in Table 3: the same shared and expert LoRA updates are retained, but no metadata is provided, and the experts are combined uniformly. We report

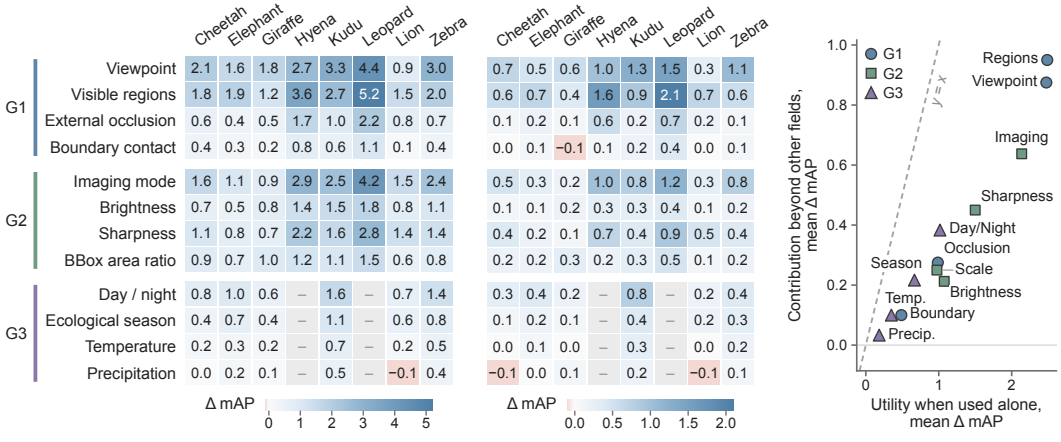
$$\Delta_f^{\text{only}} = \text{mAP}(M_f) - \text{mAP}(M_\emptyset). \tag{19}$$

A positive value indicates that the field is useful when provided on its own.

Leave-one-field-out. We next retrain MetaRouter after removing one field f from the complete metadata record while keeping all remaining fields unchanged. Let M_{All} denote the complete MetaRouter and M_{-f} the model trained without field f. We report

$$\Delta_f^{\text{LOO}} = \text{mAP}(M_{\text{All}}) - \text{mAP}(M_{-f}). \tag{20}$$

A positive value indicates that removing f reduces retrieval performance when the other metadata fields remain available. A small leave-one-field-out effect does not imply that the field is uninformative, since related fields may provide similar information, and field availability differs across images.



(a) Field-only utility. (b) Leave-one-field-out. (c) Mean comparison.

Figure D.1: **Field-level metadata analysis.** (a) Field-only effect Δ_f^{only}, measuring the change in mAP when field f is used as the only metadata input. (b) Leave-one-field-out effect Δ_f^{LOO}, measuring the decrease in mAP when field f is removed and the remaining metadata is retained. (c) Mean field-only and leave-one-field-out effects over the species for which each field is available. Results are averaged over five training seeds. Points below $y = x$ indicate that removing a field from the complete metadata produces a smaller effect than using that field alone, consistent with partial overlap with the remaining metadata. Dashes denote unavailable field–species combinations.

Figure D.1 shows clear differences across metadata fields and species. Within G1, viewpoint and visible regions produce the strongest and most consistent effects, while external occlusion and boundary contact provide smaller additional signals. Within G2, imaging mode and sharpness show the clearest effects, whereas brightness and target scale generally have smaller and more species-dependent effects. For the six species with G3 annotations, individual environmental fields have weaker effects overall; solar period and ecological season contribute more consistently than air temperature and antecedent precipitation. This pattern is consistent with the group-level analysis, in which metadata describing the visibility and capture of identity evidence provides larger gains than environmental context.

The leave-one-field-out effects are generally smaller than the corresponding field-only effects in Figure D.1c. This indicates that a field can be useful when provided alone, but becomes less critical when other fields describing related observational conditions remain available. The field-level results, therefore, suggest that the gains from G1 and G2 are distributed across multiple partially overlapping descriptors rather than depending on a single metadata field.

D.4 Metadata Availability and Robustness

The main paper evaluates how MetaRouter behaves as observed metadata fields are progressively removed at inference. Here, we provide the full evaluation protocol, separate the effects of field-level and full-record metadata dropout, and further evaluate cases in which a metadata field or source is unavailable across the test set.

Random field missingness. For each image, every originally observed metadata field is independently removed with probability $p \in \{0, 0.25, 0.50, 0.75, 1\}$. Fields that are already missing in MetaFauna remain missing. The setting $p = 0$ preserves the original metadata availability, while $p = 1$ removes all observed metadata. Query and gallery records are masked independently. For each $p \in \{0.25, 0.50, 0.75\}$, we sample 10 independent masking realizations, and the same image-level masks are used across all model variants and training seeds. The settings $p = 0$ and $p = 1$ are deterministic. Results are first averaged over the masking realizations for each training seed and then averaged over the five training seeds.

Effect of metadata dropout. To examine the roles of the two training-time dropout operations, we train four otherwise identical MetaRouter variants: no metadata dropout ($p_{\text{field}} = 0, p_{\text{rec}} = 0$), field-level dropout only ($p_{\text{field}} = 0.15, p_{\text{rec}} = 0$), full-record dropout only ($p_{\text{field}} = 0, p_{\text{rec}} = 0.10$), and the combined setting used by MetaRouter ($p_{\text{field}} = 0.15, p_{\text{rec}} = 0.10$). Field-level dropout exposes the model to different subsets of the available fields during training. Full-record dropout removes the complete metadata record and therefore trains the shared-only path used when metadata is unavailable. If field-level dropout happens to remove every available field from an image, the shared-only path is also used for that sample. All four variants are evaluated using the same test-time masks described above.

Figure D.2 shows little difference among the four training settings when the original metadata is available at $p = 0$. As metadata becomes partially unavailable, field-level dropout gives larger improvements than full-record dropout at intermediate masking rates for most species. This follows the training setting, since field-level dropout exposes the model to partial metadata records. At $p = 1$, full-record dropout becomes more important because retrieval relies entirely on the shared-only path. The record-only variant therefore matches or slightly exceeds the combined setting for several species when all metadata is removed. Using both forms of dropout maintains performance with complete metadata and provides stable performance across both partial and complete metadata loss.

Systematic metadata unavailability. Random masking removes different fields from different images, but some metadata failures affect the same field or source across an entire test set. We first evaluate single-field removal by deleting field f from every query and gallery record in which it is originally available while keeping the trained MetaRouter checkpoint and all remaining metadata unchanged. For species s, we report

$$\Delta_{f,s}^{\text{sys}} = \text{mAP}_s^{\text{native}} - \text{mAP}_{s,-f}, \tag{21}$$

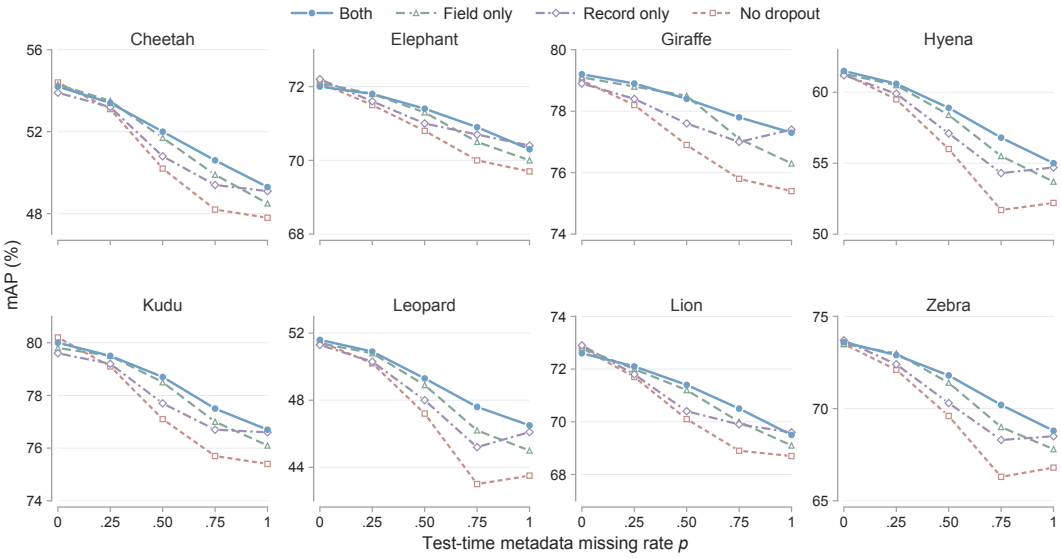


Figure D.2: **Effect of training-time metadata dropout under incomplete test-time metadata.** We compare MetaRouter trained with both field-level and full-record dropout (*Both*), field-level dropout only (*Field only*), full-record dropout only (*Record only*), and no metadata dropout (*No dropout*). Each originally observed metadata field is independently removed at inference with probability p, and the curves report mAP (%) for each species. Results are averaged over five training seeds and, for $0 < p < 1$, 10 test-time masking realizations. $p = 0$ preserves the original metadata availability, while $p = 1$ removes all observed metadata. Y-axis ranges are adjusted by species for readability.

where $\mathrm{mAP}_s^{\mathrm{native}}$ denotes performance with the original metadata and $\mathrm{mAP}_{s,-f}$ denotes performance after removing field f at inference. No retraining is performed, and means are computed over the species for which the corresponding field is available.

We also evaluate three cases in which several fields become unavailable because they depend on the same metadata source. *BBox unavailable* removes boundary contact and target scale while retaining the ReID image; this setting represents cases in which source-frame bounding-box coordinates are unavailable to the metadata construction pipeline. *ERA5-Land unavailable* removes air temperature and antecedent precipitation while retaining capture time and location, representing failure to obtain or match the external weather data. *Location unavailable* removes the location-dependent environmental metadata together with the associated capture context used in our G3 construction. Hyena and Leopard are excluded from the two G3-related settings as G3 metadata is not available for them.

Figure D.3(a) shows that sensitivity to systematic field removal varies across both fields and species. Among G1 fields, viewpoint and visible regions produce the largest mean drops, with stronger effects on Cheetah, Hyena, Leopard, and Zebra. Removing visible regions decreases mAP by 2.1 points on Hyena and 2.3 points on Leopard, while viewpoint removal decreases mAP by 1.4–1.8 points on Cheetah, Hyena, Leopard, and Zebra. External occlusion has a smaller effect, and boundary contact has little effect for most species. Within G2, imaging mode and sharpness produce the largest drops, particularly for Giraffe, Kudu, and Lion. The G3 fields have smaller effects overall, with solar period showing the largest mean drop among the four environmental fields. These results follow the species-level differences observed in the group analysis while showing that the sensitivity is not uniform within each metadata group.

Figure D.3(b) further shows that the effect of missing metadata depends on which upstream information is unavailable. Removing the two bbox-derived fields decreases mAP by 0.64 points on average, while removing the two ERA5-Land weather fields causes only a 0.18-point decrease over the six applicable species. When the location-related context is unavailable, the mean decrease is 0.70 mAP. The number of missing fields alone therefore does not explain the performance drop. For example, the two bbox-derived fields produce a larger drop than the two weather fields, while losing four location-dependent fields produces a similar average drop to losing the bbox-derived pair.

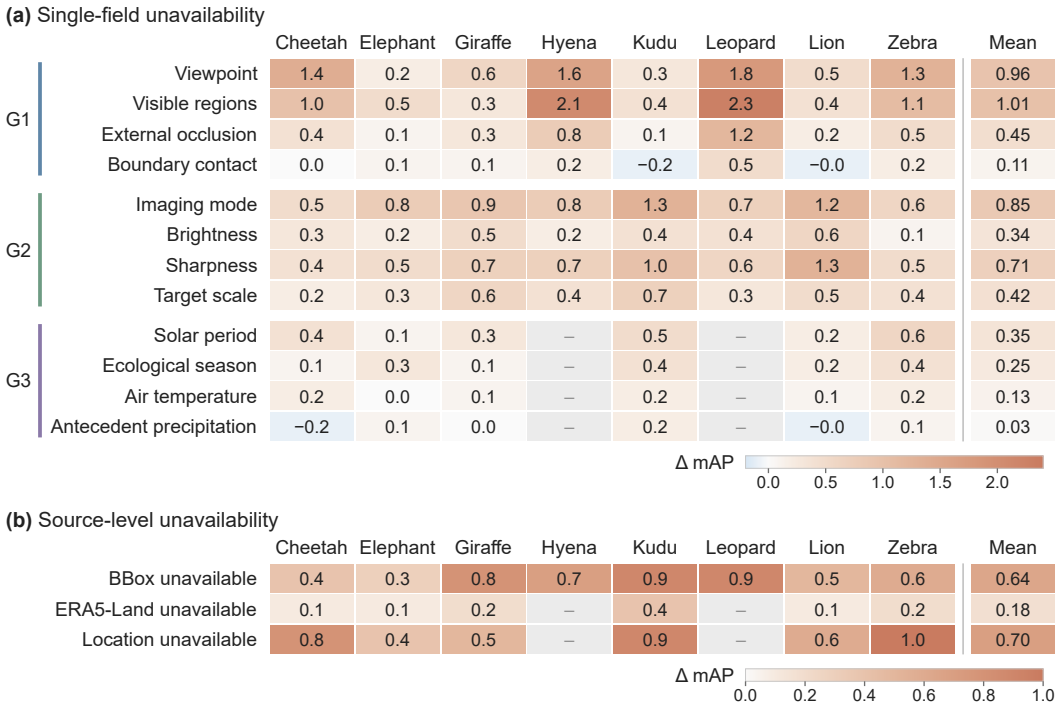


Figure D.3: **Robustness to systematic metadata unavailability.** (a) mAP decrease when one metadata field is removed from every query and gallery record in which it is originally available. (b) mAP decrease when multiple fields that depend on the same upstream metadata or context become unavailable together. The trained MetaRouter checkpoint is fixed and no retraining is performed. Positive values indicate performance degradation after metadata removal. Means are computed over species for which the corresponding metadata is available, and dashes denote unavailable field–species combinations.

Table E.1: **Across-seed results for the controlled attribution study.** Entries report mean mAP ± standard deviation (%) over five training seeds. The best result for each species is shown in bold.

Variant	Cheetah	Elephant	Giraffe	Hyena	Kudu	Leopard	Lion	Zebra
Shared LoRA	49.1 ± .5	70.5 ± .4	77.1 ± .4	55.3 ± .6	76.9 ± .5	46.4 ± .7	69.9 ± .5	69.2 ± .6
Static Expert Mix	49.8 ± .5	70.4 ± .3	77.0 ± .4	55.9 ± .5	77.5 ± .4	47.2 ± .7	69.7 ± .4	69.8 ± .6
Feature Fusion	51.7 ± .4	71.1 ± .4	78.1 ± .3	58.9 ± .5	78.3 ± .4	49.3 ± .6	71.0 ± .4	71.8 ± .5
Visual Router	50.7 ± .5	71.0 ± .4	78.4 ± .4	58.0 ± .6	78.8 ± .4	48.5 ± .6	71.4 ± .5	72.0 ± .5
Shuffled MetaRouter	49.3 ± .6	70.8 ± .4	76.6 ± .5	54.8 ± .6	77.9 ± .5	46.5 ± .8	70.2 ± .5	69.3 ± .6
MetaRouter	**54.2** ± .4	**72.0** ± .3	**79.2** ± .3	**61.5** ± .4	**80.0** ± .3	**51.6** ± .5	**72.6** ± .4	**73.6** ± .4

Together with the random-missingness results, these experiments show that MetaRouter remains robust to incomplete metadata, although the effect depends on which information is unavailable.

E Additional Results

E.1 Across-Seed Stability

Main comparison. Figure E.1 reports the across-seed variation corresponding to Table 1. MetaRouter retains the highest mean mAP and Rank-1 on all eight species, while its variation across the five training seeds remains comparable to that of the other methods. The larger improvements on Cheetah, Hyena, and Leopard therefore persist across training seeds and are not specific to a single training run.

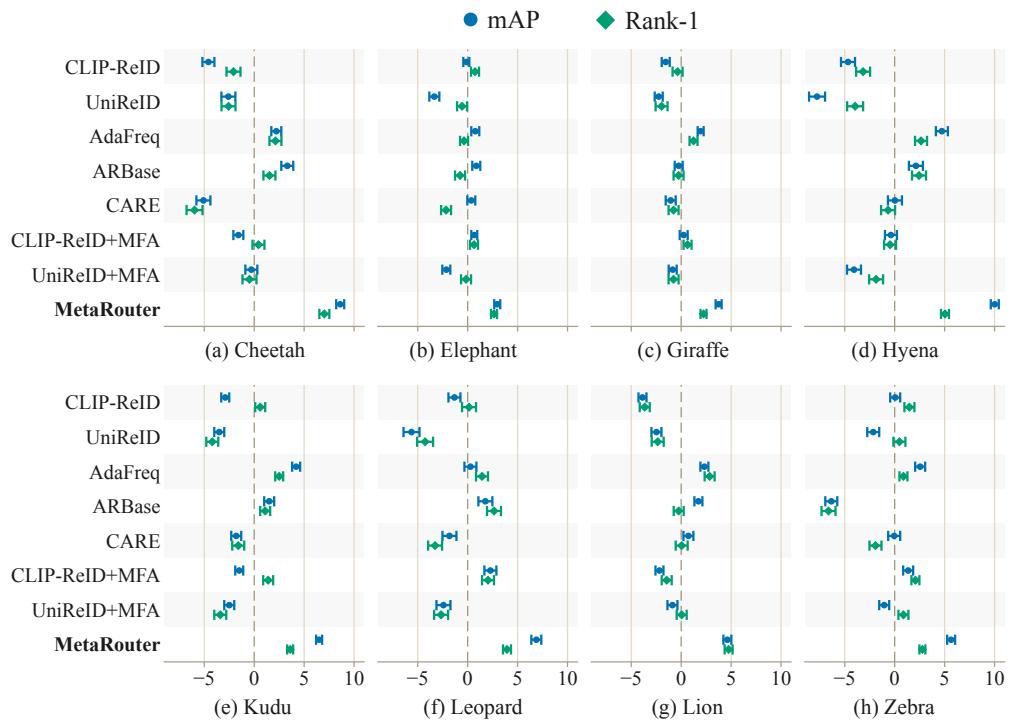


Figure E.1: **Across-seed stability of the main results.** For each species, mAP and Rank-1 are centered separately by the mean score across the eight compared methods, such that zero denotes the species-specific mean across methods and positive values indicate higher performance. Blue circles denote mAP and green diamonds denote Rank-1; horizontal error bars show one standard deviation over five training seeds.

Controlled attribution. The main attribution study in Table 2 reports mean mAP for clarity. Table E.1 provides the corresponding mean and standard deviation over five training seeds for all controlled variants. Static Expert Mix remains close to Shared LoRA, and Shuffled MetaRouter remains within 1.1 mAP of Static Expert Mix on every species. In contrast, MetaRouter improves over Static Expert Mix by 1.6–5.6 mAP and over Feature Fusion by 0.9–2.6 mAP across the eight species. The same pattern is observed across training seeds, supporting the conclusion that the improvement depends on the metadata paired with each image and its use for routing, and cannot be explained by the additional expert capacity alone.

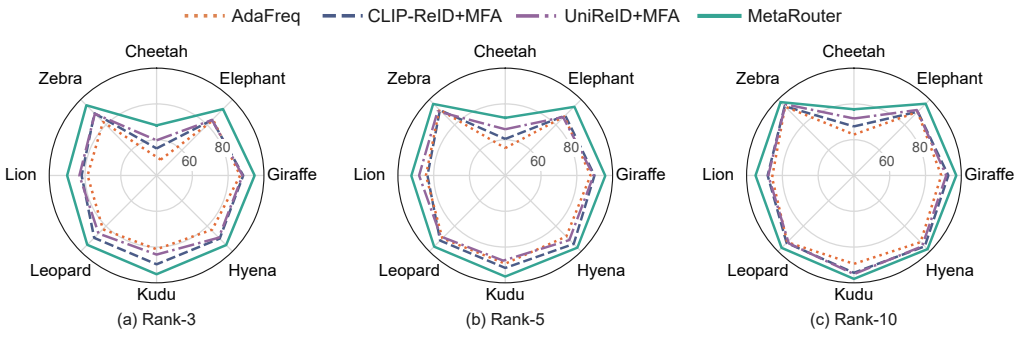


Figure E.2: Rank-k retrieval accuracy ($k = 3, 5, 10$) across the eight MetaFauna species. MetaRouter consistently achieves the highest accuracy across retrieval depths.

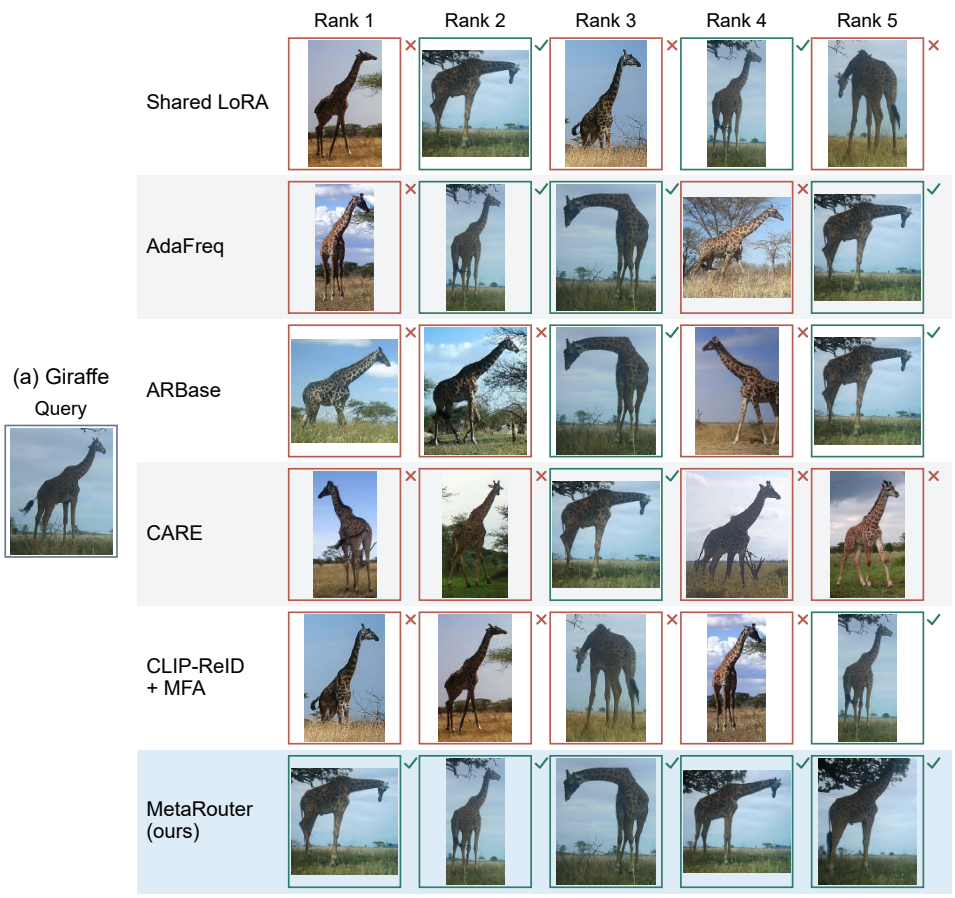


Figure E.3: **Top-5 retrieval examples for Giraffe and Lion.** Each panel shows one query image on the left and the top-5 gallery images returned by each method in rank order. Correct matches correspond to the query identity, while the remaining images are same-species distractors.

E.2 Rank-k Retrieval Accuracy

Figure E.2 extends the Rank-1 comparison in Table 1 to $k \in \{3, 5, 10\}$. MetaRouter maintains the strongest retrieval performance across different retrieval depths. The differences between methods become smaller at larger k for several species as more models recover a correct identity within a deeper candidate set, while the advantage of MetaRouter remains visible beyond Rank-1. The improvement is therefore not limited to the first retrieved image and is also reflected in the broader ranked list.

E.3 Qualitative Retrieval Examples

Figure E.3 shows representative Top-5 retrieval results on Giraffe and Lion. MetaRouter retrieves the query identity at all five ranks in both examples, while the baselines include same-species distractors, often at higher ranks. Quantitative comparisons are reported in Table 1 and Figure E.2.

E.4 Multi-Species Training

The main experiments train a separate model for each species. We further examine whether one retrieval model can be trained jointly across all eight species, and whether metadata-conditioned routing remains useful when the adaptation parameters are shared across species.

Multi-species training protocol. We pool the training partitions of all eight species and share the visual backbone, LoRA parameters, metadata encoder, routers, BNNeck, and retrieval embed-



Figure E.3: Top-5 retrieval examples for Giraffe and Lion (continued).

ding. Because identity labels are defined independently within each species, a separate identity-classification layer is used for each species during training. These classification layers provide identity supervision only and are discarded at inference, so all parameters used to produce retrieval features are shared across species. Each mini-batch contains samples from one species, with the species sampled uniformly and the batch formed by 8 identities with 4 images per identity. This keeps batch-hard triplet mining within a species and avoids using cross-species differences as trivial negative evidence. The shared models are trained for the same total number of optimization steps as the sum of the eight corresponding species-specific runs. The learning rate and checkpoint are selected using mean validation mAP across the eight species, while the MetaRouter architecture and K are kept unchanged from the species-specific setting. At inference, the shared model is evaluated separately on the original query and gallery sets of each species.

Comparisons. To evaluate the shared-model setting against established ReID approaches, we jointly train the same visual-only and metadata-aware ReID methods evaluated in Table 1 using the species-balanced sampling protocol above. To examine the effect of metadata-conditioned routing in the shared-model setting, we also train Shared LoRA, Static Expert Mix, Visual Router, and Shuffled MetaRouter using the control definitions in Appendix D. The independently trained MetaRouter models from Table 1 are included as a reference to quantify the performance change when the eight species-specific models are replaced by one shared model.

Results. Table E.2 shows that a single multi-species MetaRouter retains most of the performance of the independently trained models. Cheetah and Hyena improve slightly under joint training, while the other six species decrease by 0.6–2.8 mAP and 0.6–1.8 Rank-1. We find that the effect of multi-species training depends on the amount of species-specific training data. For Cheetah and Hyena, which have among the fewest training images in MetaFauna, the shared model slightly improves over independent training, whereas larger subsets such as Giraffe, Leopard, and Zebra show modest decreases. This suggests that additional cross-species supervision can be beneficial when species-specific data are more limited, while its benefit diminishes as sufficient within-species supervision becomes available.

The relatively small overall gap is also consistent with the use of a frozen pretrained backbone. DINOv3 already provides a strong transferable visual representation, so the multi-species experiment can share only the lightweight adaptation and metadata-conditioning modules rather than relearning general visual features across species. Despite this shared adaptation capacity, MetaRouter outperforms the jointly trained external baselines and improves upon Visual Router and Static Expert Mix across all species. Shuffling the metadata substantially reduces this advantage, indicating that the

Table E.2: **Multi-species training on MetaFauna.** All models in the first two groups use one shared retrieval model across the eight species and are evaluated separately on the original species-specific query and gallery sets. The final row reports the independently trained MetaRouter models from Table 1. Each entry reports mAP / Rank-1 (%). Bold values indicate the best result among models trained jointly across species. MetaR. denotes MetaRouter.

Method	Cheetah	Elephant	Giraffe	Hyena	Kudu	Leopard	Lion	Zebra
Existing methods trained jointly across all species								
CLIP-ReID	41.8/44.8	68.4/80.8	73.5/83.3	47.6/75.2	70.9/81.8	42.7/79.5	64.4/71.1	67.2/83.6
UniReID	44.6/45.0	66.5/80.4	74.0/82.9	45.5/75.0	71.2/78.0	39.8/76.4	66.5/73.0	66.2/83.5
AdaFreq	48.4/48.8	69.4/79.8	77.0/85.0	56.8/80.6	77.8/83.6	44.5/81.0	70.4/77.5	69.8/83.1
ARBase	49.5/48.1	70.2/79.9	75.5/84.1	54.6/80.6	75.5/82.4	45.9/82.2	70.1/74.7	62.0/76.4
CARE	41.2/40.8	69.2/78.0	74.5/83.5	52.3/77.5	72.1/79.7	42.4/76.3	68.9/74.8	67.4/80.3
CLIP-ReID$_{+\mathrm{MFA}}$	44.3/47.0	67.7/80.4	75.4/84.8	50.5/77.0	72.6/83.0	45.0/81.1	66.8/73.5	68.8/**84.4**
UniReID$_{+\mathrm{MFA}}$	46.3/47.0	67.6/81.0	75.5/84.0	48.0/76.5	72.7/79.3	41.8/78.0	68.5/75.3	67.5/84.2
MetaRouter and controls trained jointly across all species								
Shared LoRA	49.8/49.0	70.4/77.9	76.6/82.2	56.1/77.2	77.3/78.9	45.9/76.4	70.2/73.6	68.3/80.5
Static Expert Mix	50.6/49.5	70.2/78.5	76.5/82.7	56.6/78.0	77.6/79.7	46.6/77.2	69.8/74.2	69.1/80.9
Visual Router	51.3/51.7	**71.4**/80.0	77.6/84.3	58.7/80.4	78.5/82.3	47.5/81.1	71.5/76.6	**70.8**/83.0
Shuffled MetaR.	49.9/49.3	70.5/78.3	76.0/82.6	55.4/77.6	77.8/79.4	45.6/76.9	70.2/74.0	68.5/80.7
MetaRouter	**54.5/53.2**	70.7/**81.8**	**77.8/85.2**	**61.7/82.9**	**79.1/83.7**	**49.7/82.4**	**72.0/78.6**	**70.8**/83.7
Species-specific MetaRouter								
MetaRouter	54.2/53.0	72.0/83.0	79.2/86.4	61.5/82.6	80.0/84.5	51.6/84.0	72.6/79.2	73.6/85.5

benefit under multi-species training still depends on the correspondence between each image and its metadata rather than on shared expert capacity alone.