Region-Aware Multimodal Interleaving for Animal Re-Identification

Yihao Wu[1], Di Zhao[1], Wayne Marcus Getz[2], Lingqiao Liu[3], Gillian Dobbie[1], Daniel Wilson[1], and Yun Sing Koh[1]

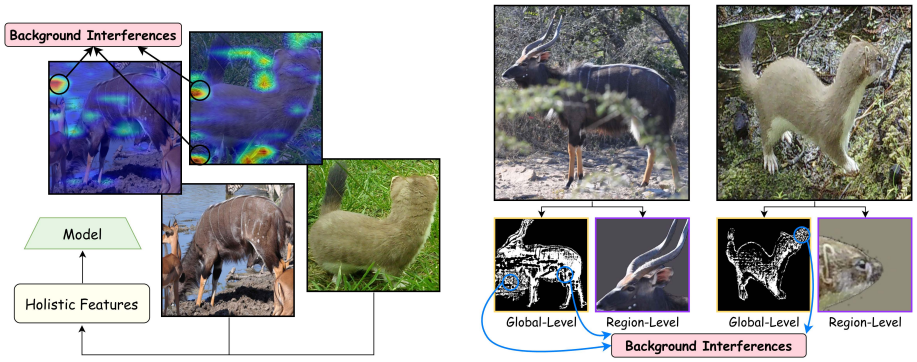
[1] School of Computer Science, University of Auckland, New Zealand
{yihao.wu,di.zhao,g.dobbie,daniel.wilson,y.koh}@auckland.ac.nz
[2] Dept. ESPM, University of California, Berkeley, CA 94720, USA
wgetz@berkeley.edu
[3] School of Computer Science & Information Technology, Adelaide University, Australia
lingqiao.liu@adelaide.edu.au

Abstract. Animal Re-Identification (Animal ReID) is crucial for ecological monitoring in the wild, yet remains challenging due to the fine-grained and diverse visual features of animals, including intricate textures, distinctive patterns, and various pose or form representations. While current Animal ReID methods mostly rely on holistic global features, they frequently overfit to the surrounding environment rather than capturing localised, identity-relevant cues. Existing pattern-based methods primarily focus on global generalised signals that are sensitive to background interventions and often fail to highlight identity evidence for individual discrimination. Additionally, these methods are mainly constrained to the visual space only and lack pattern-based semantic interactions from the multimodal space. To address these challenges, we introduce a **R**egion-**A**ware **M**ultimodal **I**nterleaving (RAMI) framework, which formulates Animal ReID as a visual-textual semantic interaction over informative, localised regions. Specifically, RAMI projects segmentation-derived region tokens into a shared token space and interleaves them with semantic tokens within a Transformer-based module to facilitate dense cross-modal interactions. Given the lack of region-level data, we design a simple pipeline to produce biologically representative regions from segmented animals. To the best of our knowledge, we are the first to formulate multimodal interleaving tailored for fine-grained discriminative tasks in Vision-Language Models (VLMs). Experiments show that RAMI outperforms state-of-the-art methods across both benchmark and in-the-wild datasets, with an average mAP gain of up to 18.7%. Our code is available at https://github.com/ML-4-SocialGood/RAMI.git.

Keywords: animal re-identification · multimodal interleaving · region modelling · computer vision for social good

1 Introduction

Animal Re-Identification (Animal ReID) aims to identify individual animals across multiple scenes over time, playing a critical role in ecological monitor-



(a) Challenges in current Animal ReID methods, *e.g.*, CARE [52], that leverage holistic visual features to capture identity-relevant cues. We employ Grad-CAM [44] to visualise the attention driven from the model when identifying individual animals. The focus on the surrounding environment, such as grass and background elements, introduces spurious features, limiting the model's generalisability.

(b) Challenges in existing pattern-based methods for Animal ReID. The global pattern extraction methods, *e.g.*, GloPER [6], are often sensitive to background interventions and occlusion, which can unintentionally incorporate spurious features into models. Conversely, the decomposed region-level cues potentially provide fine-grained, identity-relevant features for characterising individuals.

Fig. 1: Illustration of challenges in existing Animal ReID methods.

ing, specifically enabling population estimation [43], movement analysis [27], and long-term behavioural studies [41] from camera traps. Despite some progressive efforts having been made in Animal ReID, it remains an open challenge stemming from diverse and fine-grained visual representations of animals, including extreme pose and form variations, intricate markings, frequent occlusions, and significant domain shifts across cameras [15,22,39,57–59]. Existing computer vision methods [10,36,37,42] primarily leverage holistic visual features to capture identity-relevant cues, which are often susceptible to background interference and can introduce suspicious correlations into models. Figure 1a shows the challenges of relying on holistic visual concepts in current methods. Specifically, the Grad-CAM [44] visualisations indicate that the models mainly highlight surrounding elements within the holistic features rather than concentrating on biologically representative animal regions when identifying individuals. Moreover, focusing solely on these holistic traits often restricts the model's capability to recognise the same individual under high visual variation, *e.g.*, extreme pose changes or occlusions [14,52].

While pattern-based methods offer a potential orientation in Animal ReID, they mainly rely on global signals that capture generalised animal features rather than fine-grained identity evidence [4,48]. Additionally, the global pattern extraction of these methods is particularly sensitive to background interference and occlusion, inadvertently leveraging noisy information to characterise identities. In contrast, decomposing the whole image into localised regions with skin markings, complex textures, or distinctive body parts often provides identity-discriminative cues for distinguishing individuals [31,45,54]. Figure 1b illustrates

the challenges in existing pattern-based methods that leverage global, generalised patterns for Animal ReID, as well as examples of localised regions. Furthermore, current pattern-based methods [6, 7, 17, 28] are mostly limited to a unimodal space, lacking semantic context and interactions across patterns.

Recently, exploiting the potential of language guidance in Animal ReID through pretrained Vision-Language Models (VLMs), *e.g.*, CLIP [40], to inject semantic context into visual concept learning has shown promise [5, 16, 22, 52]. However, effectively incorporating semantic context from language to improve the ability to capture identity-relevant cues remains a key bottleneck in current methods. Specifically, some identity evidence for animals, such as intricate textures, complex markings, intrinsic pattern colours, or distinctive body parts, is excessively difficult to completely convey in plain language. In addition, existing methods that rely on post hoc feature fusion or simple cross-modal alignment to enhance visual cues lack dense cross-modal interactions for Animal ReID.

To address these limitations, we introduce a novel **R**egion-**A**ware **M**ultimodal **I**nterleaving (RAMI) framework that formulates Animal ReID as dense cross-modal interactions over potentially identity-discriminative regions. RAMI first grounds segmentation-derived, fine-grained regions with semantic context in region-referenced interleaved tokens and then structures the multimodal interleaving via a shallow Transformer-based Interleaver module, enforcing dense cross-modal interactions via the attention mechanism. While multimodal interleaving has made considerable strides in Large Multimodal Models (LMMs) tailored for generation or reasoning tasks in multi-image scenarios [18, 53], its potential for tackling fine-grained discriminative tasks, *e.g.*, Animal ReID, in VLMs is still underexplored. To bridge this gap, we leverage a SigLIP-inspired objective [55] to structure dense cross-modal interactions between semantic context and localised regions, thereby enhancing guidance towards identity-relevant visual cues for discriminating individual animals.

The core contributions of this work are summarised as: (1) We introduce a simple pipeline to autonomously generate segmentation-derived region-level data that prioritises the biologically representative regions with potentially identity-discriminative features for Animal ReID. (2) We formulate region-aware Animal ReID in a multimodal space, enriching semantic context and region-level interactions. (3) To the best of our knowledge, we are the first to integrate multimodal interleaving tailored for fine-grained discriminative tasks into pretrained foundation models. Our proposed RAMI framework enhances cross-modal interactions across regions, offering valuable insights into the challenges in Animal ReID. Extensive experiments on benchmark and in-the-wild datasets validate RAMI's effectiveness and its potential for ecological monitoring.

2 Related Work

Computer Vision for Animal ReID. Harnessing the power of computer vision has recently advanced Animal ReID for ecological monitoring, yet it is still struggling with the diverse visual representations of animals in the wild. Early

efforts primarily leverage CNN- and ViT-based approaches to re-identify zebras [47], African elephants [50], yaks [56], pandas [48], and dairy cows [2]. Existing pattern-based methods, such as ALFRE-ID [29] and GloPER [6], adopt local visual feature aggregation and global unsupervised segmentation to extract identity evidence, lacking semantic context across animal patterns. Currently, more attention has been shifted to VLMs, which exploit language guidance [34, 35, 38] to enhance visual representation learning for characterising identities. MetaWild [22] incorporates environmental metadata into CLIP [40] to supplement the semantic context for Animal ReID. UniReID [16] relies on separate visual and textual guidance to capture discriminative features for individual identification, while CARE [52] focuses on image-conditioned prompt learning and semantic feature alignment to explicitly address the inherent visual challenges in Animal ReID. However, these methods structure visual and textual features separately and rely solely on a simple cross-modal alignment to highlight identity-relevant cues, without dense cross-modal interactions, limiting their ability to provide enriched semantic guidance towards identity evidence.

Multimodal Interleave. The multimodal interleaving paradigm has been recently adopted in several LMMs [23, 33, 46, 53]. LLaVA-Interleave [18] is capable of various real-world multi-image tasks under diverse scenarios, *e.g.*, video and 3D environment. Nevertheless, enabling multimodal interleaving for fine-grained discriminative representation learning within pretrained foundation models remains underexplored.

3 Region-Aware Multimodal Interleaving Framework

This section begins with a formal definition of Animal ReID (Section 3.1). We then present the proposed **R**egion-**A**ware **M**ultimodal **I**nterleaving (RAMI) framework that structures dense cross-modal interactions at the region level for Animal ReID, comprising region-aware visual tokenisation and region-referenced interleaved token grounding (Section 3.2) and multimodal interleaving within a shallow Transformer-based Interleaver module (Section 3.3). We finally summarise the overall training objective of RAMI and the inference protocol in Section 3.4. Figure 2 shows an overview of the proposed RAMI framework.

3.1 Preliminaries

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, $y_i^c \in \mathbb{R}$ is the corresponding identity label with the 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, $y_j^c \in \mathbb{R}$ and $y_k^c \in \mathbb{R}$ are the corresponding identity labels with the class c, and N_G and N_Q denote the sample size in the Gallery and Query Sets. At inference time, the identities in $\mathcal{Q}$ and $\mathcal{G}$ are unseen during training, *i.e.*, $\mathcal{Y}_\mathcal{D} \cap \mathcal{Y}_\mathcal{G} = \emptyset$, $\mathcal{Y}_\mathcal{D} \cap \mathcal{Y}_\mathcal{Q} = \emptyset$, while $\mathcal{Y}_\mathcal{G} \cap \mathcal{Y}_\mathcal{Q} \neq \emptyset$.

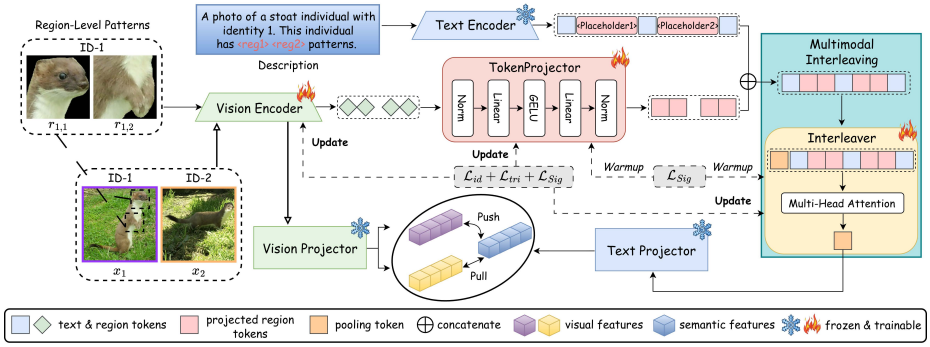


Fig. 2: Overview of the proposed RAMI framework. Given an input image sample, the corresponding localised regions are encoded into region tokens through SigLIP's vision encoder, then projected to a shared token space via a lightweight TokenProjector. Meanwhile, a textual description corresponding to the given input is encoded into text tokens, interleaved with placeholders, through SigLIP's text encoder. The placeholders are the positions where the projected region tokens interleave. The localised visual cues are finally grounded in region-referenced interleaved tokens through concatenation. To enable multimodal interleaving, we first prepend a learnable pooling token to the interleaved token sequence. A shallow Transformer-based Interleaver module then structures dense cross-modal interactions over regions. SigLIP's vision and text projectors project visual tokens of the given input sample (x_i) and a pooling token from cross-modal interactions into a shared embedding space, aligning paired visual and semantic features for Animal ReID.

The goal is to learn a matching function $f : \mathcal{Q} \times \mathcal{G} \mapsto \mathbb{R}^{N_Q \times N_G}$ that assigns a similarity score to each pair (α_k, β_j), where $1 \leq k \leq N_Q$ and $1 \leq j \leq N_G$, such that

$$f(\alpha_q, \beta_p) > f(\alpha_q, \beta_r), \text{ if } y_q^c = y_p^c \text{ and } y_q^c \neq y_r^c, \tag{1}$$

where α_q is a query sample with the identity y_q^c, β_p and β_r are two distinct gallery samples with the identities y_p^c and y_r^c, respectively, $1 \leq q \leq N_Q$, and $1 \leq p, r \leq N_G$.

3.2 Region-Aware Visual Tokenisation and Region-Referenced Interleaved Token Grounding

While global pattern representations can capture generalised features of animals, they frequently fail to highlight identity-relevant cues and are susceptible to background interference, occlusions, and pose variations. In contrast, localised, fine-grained visual cues, such as coat patterns, intricate textures, unique markings, or distinctive body parts, often serve as discriminative features that characterise animal identities. To produce segmentation-derived region-level data, we design a simple pipeline that autonomously generates biologically informative regions from each image sample. The details of our proposed region mining pipeline are discussed in Section 4.1.

Region-Aware Visual Tokenisation. Given an image sample x_i, we define a set of biologically meaningful regions $\mathcal{R}_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,N}\}$ obtained by our proposed pipeline. Each region corresponds to a localised area that may contain identity-relevant appearance. The segmentation-derived regions are initially encoded into region tokens $\mathbf{r}_i \in \mathbb{R}^{N \times pats \times d_v}$ via SigLIP's vision encoder $\mathcal{I}(\cdot)$, where N denotes the number of regions associated with the sample x_i, *pats* represents the number of patches per region, and d_v is the hidden dimension of visual tokens. To enable cross-modal interleaving, the region tokens are then projected into a shared token space through a lightweight TokenProjector $\mathcal{P}(\cdot)$:

$$\mathbf{z}_i = \mathcal{P}\left(\mathbf{r}_i\right) \in \mathbb{R}^{N \times pats \times d_t}, \tag{2}$$

where d_t represents the hidden dimension of text tokens.

Region-Referenced Interleaved Token Grounding. To provide a semantic grounding over regions, each image sample is associated with a region-referenced textual description, interleaved with placeholders for regions. Specifically, the region-referenced textual prompts are formatted with a generic template, *e.g.*, "A photo of a $\{species\}$ individual with identity $\{id\}$. This individual has $\langle reg1\rangle \ldots \langle regN\rangle$ patterns.". The $\{species\}$ and $\{id\}$ are the corresponding species class (*e.g.*, stoat) and identity label of the image sample, whereas $\langle reg1\rangle \ldots \langle regN\rangle$ represent placeholders for each localised region.

SigLIP's text encoder $\mathcal{T}(\cdot)$ is employed to encode the region-referenced textual prompt and obtain the text tokens $\mathbf{z}_i^{(t)}$ with placeholders. The regions corresponding to the image sample x_i are grounded into region-referenced interleaved tokens by interleaving projected region tokens with text tokens:

$$\kappa_i = \mathbf{z}_i^{(t)} \oplus \mathbf{z}_i \in \mathbb{R}^{m \times d_t}, \tag{3}$$

where m denotes the number of tokens, and d_t represents the hidden dimension of text tokens.

The proposed TokenProjector $\mathcal{P}(\cdot)$ is optimised with a SigLIP-style contrastive objective [55], as shown in Equation (4), to establish stable multimodal grounding across regions. However, the SigLIP's text encoder remains frozen throughout training. The SigLIP-style contrastive objective leverages a sigmoid-based contrastive loss to align each matched pair of text and visual features in a shared embedding space.

$$\mathcal{L}_{Sig} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\log\frac{1}{1+\exp\left(-\sigma_{i,j}\cdot s(v_i,t_j)\right)}, \tag{4}$$

where B denotes the batch size, $\sigma_{i,j}$ represents a signed binary indicator, equals 1 if $i = j$ and -1 otherwise, and $s(v_i, t_j)$ is a similarity measure of each visual and semantic features pair.

3.3 Multimodal Interleaving

Traditional multimodal fusion approaches [11, 12, 19] typically integrate modalities at specific layers or at the decision stage, producing a unified embedding feature that limits the potential for cross-modal deep interaction over fine-grained

regions in Animal ReID. While the multimodal interleaving paradigm has been widely applied in LMMs tailored for various generation or reasoning tasks recently [18,33,49], its effectiveness in addressing fine-grained discriminative tasks, *e.g.*, re-identification, in pretrained foundation models remains underexplored. To fill the gap, we introduce a shallow Transformer-based Interleaver module that explicitly models structured cross-modal interactions at the region level with the multimodal interleaving mechanism. The shallow Transformer-based Interleaver module consists of a single multi-head self-attention layer, maintaining low computational overhead.

For an image sample x_i, we construct a region-referenced interleaved token sequence prepended with a learnable pooling token $\mathbf{z}_i^{pool}$:

$$\kappa_i = \left[\mathbf{z}_i^{pool}, \mathbf{z}_{i,1}^{(t)}, \mathbf{z}_{i,1}, \mathbf{z}_{i,2}^{(t)}, \mathbf{z}_{i,2}, \ldots, \mathbf{z}_{i,N}, \mathbf{z}_{i,M}^{(t)}\right], \tag{5}$$

where $\mathbf{z}_{i,M}^{(t)}$ denotes the M-th text token in the sequence. The sequence structure enables dense cross-modal interactions through the proposed Interleaver $\mathcal{F}(\cdot)$ with a self-attention mechanism:

$$\overline{\kappa_i} = \mathcal{F}\left(\kappa_i\right) \in \mathbb{R}^{d_t}, \tag{6}$$

where $\overline{\kappa_i}$ is the final representation of the pooling token with dense cross-modal interactions. The Interleaver module is optimised with the sigmoid loss, as illustrated in Equation (4), to stabilise the multimodal interleaving capability.

This multimodal interleaving design not only preserves the original semantic context of textual descriptions but also enhances guidance towards fine-grained, identity-relevant visual cues that can be challenging to fully convey in plain language. As a result, our approach, which leverages multimodal interleaving to enable dense cross-modal interactions across regions, highlights identity-discriminative features to distinguish individuals.

3.4 Overall Training Objective and Inference

To establish stable region-referenced interleaved token grounding and cross-modal interleaving, we first warm up the proposed TokenProjector and Interleaver using a sigmoid-based contrastive loss [55], as demonstrated in Equation (4), while the pretrained SigLIP model remains frozen. We then jointly optimise them with SigLIP's vision encoder to learn the discriminative features for Animal ReID under an overall training objective:

$$\mathcal{L}_{overall} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \mathcal{L}_{Sig}, \tag{7}$$

where $\mathcal{L}_{id}$ and $\mathcal{L}_{tri}$ represent the identity and triplet losses. Meanwhile, we freeze SigLIP's text encoder, vision projector, and text projector throughout training. Our progressive optimisation strategy allows the TokenProjector and Interleaver to transition from stable cross-modal alignment to identity-discriminative feature

specialisation. The identity loss is defined as:

$$\mathcal{L}_{id} = -\sum_{\phi=1}^{\Phi} q_\phi \log(p_\phi), \tag{8}$$

where $q_\phi = (1-\epsilon) \cdot \delta_{\phi,y_i^c} + \epsilon/\Phi$ represents the ground-truth identity distribution, and p_ϕ is the predicted probability for identity ϕ. Following prior work [13,15,21], we employ label smoothing to reduce the negative impact of potentially noisy labels in the training data. The smoothing factor (ϵ) targets to distribute a small fraction of probability mass uniformly across all identities, and the Kronecker delta (δ_{ϕ,y_i^c}), *e.g.*, a binary indicator, equals 1 when $\phi = y_i^c$; otherwise, it is 0. The triplet loss is formulated as:

$$\mathcal{L}_{tri} = \max\left(0, d_{pos} - d_{neg} + \gamma\right), \tag{9}$$

where d_{pos} and d_{neg} are the visual feature distances of the positive and negative sample pairs, and γ quantifies the margin that prevents the case when $d_{pos} = d_{neg}$ by enforcing a minimum separation between each pair.

During inference, we adopt the standard evaluation approach for query samples in $\mathcal{Q}$, rather than the region-aware multimodal interleaving used during training, which closely aligns with practical deployment settings. Specifically, we obtain the visual features of the query and gallery samples from SigLIP's vision encoder, while we do not incorporate the textual modality into the inference. The pseudocode of RAMI is present in the supplementary materials.

4 Experiments

Datasets and Baselines. We evaluate our framework on four publicly available Animal ReID benchmark datasets: (1) ATRW [20], (2) Giraffes [25], (3) NyalaData [8], and (4) SealID [30], as well as on a recently collected in-the-wild Stoat dataset. We compare our proposed framework with five representative state-of-the-art methods: (1) TransReID [13], (2) CLIP-ReID [21], (3) GloPER [6], (4) MDReID [9], and (5) CARE [52].

Evaluation Metrics. We employ two evaluation metrics: mean Average Precision (mAP) [51] and Cumulative Matching Characteristics (CMC) [26]. To evaluate the comprehensive re-identification capability of ReID models, mAP assesses the model's overall retrieval performance by averaging precision across all query samples in the Query Set $\mathcal{Q}$. However, CMC aims to measure the model's ability to identify the correct match among the top-k ranked gallery samples in the Gallery Set $\mathcal{G}$. We report both mAP and CMC accuracy at Rank-1 (%) averaged over ten runs, along with the corresponding 95% confidence intervals in our experiments.

Implementation Details. We leverage the pretrained SigLIP model to extract the text and visual features through text and vision encoders. Specifically, the text encoder has a Transformer-based architecture, and the vision encoder is

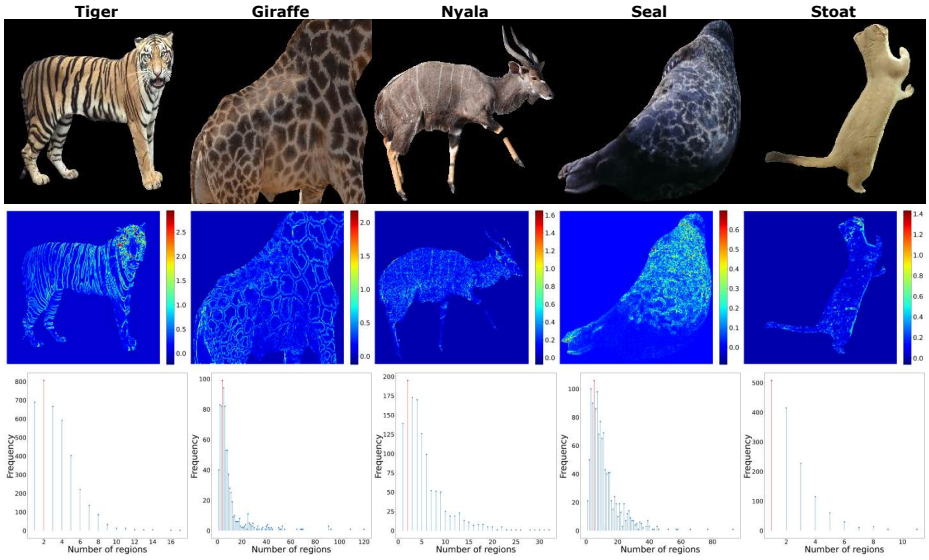


Fig. 3: Qualitative visualisation of image-driven region mining on benchmark and Stoat datasets. The top row shows the animal foreground mask produced by the pretrained segmentation model. The middle row depicts the pixel-wise texture importance score map within the foreground mask. The bottom row shows the frequency of candidate-region counts per image. The red lines represent the frequency modes in the datasets.

built on a variant of ViT. In our experiments, we employ the vision encoder with architecture SoViT-400m [1]. We implement the RAMI framework using PyTorch libraries. All experiments are conducted on NVIDIA HGX H200 GPUs. During the experiments, all images are resized to 384×384 with padding, and random horizontal flipping, cropping, and erasing [60] are applied to train both RAMI and state-of-the-art baselines. We follow the same training and test splits for the benchmark datasets as described in the papers. For our newly collected Stoat dataset, we allocate 50% of the data to training and 50% to testing. We use the Adaptive Moment Estimation with Weight Decay (AdamW) [24] optimiser to optimise all trainable components of RAMI. We warm up the TokenProjector and Interleaver for 10 epochs with a learning rate initialised at 1×10^{-3} and decayed by a cosine scheduler. We then jointly optimise SigLIP's vision encoder with TokenProjector and Interleaver for 60 epochs. The learning rate for SigLIP's vision encoder is initialised at 1×10^{-5}, while it is 1×10^{-4} for TokenProjector and Interleaver during the joint optimisation. Both are decayed using a cosine scheduler. The batch size is set to 32 for all the datasets during the training. Following the empirical experiments from prior work [52], we set the smoothing factor (ϵ) for the identity loss to 0.1. The margin (γ) for the triplet loss is set to 0.7 in the experiments.

4.1 Image-Driven Region-Level Pattern Mining

To construct informative, fine-grained region-level patterns potentially relevant to identity discrimination, we present an autonomous pipeline that employs lightweight, off-the-shelf image-processing operators without introducing additional trainable modules, thereby highly reducing computational overhead. We first extract the animal foreground from each image sample using a pretrained segmentation model, *e.g.*, SAM 3 [3], to prevent intervention from the surrounding environment. We then derive a pixel-wise texture importance score map that highlights regions within the foreground mask with strong local intensity gradients, such as distinctive coat markings or body parts. Furthermore, grayscale normalisation is applied to eliminate colour bias and constrain the scoring to the structural content of the foreground mask. Specifically, we compute the Sobel gradient magnitude [32] of the grayscale mask to estimate local texture strength and variation at the pixel level, effectively prioritising fine-grained regions that typically contain identity evidence. As shown in the middle row of Figure 3, regions with strong local intensity gradients have higher importance scores, such as the coat textures or markings, facial contours, and head regions of animals.

Due to the high variability in animal visual representations [52], such as pose or form changes and occlusions, we derive an image-driven score map computation with statistical tests to reduce the False Discovery Rate (FDR). Given an image sample x_i, we define a set of pixel-wise texture importance scores $S_i = \{s_1, s_2, \ldots, s_L\}$. To measure the significance and extremeness of the pixel-wise score map, we compute the p-value of each score in S_i as:

$$p\left(s_l\right) = \frac{1 + \left|\left\{s_k \in S_i : s_k \geq s_l\right\}\right|}{1 + L}, \tag{10}$$

where $1 \leq l \leq L$ and L denotes the cardinality of the set S_i. We set the significance level for the tests to 0.01% to mitigate FDR, given the extremely large number of pixels in images. Furthermore, we employ the Benjamini–Hochberg (BH) procedure to control FDR of the statistical tests at the 5% level for each image sample. To ensure stable and consistent region mining across all images within each dataset, we finally adopt the mode of the statistically significant candidate counts observed per image as the dataset-level region counts, as demonstrated in the bottom row of Figure 3. Specifically, the candidate regions with higher importance scores are prioritised based on dataset-level region counts.

4.2 Comparison with state-of-the-art Baselines

The evaluation results of the proposed RAMI framework and state-of-the-art baselines on four benchmark datasets and our Stoat dataset are presented in Table 1. For each metric in each dataset, the highest values are highlighted in bold, and the second-highest values are underscored. The results indicate that RAMI outperforms all state-of-the-art methods on mAP across all datasets, achieving an approximate 18.7% mAP gain on the NyalaData dataset.

Table 1: Comparison with state-of-the-art baselines across five Animal ReID datasets.

Method		ATRW	Giraffes	NyalaData	SealID	Stoat
TransReID [13]	mAP	44.6±0.9	25.1±1.7	9.0±0.1	11.2±0.4	58.1±1.4
	Rank-1	89.7±1.0	51.6±4.1	12.9±0.5	26.5±3.5	85.7±2.0
CLIP-ReID [21]	mAP	54.1±0.3	48.9±0.1	13.4±0.1	20.4±0.2	62.8±0.3
	Rank-1	94.2±0.2	77.2±0.8	21.9±0.3	57.0±0.5	**91.5±0.7**
GloPER [6]	mAP	37.8±0.0	20.7±0.0	9.3±0.0	10.4±0.0	28.9±0.0
	Rank-1	81.4±0.0	44.7±0.0	12.2±0.0	25.0±0.0	63.8±0.0
MDReID [9]	mAP	49.5±0.7	34.9±0.8	6.4±0.2	12.2±0.9	57.8±0.7
	Rank-1	91.9±0.6	66.7±1.8	7.7±0.8	34.6±2.0	89.3±0.4
CARE [52]	mAP	57.0±0.4	60.0±0.5	16.6±0.1	21.6±0.3	64.1±0.2
	Rank-1	95.4±0.2	80.9±0.7	29.0±0.2	58.5±1.1	**91.9±0.4**
RAMI (Ours)	mAP	**63.2±0.1**	**69.3±0.4**	**35.3±0.2**	**30.1±0.2**	**66.8±0.5**
	Rank-1	**97.4±0.1**	**88.0±0.2**	**58.8±1.1**	**71.0±0.1**	**91.6±0.5**

Table 2: Analysis of (i) the effectiveness of multimodal interleaving for Animal ReID and (ii) the comparison between global and regional patterns. SigLIP-FT fine-tunes only SigLIP's vision encoder without incorporating the multimodal interleaving mechanism. RAMI-Global relies on global animal patterns to enable the cross-modal interactions with our proposed framework.

Dataset	SigLIP-FT		RAMI-Global		RAMI	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
ATRW	55.1±0.3	94.2±0.2	61.4±0.3	96.5±0.2	**63.2±0.1**	**97.4±0.1**
Giraffes	67.2±0.5	87.1±0.6	69.0±0.2	88.2±0.5	69.3±0.4	88.0±0.2
NyalaData	25.8±0.8	41.8±1.5	34.1±0.3	56.6±0.8	**35.3±0.2**	**58.8±1.1**
SealID	20.2±0.4	48.8±0.4	28.8±0.2	69.6±0.4	**30.1±0.2**	**71.0±0.1**
Stoat	61.7±0.6	91.3±0.4	64.8±0.4	90.4±0.2	**66.8±0.5**	91.6±0.5

4.3 Ablation Studies

Effectiveness of Multimodal Interleaving. Table 2 analyses the impact of region-aware multimodal interleaving for Animal ReID by comparing RAMI with SigLIP-FT, which fine-tunes SigLIP's vision encoder only on pure textual descriptions via a simple cross-modal alignment without considering dense cross-modal interactions. The training for SigLIP-FT remains consistent with RAMI, including SigLIP-style alignment and ReID objectives. RAMI consistently outperforms SigLIP-FT across all datasets, demonstrating that structured region-level interleaving and dense cross-modal interactions substantially improve discriminative representation learning for Animal ReID.

Comparison of Global and Region-Level Patterns for Animal ReID. To illustrate the necessity of localised region modelling for Animal ReID, we compare the proposed region-aware design against a global pattern-based approach

under an identical multimodal interleaving mechanism in Table 2. RAMI-Global leverages generalised global patterns extracted by employing GloPER [6], a state-of-the-art animal pattern extraction method, to structure dense cross-modal interactions via an interleaving mechanism. The comparison results demonstrate that the localised, region-level animal patterns often enrich fine-grained identity-relevant cues, thereby enhancing the discrimination across individuals.

4.4 Further Analysis

Counterfactual Region-Level Reasoning. Models trained with strong auxiliary cues, *e.g.*, region-referenced interleaved tokens, often yield improved identity predictions. Yet, it remains unclear how they causally harness these cues to distinguish individuals rather than exploit spurious correlations. While attention weights derived from models provide indirect signals, they struggle with reasoning about whether a localised region is necessary for matching identities. We therefore incorporate a counterfactual reasoning mechanism over regions that evaluates how the paired similarity changes when each localised region is removed. To explicitly quantify the causal contribution of the most influential region in cross-modal interactions, we mask the region whose removal causes the largest decline in paired similarity.

We enable the counterfactual intervention within RAMI by simulating the removal of each region. Given an image sample x_i with its projected region tokens produced by the TokenProjector, we construct a counterfactual intervention by removing the n-th region, such as $\{\mathbf{z}_{i,1}, \mathbf{z}_{i,2}, \ldots, \mathbf{z}_{i,N}\} \setminus \{\mathbf{z}_{i,n}\}$, where $1 \leq n \leq N$. Meanwhile, we keep other projected region tokens and dense cross-modal interactions unchanged. We then compute the pairwise similarity between visual and semantic features in both the original and counterfactual settings. The counterfactual impact of region $r_{i,n}$ is thus defined as:

$$\Delta_{i,n}(x_i) = s\left(v_i, t_i\right) - s\left(v_i, t_i^{\setminus n}\right), \tag{11}$$

where v_i and t_i represent the visual and semantic features, $t_i^{\setminus n}$ is the counterfactual semantic features, and $s(\cdot,\cdot)$ denotes a similarity measure function. A large $\Delta_{i,n}$ suggests that the region $r_{i,n}$ provides necessary evidence and a large contribution to identifying individual animals. For each image sample x_i with its corresponding set of regions, we finally mask the most influential region $r_{i,n}$ by computing:

$$\arg\max_{n} \Big\{ \Delta_{i,n}(x_i) \Big\}. \tag{12}$$

To enable counterfactual regularisation, we introduce the counterfactual loss for positive and negative feature pairs, such as:

$$\mathcal{L}_{cf}^{+}(i) = \max\left(0, \tau + s\left(v_i, t_i^{\setminus n}\right) - s\left(v_i, t_i\right)\right)$$

and

$$\mathcal{L}_{cf}^{-}(i) = \sum_{j \neq i} \max\left(0, \tau + s\left(v_i, t_j\right) - s\left(v_i, t_j^{\setminus n}\right)\right),$$

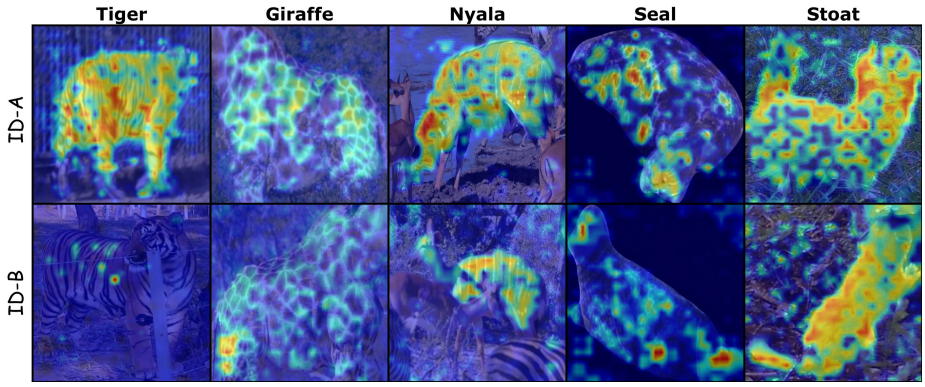


Fig. 4: Grad-CAM [44] visualisations of RAMI on distinct individuals across five Animal ReID datasets.

Table 3: Analysis of the causal contribution of the most influential region in RAMI.

Dataset	RAMI-CF		RAMI	
	mAP	Rank-1	mAP	Rank-1
ATRW	62.5±0.2	96.9±0.1	**63.2±0.1**	**97.4±0.1**
Giraffes	68.8±0.4	87.3±0.4	69.3±0.4	**88.0±0.2**
NyalaData	34.4±0.5	56.9±0.6	**35.3±0.2**	**58.8±1.1**
SealID	29.0±0.2	69.0±0.5	**30.1±0.2**	**71.0±0.1**
Stoat	65.6±0.4	91.9±0.3	**66.8±0.5**	91.6±0.5

where τ is a margin controlling the expected impact of regions. The full counterfactual loss is then given by:

$$\mathcal{L}_{cf} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{cf}^{+}(i) + \mathcal{L}_{cf}^{-}(i), \tag{13}$$

where B denotes the batch size.

Table 3 presents a counterfactual reasoning analysis with a comparison between RAMI-CF and RAMI. RAMI-CF integrates a counterfactual regularisation into RAMI to study the causal contribution of the most influential region for discriminating individuals. The performance degradation causally quantifies the gain of including the most influential region for Animal ReID in our framework.

Grad-CAM. We use Grad-CAM [44] to visualise where the model focuses for distinguishing individuals within each animal species. Figure 4 shows a comparison of the model's attention on distinct individuals of each species, indicating both variation in attention patterns and common species-level discriminative features. For instance, in tigers, ID-A highlights leg textures, while ID-B concentrates more on the body strides. For stoats, the model focuses on the head in ID-A, whereas in ID-B it shifts to the tail and fur. Additionally, the comparison

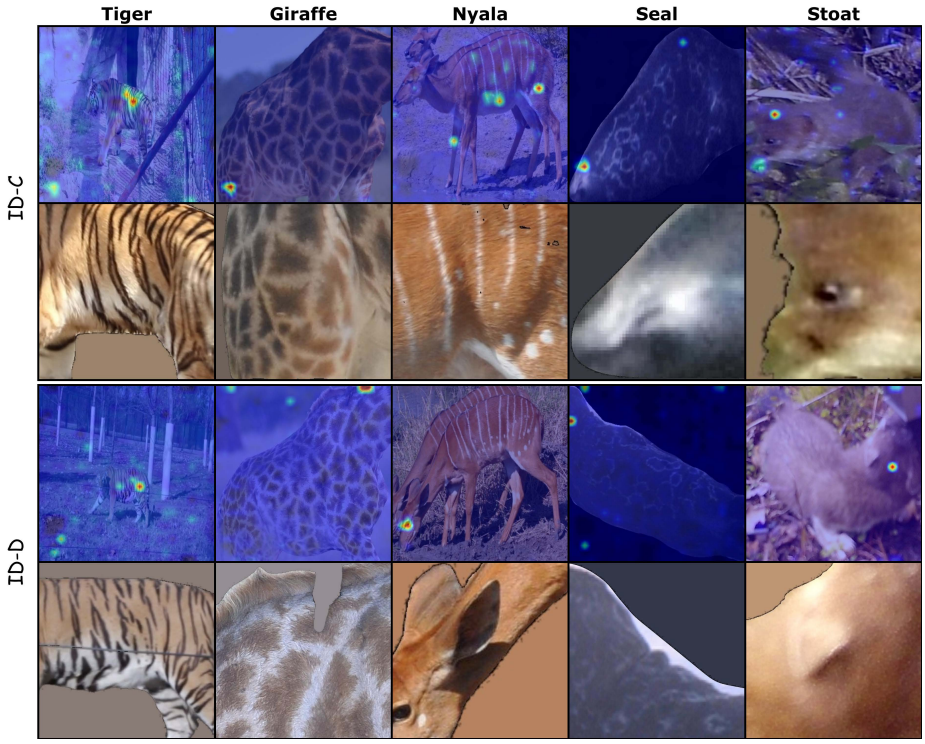


Fig. 5: Region samples from distinct individuals in the Query Set generated by using our proposed image-driven region mining pipeline. The Grad-CAM [44] is employed to illustrate the consistency between the regions and the model's attention for characterising identities.

reveals that the models consistently capture similar anatomical regions across individuals of the same species. For example, the coat patterns of giraffes, the body stripes of nyalas, and the skin markings of seals, each associated with various attention distributions, are attended to by the models, demonstrating that RAMI leverages reliable species-level discriminative cues to identify individuals.

Qualitative Visualisations of Animal Regions. To validate the effectiveness of region-level patterns for Animal ReID, we compare the regions of query samples with their corresponding Grad-CAM [44] visualisations. As shown in Figure 5, the localised regions for distinct individuals generated by our proposed region-mining pipeline are highly consistent with the regions highlighted by the models during individual discrimination. Furthermore, the region samples for all species also align with the anatomical regions shown in Figure 4. This qualitative analysis further demonstrates that RAMI consistently leverages identity-relevant visual cues enhanced by localised regions to identify individuals.

Computational Cost Analysis. Table 4 reports the computational cost and backward FLOPs of RAMI compared to the SigLIP-FT baseline. We explic-

Table 4: Computational cost and backward FLOPs analysis.

Model	# Total Parameters	# Trainable Parameters	Backward FLOPs
SigLIP-FT	878.01M	428.27M	1440.23G
RAMI	896.61M	446.87M	1504.84G
TokenProjector	2.66M	2.66M	–
Interleaver	15.94M	15.94M	–

itly show the memory overhead of the proposed TokenProjector and Interleaver in RAMI, demonstrating that our framework's memory overhead is negligible relative to the large number of parameters in the pretrained backbone.
Others. We further provide supporting evidence by conducting detailed sensitivity analyses in the supplementary materials (Appendix C).

5 Conclusion

This work presents a novel Region-Aware Multimodal Interleaving framework that leverages informative, fine-grained, localised regions with structure-rich visual cues to embed Animal ReID into the multimodal space and strengthen dense cross-modal interactions across regions via an interleaving mechanism. This approach explicitly enhances guidance towards complex, identity-discriminative visual cues, which are frequently challenging to describe in natural language. Furthermore, post-hoc counterfactual reasoning over regions causally illustrates the contribution of localised regions in RAMI. In addition, we propose an autonomous region-mining pipeline to prioritise fine-grained regions for distinguishing individual animals. Extensive experiments on benchmark and in-the-wild datasets demonstrate that RAMI consistently outperforms state-of-the-art baselines. Our work provides principled insights into multimodal interleaving for fine-grained discriminative tasks in VLMs and establishes a promising foundation for addressing real-world challenges in Animal ReID for ecological monitoring.

Acknowledgements

This research was partially supported by Royal Society Te Apārangi Catalyst: Seeding General (25-UOA-023-CSG). We also gratefully acknowledge the Centre of Machine Learning for Social Good and the Advanced Machine Learning and Data Analytics Research (MARS) Lab at the University of Auckland.

References

1. Alabdulmohsin, I., Zhai, X., Kolesnikov, A., Beyer, L.: Getting vit in shape: scaling laws for compute-optimal model design. In: The Thirty-Seventh Annual Conference on Neural Information Processing Systems (2023)

2. Bergman, N., Yitzhaky, Y., Halachmi, I.: Biometric identification of dairy cows via real-time facial recognition. animal (2024)
3. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Coll-Vinent, D.S., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., HAZRA, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollar, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment anything with concepts. In: The Fourteenth International Conference on Learning Representations (2026)
4. Cermak, V., Picek, L., Adam, L., Neumann, L., Matas, J.: Wildfusion: Individual animal identification with calibrated similarity fusion. In: European Conference on Computer Vision Workshops. Springer (2024)
5. Čermák, V., Picek, L., Adam, L., Papafitsoros, K.: Wildlifedatasets: An open-source toolkit for animal re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
6. Chen, B., Koh, Y.S., Dobbie, G.: Gloper: Unsupervised animal pattern extraction from local reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
7. Chen, P., Swarup, P., Matkowski, W.M., Kong, A.W.K., Han, S., Zhang, Z., Rong, H.: A study on giant panda recognition based on images of a large proportion of captive pandas. Ecology and Evolution (2020)
8. Dlamini, N., Zyl, T.L.v.: Automated identification of individuals in wildlife population using siamese neural networks. In: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI) (2020)
9. Feng, Y., Li, J., Hu, J., Zhang, Y., Tan, L., Ji, J.: MDReID: Modality-decoupled learning for any-to-any multi-modal object re-identification. In: The Thirty-Ninth Annual Conference on Neural Information Processing Systems (2025)
10. Gómez-Vargas, N., Alonso-Fernández, A., Blanquero, R., Antelo, L.T.: Re-identification of fish individuals of undulate skate via deep learning within a few-shot context. Ecological Informatics (2023)
11. Guo, C., Zhang, L.: A model-level fusion-based multi-modal object detection and recognition method. In: 2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT) (2023)
12. Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., Wang, C., Guo, L., Meng, W., Zhang, X., Xu, R., Xu, S.: Multimodal fusion and vision–language models: A survey for robot vision. Information Fusion (2025)
13. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
14. He, Z., Qian, J., Yan, D., Wang, C., Xin, Y.: Animal re-identification algorithm for posture diversity. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2023)
15. Hou, S., Huang, P., Wang, Z., Liu, Y., Li, Z., Zhang, M., Huang, Y.: Openanimals: Revisiting person re-identification for animals towards better generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
16. Jiao, B., Liu, L., Gao, L., Wu, R., Lin, G., WANG, P., Zhang, Y.: Toward re-identifying any animal. In: The Thirty-Seventh Annual Conference on Neural Information Processing Systems (2023)
17. Li, D., Su, H., Jiang, K., Liu, D., Duan, X.: Fish face identification based on rotated object detection: dataset and exploration. Fishes (2022)

18. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In: The Thirteenth International Conference on Learning Representations (2025)
19. Li, L., Chen, B., Zou, X., Xing, J., Tao, P.: Uv-mamba: A dcn-enhanced state space model for urban village boundary identification in high-resolution remote sensing images. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2025)
20. Li, S., Li, J., Tang, H., Qian, R., Lin, W.: Atrw: A benchmark for amur tiger re-identification in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia (2020)
21. Li, S., Sun, L., Li, Q.: Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
22. Li, Y., Zhao, D., Qiao, T., Wu, Y., Pang, B., Koh, Y.S.: Metawild: A multimodal dataset for animal re-identification with environmental metadata. In: Proceedings of the 33rd ACM International Conference on Multimedia (2025)
23. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
25. Miele, V., Dussert, G., Spataro, B., Chamaillé-Jammes, S., Allainé, D., Bonenfant, C.: Revisiting animal photo-identification using deep metric learning and network analysis. Methods in Ecology and Evolution (2021)
26. Moon, H., Phillips, P.J.: Computational and performance aspects of pca-based face-recognition algorithms. Perception (2001)
27. Moosa, M., Cheikh, F.A., Beghdadi, A., Ullah, M.: Self-supervised animal detection, tracking & re-identification. In: 2024 IEEE Thirteenth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE (2024)
28. Moskvyak, O., Maire, F., Dayoub, F., Baktashmotlagh, M.: Learning landmark guided embeddings for animal re-identification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (2020)
29. Nepovinnykh, E., Chelak, I., Eerola, T., Immonen, V., Kälviäinen, H., Kholiavchenko, M., Stewart, C.V.: Species-agnostic patterned animal re-identification by aggregating deep local features. International Journal of Computer Vision (2024)
30. Nepovinnykh, E., Eerola, T., Biard, V., Mutka, P., Niemi, M., Kunnasranta, M., Kälviäinen, H.: Sealid: Saimaa ringed seal re-identification dataset. Sensors (2022)
31. Nepovinnykh, E., Vilkman, A., Eerola, T., Kälviäinen, H.: Re-identification of saimaa ringed seals from image sequences. In: Scandinavian Conference on Image Analysis (SCIA). Springer (2023)
32. Nguyen, T.P., Chae, D.S., Park, S.J., Yoon, J.: A novel approach for evaluating bone mineral density of hips based on sobel gradient-based map of radiographs utilizing convolutional neural network. Computers in Biology and Medicine (2021)
33. Nie, M., Wang, C., Han, J., Xu, H., Zhang, L.: Towards unified multimodal interleaved generation via group relative policy optimization. In: The Thirty-Ninth Annual Conference on Neural Information Processing Systems (2025)
34. Pang, B., Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Cabin: Debiasing vision-language models using backdoor adjustments. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25 (2025)

35. Pang, B., Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Libra: Measuring bias of large language model from a local context. In: European Conference on Information Retrieval. Springer (2025)
36. Patton, P.T., Cheeseman, T., Abe, K., Yamaguchi, T., Reade, W., Southerland, K., Howard, A., Oleson, E.M., Allen, J.B., Ashe, E., et al.: A deep learning approach to photo-identification demonstrates high performance on two dozen cetacean species. Methods in Ecology and Evolution (2023)
37. Perneel, M., Adriaens, I., Verwaeren, J., Aernouts, B.: Dynamic multi-behaviour, orientation-invariant re-identification of holstein-friesian cattle. Sensors (2025)
38. Qiao, T., Walker, C., Cunningham, C., Koh, Y.S.: Thematic-lm: a llm-based multi-agent system for large-scale thematic analysis. In: Proceedings of the ACM on Web Conference 2025 (2025)
39. Qiao, T., Zhao, D., Walker, C., Cunningham, C., Koh, Y.S.: Zero-shot domain generalisation via prompt-driven feature refinement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2026)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
41. Ravoor, P.C., Sudarshan, T.: Deep learning methods for multi-species animal re-identification and tracking - a survey. Computer Science Review (2020)
42. Schneider, S., Taylor, G.W., Kremer, S.C.: Similarity learning networks for animal individual re-identification: an ecological perspective. Mammalian Biology (2022)
43. Schneider, S., Taylor, G.W., Linquist, S., Kremer, S.C.: Past, present and future approaches using computer vision for animal re-identification from camera trap data. Methods in Ecology and Evolution (2019)
44. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
45. de Silva, E.M., Kumarasinghe, P., Indrajith, K.K., Pushpakumara, T.V., Vimukthi, R.D., de Zoysa, K., Gunawardana, K., de Silva, S.: Feasibility of using convolutional neural networks for individual-identification of wild asian elephants. Mammalian Biology (2022)
46. Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Wang, Y., Rao, Y., Liu, J., Huang, T., Wang, X.: Generative multimodal models are in-context learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
47. Van Zyl, T.L., Woolway, M., Engelbrecht, B.: Unique animal identification using deep transfer learning for data fusion in siamese networks. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION) (2020)
48. Wang, L., Ding, R., Zhai, Y., Zhang, Q., Tang, W., Zheng, N., Hua, G.: Giant panda identification. IEEE Transactions on Image Processing (2021)
49. Wang, X., Cui, Y., Wang, J., Zhang, F., Wang, Y., Zhang, X., Luo, Z., Sun, Q., Li, Z., Wang, Y., Yu, Q., Zhao, Y., Ao, Y., Min, X., Men, C., Wu, B., Zhao, B., Zhang, B., Wang, L., Liu, G., He, Z., Yang, X., Liu, J., Lin, Y., Wang, Z., Huang, T.: Multimodal learning with next-token prediction for large multimodal models. Nature (2026)
50. Weideman, H., Stewart, C., Parham, J., Holmberg, J., Flynn, K., Calambokidis, J., Paul, D.B., Bedetti, A., Henley, M., Pope, F., et al.: Extracting identifying

contours for african elephants and humpback whales using a learned appearance model. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020)
51. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
52. Wu, Y., Zhao, D., Li, Y., Alajas, M., Glen, A.S., Zhang, J., Dobbie, G., Wilson, D., Koh, Y.S.: 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 (2026)
53. Xia, P., Han, S., Qiu, S., Zhou, Y., Wang, Z., Zheng, W., Chen, Z., Cui, C., Ding, M., Li, L., Wang, L., Yao, H.: MMIE: Massive multimodal interleaved comprehension benchmark for large vision-language models. In: The Thirteenth International Conference on Learning Representations (2025)
54. Xiao, Z., Dai, W., Li, C., Liang, W., Chen, X.: Enhanced multi-breed cattle face recognition in complex environments using attention-based deep learning. The Visual Computer (2025)
55. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
56. Zhang, T., Zhao, Q., Da, C., Zhou, L., Li, L., Jiancuo, S.: Yakreid-103: A benchmark for yak re-identification. In: 2021 IEEE International Joint Conference on Biometrics (IJCB) (2021)
57. Zhao, D., Koh, Y.S., Dobbie, G., Hu, H., Fournier-Viger, P.: Symmetric self-paced learning for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
58. Zhao, D., Zhang, J., Hu, H., Fournier-Viger, P., Dobbie, G., Koh, Y.S.: Balancing invariant and specific knowledge for domain generalization with online knowledge distillation. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25 (2025)
59. Zhao, D., Zhang, J., Hu, H., Fournier-Viger, P., Dobbie, G., Koh, Y.S.: Unlearning during training: Domain-specific gradient ascent for domain generalization. In: The Fourteenth International Conference on Learning Representations (2026)
60. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)