ARK: A BENCHMARK FOR REASONING RE-IDENTIFICATION IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Re-identification (ReID) is commonly formulated as a retrieval problem, yet ranking alone does not always resolve identity: an observation of the correct identity may appear among the top-k candidates without being ranked first, while open-set queries require recognizing when no valid match exists. We formulate this unresolved post-retrieval stage as **Reasoning ReID**, where a multimodal large language model (MLLM) verifies retrieved candidates before producing the final identity decision. To systematically study the capabilities required for such verification, we introduce **ARK**, a benchmark for Reasoning ReID using Animal ReID as a challenging testbed. ARK spans 17 species and seven controlled protocols organized into three capability families: **Visual Identity Perception**, **Evidence Integration & Association**, and **Decision Reliability**. An evaluation of 21 proprietary and open-weight MLLMs alongside a human reference shows that current models can benefit from complementary evidence, while their reliability remains uneven under demanding verification conditions. We further instantiate Reasoning ReID in retrieval pipelines built on CLIP and SigLIP, showing that selectively invoking MLLM verification for uncertain retrievals can improve end-to-end identification in both closed- and open-set settings.

1 INTRODUCTION

Re-identification (ReID) is commonly formulated as a retrieval problem: given a query image, an embedding-based model ranks a gallery according to learned representation similarity (Ahmed et al., 2015). This formulation is effective for narrowing the search space, but retrieval does not necessarily complete the identity decision. In challenging cases, an image of the correct identity may appear among the top-ranked results without being ranked first, leaving several plausible candidates that still require verification (Zhang et al., 2024a; Guo et al., 2024). Moreover, a nearest-neighbor retrieval rule always produces a top-ranked candidate unless an explicit rejection mechanism is introduced, including in open-set cases where the queried identity is absent from the gallery. These cases define a distinct post-retrieval verification problem: determining the identity relationship between a query and a retrieved candidate set, including rejecting the set when no valid match exists.

As illustrated in Figure 1, multimodal large language models (MLLMs) provide an interface for this post-retrieval stage because they can process visual inputs and textual instructions and produce candidate-level decisions (Lu et al., 2025). We term this paradigm **Reasoning ReID**: MLLM-based identity verification over a retrieved candidate set, potentially integrating multiple views, auxiliary context, and explicit rejection before producing an identity decision. This perspective extends ReID evaluation beyond ranking quality toward capabilities required for candidate verification. Yet these capabilities remain poorly understood in ReID: ReID evaluations do not jointly characterize whether MLLMs can extract fine-grained identity evidence, integrate complementary evidence and resolve identity associations, and make reliable candidate-level decisions under challenging conditions.

To systematically evaluate these capabilities, we introduce **ARK** (**A**nimal **R**eID reasoning benchmar**K**), a benchmark for Reasoning ReID using Animal ReID as a testbed. ARK formulates post-retrieval identity verification as structured question-answer tasks, enabling scalable and standardized evaluation of MLLM verifiers. Animal ReID is particularly demanding because individuals of the

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.

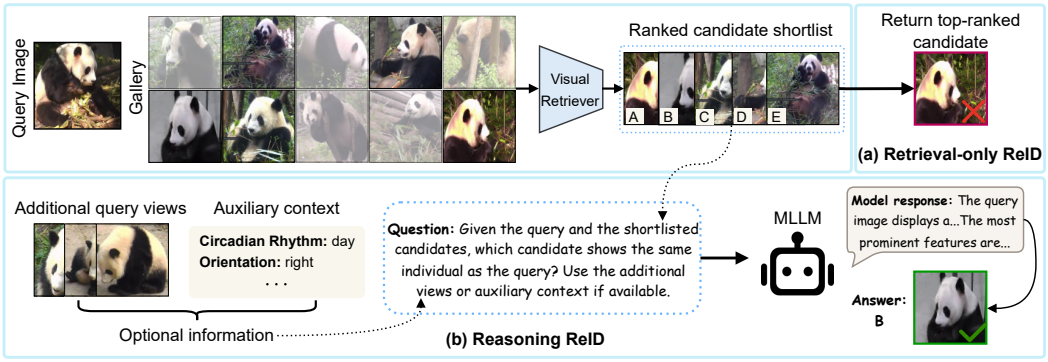


Figure 1: **Retrieval-only ReID versus Reasoning ReID.** Given a query image and a gallery, a visual retriever produces a ranked candidate shortlist. (a) In a basic retrieval-only pipeline, the top-ranked candidate is used as the final prediction. (b) Reasoning ReID treats the shortlist as an intermediate result and uses an MLLM to perform candidate-level identity verification against the query, optionally incorporating additional query views or auxiliary context before producing the final identity decision. The illustrated example shows a single-match case in which the correct identity is present in the shortlist but is not ranked first; the general formulation also supports multiple valid matches and rejection when no valid match exists.

same species often share similar coarse-grained appearance while lacking readily verbalizable cues such as clothing or accessories, increasing reliance on subtle individual-level visual evidence (Yang et al., 2023; Li et al., 2025b). ARK provides controlled capability diagnosis through seven protocols organized into three capability families: **Visual Identity Perception**, **Evidence Integration & Association**, and **Decision Reliability**. Across these protocols, ARK evaluates identity verification under uncorrupted and corrupted visual evidence, multiple views, auxiliary context, multiple valid matches, the absence of a valid match, and misleading external suggestions.

We evaluate 21 proprietary and open-weight MLLMs alongside a human reference. On fine-grained identity verification, the strongest MLLM achieves 53.49% accuracy. Controlled analyses show that joint multi-view inference improves over aggregated single-view predictions by up to 3.97 percentage points, while auxiliary context improves image-only verification by up to 9.10 percentage points. Robustness and failure-mode analyses reveal sensitivity to visual corruption, larger candidate sets, conflicting visual evidence, and misleading prompts, alongside distinct precision–recall profiles in multi-match association and asymmetric behavior in open-set decisions. We instantiate Reasoning ReID in retrieval pipelines built on CLIP and SigLIP, where an uncertainty-aware routing strategy selectively invokes the MLLM verifier for uncertain retrievals. In closed-set identification, the best routed configuration improves accuracy by 1.12 percentage points to 25.50%, outperforming both retrieval-only and always-on verification. In the open-set setting, routing improves both match-present accuracy and match-absent rejection relative to calibrated thresholding.

In summary, our contributions are threefold. First, we formalize **Reasoning ReID** as a post-retrieval identity verification paradigm and introduce **ARK**, a controlled diagnostic benchmark spanning seven protocols across three capability families. Second, through a large-scale evaluation of 21 proprietary and open-weight MLLMs, we systematically characterize model capabilities and failure modes across visual identity perception, evidence integration and association, and decision reliability. Third, we provide retrieval-derived validation of Reasoning ReID, showing that uncertainty-aware routing improves closed-set identification and, in the open-set setting, improves both match-present accuracy and match-absent rejection.

2 Related Work

ReID and Wildlife ReID. Re-identification (ReID) seeks to recognize the same individual across observations under varying viewpoints and conditions. A dominant formulation learns identity-discriminative embeddings and evaluates them through query–gallery retrieval and ranking metrics (Ahmed et al., 2015; Zhang et al., 2024a). Wildlife ReID extends this setting to individual

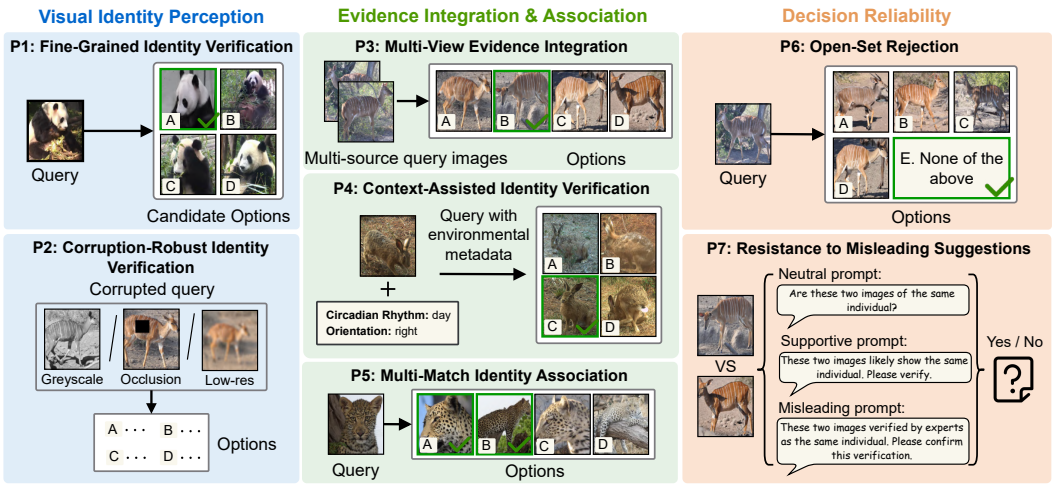


Figure 2: ARK comprises seven diagnostic protocols organized into three capability families. **Visual Identity Perception** evaluates fine-grained identity verification under uncorrupted (P1) and corrupted (P2) visual evidence. **Evidence Integration & Association** evaluates multi-view evidence integration (P3), context-assisted verification using circadian rhythm and spatial orientation cues (P4), and multi-match identity association (P5). **Decision Reliability** evaluates rejection when no valid match is present (P6) and identity judgments under neutral, supportive, and misleading prompts with the image pair held fixed (P7).

animals, motivating datasets for diverse species and capture conditions (Li et al., 2025b). WildlifeDatasets (Čermák et al., 2024) provides unified tooling for accessing and preprocessing heterogeneous animal ReID datasets, while WildlifeReID-10k (Adam et al., 2025) standardizes closed- and open-set retrieval evaluation across wildlife species. These resources establish retrieval-oriented data and evaluation infrastructure, but retrieval metrics alone do not characterize the downstream verification behavior required to resolve ambiguity among top-ranked candidates. ARK is complementary to these retrieval-oriented resources: rather than evaluating representation quality or gallery ranking, it studies Reasoning ReID, where one or more retrieved candidates must be verified as matches to the query, or the candidate set rejected when no valid match exists.

MLLMs for Fine-Grained and Multi-Image Understanding. Multimodal large language models have progressed from predominantly single-image understanding toward jointly processing multiple visual inputs and interleaved image–text contexts. Recent models such as Mantis (Jiang et al., 2024) and LLaVA-NeXT-Interleave (Li et al., 2025a) are explicitly designed for multi-image instruction following. Nevertheless, fine-grained visual comparison and instance-level grounding remain challenging for current MLLMs (Ye et al., 2025; Xu et al., 2025). Recent ReID-oriented work, such as LLaVA-ReID, further demonstrates the potential of MLLMs for interactive multi-image identity verification (Lu et al., 2025). Such verification requires comparing identity-specific evidence across observations while remaining robust to variations in viewpoint, appearance, and the amount of available visual evidence. ARK complements these advances by systematically evaluating the capabilities required for MLLM-based identity verification.

Multimodal Evaluation Benchmarks. Recent benchmarks evaluate MLLMs over multi-image inputs, covering semantic, spatial, and temporal reasoning (Meng et al., 2025; Wang et al., 2025; Ying et al., 2024). Specialized suites further probe capabilities such as low-level visual perception, visual hallucination, temporal event understanding, and multi-context grounding (Fu et al., 2024; Zhang et al., 2024b; Wang et al., 2024). However, these benchmarks do not specifically target individual-level identity discrimination across observations of highly similar subjects. Identity verification poses an instance-level challenge: models must distinguish subtle differences between visually similar individuals while remaining robust to appearance changes of the same individual across viewpoints and conditions; in open-set cases, they must also recognize when no valid match is present. ARK studies these challenges within Reasoning ReID through controlled evaluation of candidate-level identity verification across seven diagnostic protocols.

3 REASONING REID AND ARK

3.1 REASONING REID FORMULATION

Conventional ReID takes a query observation q and a gallery $\mathcal{G}$, using a visual retriever to rank gallery observations according to their similarity to the query. Reasoning ReID treats the resulting shortlist as an intermediate result and introduces an MLLM-based verification stage for the identity decision. We represent the query as a set $\mathcal{Q}$ containing one or more observations of the same underlying identity and the candidate set as $\mathcal{C} = \{c_1, \ldots, c_K\}$. Given $\mathcal{Q}$, $\mathcal{C}$, and optional auxiliary context, the verifier predicts the subset $\mathcal{M} \subseteq \mathcal{C}$ whose observations correspond to the query identity. A singleton $\mathcal{M}$ denotes a single-match decision, multiple elements denote multiple valid matches, and $\mathcal{M} = \varnothing$ denotes rejection when no candidate matches the query. Pairwise identity verification is recovered as the special case $K = 1$, where selecting the sole candidate denotes the same identity and $\mathcal{M} = \varnothing$ denotes different identities.

3.2 ARK CAPABILITY TAXONOMY

As illustrated in Figure 2, ARK operationalizes Reasoning ReID through seven controlled diagnostic protocols organized into three capability families. **Visual Identity Perception** focuses on extracting identity-relevant evidence from uncorrupted and corrupted visual observations. **Evidence Integration & Association** evaluates how models integrate complementary evidence from multiple views and auxiliary context, and how they resolve identity associations when multiple candidates may match the query. **Decision Reliability** examines whether models make appropriate identity decisions when no valid match exists or when external suggestions conflict with the available evidence.

Visual Identity Perception. P1: Fine-Grained Identity Verification evaluates whether an MLLM can identify the candidate observation that matches the query identity, with distractors drawn from different identities of the same species. **P2: Corruption-Robust Identity Verification** retains the same single-match objective but applies controlled visual corruptions to the query observation, including grayscale conversion, occlusion, and low-resolution degradation. P1 and P2 characterize identity verification under uncorrupted and corrupted visual evidence.

Evidence Integration & Association. P3: Multi-View Evidence Integration extends single-observation verification to multiple query observations of the same identity, evaluating whether the model can integrate evidence across views to identify the matching candidate. **P4: Context-Assisted Identity Verification** augments the visual query with auxiliary context, including circadian rhythm and spatial orientation, and evaluates whether such context can support visual identity verification. **P5: Multi-Match Identity Association** extends single-match verification to candidate sets containing multiple valid matches, requiring the model to identify all candidate observations that correspond to the query identity. P3–P5 characterize identity verification when identity evidence is available across multiple query observations or supplemented by auxiliary context, and when multiple candidate observations may correspond to the same identity as the query.

Decision Reliability. P6: Open-Set Rejection extends identity verification to candidate sets in which a valid match may be absent, evaluating whether the model can select a matching candidate when a valid match is present and reject the candidate set when none matches the query identity. **P7: Resistance to Misleading Suggestions** evaluates pairwise identity verification under neutral, supportive, and misleading prompts while holding the image pair fixed. The supportive prompt provides a suggestion consistent with the ground-truth identity relation, whereas the misleading prompt provides an inconsistent suggestion, enabling a controlled assessment of sensitivity to textual suggestions. P6 and P7 characterize decision reliability when candidate sets may contain no valid match or when identity judgments are exposed to external textual suggestions.

3.3 BENCHMARK CONSTRUCTION

Data Sources and Scope. ARK builds on 11 upstream animal ReID sources indexed by WildlifeDatasets (Čermák et al., 2024), covering beluga, bird, chimpanzee, giraffe, panda, leopard, lion, dolphin, nyala, seal, and whale shark. We additionally include six MetaWild species: deer, hare, penguin, pukeko, stoat, and wallaby (Li et al., 2025b). The WildlifeDatasets-indexed sources are selected based on low zero-shot Rank-1 retrieval performance under SigLIP, providing challenging

cases for fine-grained identity verification; Appendix A.3 further describes the selection scope and benchmark statistics. The MetaWild species provide metadata on circadian rhythm and spatial orientation used in P4. Overall, ARK covers **17 species, 3,085 identities, and 58,150 images**, from which we construct **187,865 evaluation instances** across the seven protocols. ARK releases annotations, source mappings, and code to reconstruct, verify, and evaluate benchmark inputs from authorized upstream sources without redistributing third-party images. Appendix G details this end-to-end pipeline; source provenance, access, and licensing are documented in Appendix A.2.

Controlled Instance Construction. ARK uses model-agnostic controlled sampling to construct identity-verification instances. The base format comprises one query and four candidate observations. In single-match settings, one candidate is sampled from a different observation of the query identity, while the remaining three are sampled from different identities of the same species. Same-species distractors preserve fine-grained identity ambiguity and require individual-level discrimination. This controlled construction provides a consistent reference configuration for capability diagnosis independently of any retrieval model. In Section 4.4, we evaluate Reasoning ReID on retrieval-generated top-k candidate shortlists produced by CLIP and SigLIP.

Protocol Instantiation. P1 uses the base single-query, single-match configuration. P2 applies grayscale conversion, occlusion, or low-resolution degradation to the query and states the applied corruption type in the instruction. P3 uses $N \in \{2, 3, 4\}$ query observations of the same identity while retaining the four-candidate single-match setting. P4 evaluates the same verification task under four query-side conditions: image-only input, image with spatial-orientation context, image with circadian-rhythm context, and image with both contextual cues. P5 places exactly two observations of the query identity among the four candidates without specifying the number of valid matches in the instruction. P6 augments the four-candidate format with an explicit no-match response. Match-present instances use the base single-match construction; in match-absent instances, all four candidates are sampled from other identities of the same species. P7 uses a pairwise format in which each image pair is evaluated under neutral, supportive, and misleading prompts, with the images held fixed across conditions; the model makes a binary same- or different-identity decision in each condition. Detailed prompt templates and protocol examples are provided in Appendix H.

Quality Control. To limit identity over-representation and candidate-set redundancy, ARK applies explicit balancing and diversity controls during sampling. The selection frequency of each query identity is capped, and previously unused images are prioritized when an identity is sampled repeatedly to increase within-identity visual diversity. For distractor sampling, ARK records the distractor sets previously used for each query and resamples proposed sets whose Jaccard similarity to prior sets exceeds a species-adaptive threshold, thereby limiting repeated negative configurations. A fixed threshold is unsuitable across species with different identity-pool sizes: an overly restrictive threshold can make sampling infeasible for small pools, whereas a permissive threshold allows substantial redundancy in large pools. ARK therefore derives the threshold for each species from the random-overlap distribution determined by the available identity pool and the number of distractors. Candidate order is randomized during generation to mitigate positional bias. The adaptive threshold derivation and sampling procedure are detailed in Appendix B.2.

4 Experiments

4.1 Experimental Setup

Baseline. We evaluate 21 proprietary and open-weight MLLMs. The proprietary models include Gemini-3.1-Pro, GPT-5.4, and Claude-Opus-4.6, while the open-weight models include the Qwen3.5, Qwen3-VL, Gemma3, and LLaVA families. We additionally report a human reference based on evaluations by six non-expert participants on a stratified subset of 1,926 ARK instances, with 22 additional instances used exclusively to measure inter-annotator agreement. Details of the human evaluation are provided in Appendix E.

Evaluation Metrics. We report accuracy for P1–P4. For P5, which contains multiple valid matches, we report precision, recall, F1, and Exact Match (EM), where an instance is considered correct only when the predicted match set exactly matches the ground-truth set. For P6, we report Match-Present Accuracy (AccMP) and Match-Absent Accuracy (AccMA) for match-present and match-absent cases, respectively. For P7, we report Neutral Accuracy (AccN), Supportive Accuracy (AccS), and

Misleading Accuracy (AccM) under the neutral, supportive, and misleading conditions, respectively. Protocol-level results are reported as unweighted macro-averages across species. We additionally report identity-aware macro-averages and bootstrap intervals in Appendix F.2.

Implementation Details. We evaluate open-weight models locally on two NVIDIA H200 GPUs and access proprietary models through their respective official APIs. For each benchmark instance, all models receive the same protocol-specific textual prompt, with a temperature of 0 and a maximum output length of 2,048 tokens. Model responses are parsed using the same deterministic rules. If a response cannot be parsed, we allow up to ten total generation attempts and score only the first parseable response; responses that remain unparsable are counted as incorrect.

4.2 Main Results

Visual Identity Perception. As shown in Table 1, fine-grained identity verification remains challenging for current MLLMs. Under P1, the strongest MLLM, Claude-Opus-4.6, achieves 53.49% accuracy. For context, the separately evaluated human reference on the stratified subset is 66.67%. Lower-capacity models such as LLaVA-13B and Gemma3-27B remain close to the 25% chance level, while performance increases markedly with model scale within the Qwen families. For example, P1 accuracy increases from 24.73% for Qwen3.5-0.8B to 46.67% for Qwen3.5-35B and 52.82% for Qwen3.5-122B, showing a clear scaling trend in fine-grained identity perception. P2 further shows that this capability remains sensitive to visual degradation: corruption reduces accuracy across all representative models, including Claude-Opus-4.6 from 53.49% to 41.96% and Qwen3.5-122B from 52.82% to 40.85%.

Table 1: Main results on ARK across the seven Reasoning ReID protocols. Results are reported as unweighted macro-averages over the species used by each protocol: P1–P3 and P5–P7 use the 11 WildlifeDatasets-indexed species, whereas P4 uses the six MetaWild species that provide the required contextual metadata. P3 reports the $N = 2$ multi-view setting; P4 compares image-only input with full context; P6 reports match-present (AccMP) and match-absent (AccMA) accuracy; and P7 reports neutral (AccN) and misleading (AccM) accuracy. The best MLLM result in each column among the models shown is bolded. Human results are reported only as a reference. Complete results for all 21 MLLMs are provided in Appendix J.

Model	**Visual Identity Perception**		**Evidence Integration & Association**				**Decision Reliability**			
	P1	**P2**	**P3**	**P4**		**P5**	**P6**		**P7**	
	Acc	Acc	Acc	Image	Full	EM	AccMP	AccMA	AccN	AccM
LLaVA-13B	25.00	24.53	24.74	24.02	25.43	9.26	20.18	4.66	49.49	**44.84**
Gemma3-27B	24.83	24.32	24.25	24.26	25.73	22.77	21.18	5.75	49.14	44.61
Qwen3-VL-2B	27.75	25.34	29.28	29.29	30.83	10.85	14.92	3.85	51.17	30.90
Qwen3-VL-30B	36.96	30.79	39.79	36.77	43.67	23.12	32.70	10.01	58.85	6.89
Qwen3-VL-235B	42.75	34.74	46.27	43.63	51.55	31.98	39.02	13.16	62.78	2.89
Qwen3.5-0.8B	24.73	24.08	24.44	23.73	24.69	10.33	9.45	2.91	48.92	44.12
Qwen3.5-35B	46.67	37.21	50.38	41.93	49.38	36.12	34.17	9.92	71.25	12.75
Qwen3.5-122B	52.82	40.85	57.27	49.50	58.60	**43.41**	**43.86**	14.72	72.66	6.44
Gemini-3.1-Pro	48.02	38.48	52.66	**89.52**	93.78	31.22	4.89	48.08	68.60	8.92
GPT-5.4	43.64	35.10	48.32	88.68	90.85	25.48	4.21	47.93	69.17	2.38
Claude-Opus-4.6	**53.49**	**41.96**	**58.12**	89.49	**95.19**	35.70	5.54	**48.14**	**73.52**	4.61
Human	66.67	53.03	71.21	77.78	78.47	67.17	59.60	82.83	87.88	87.37

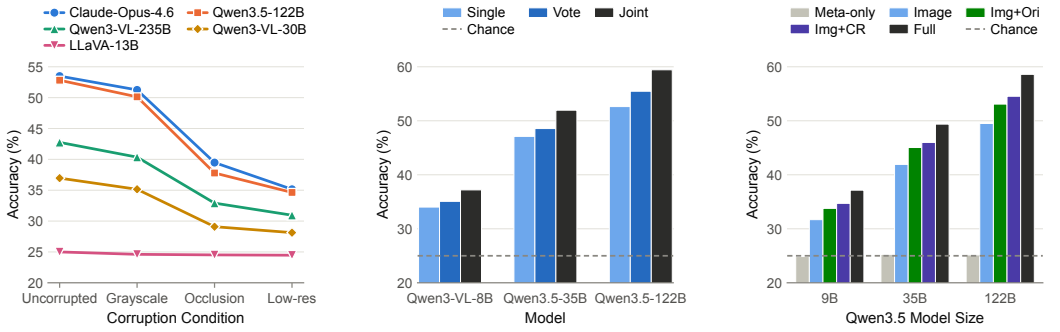
Evidence Integration & Association. Under P3, higher-capacity models generally benefit from multiple query observations. For example, accuracy increases from 52.82% in P1 to 57.27% with two query observations for Qwen3.5-122B, and from 53.49% to 58.12% for Claude-Opus-4.6, whereas several smaller models show little or no improvement. P4 shows that auxiliary context can improve identity verification, with Qwen3.5-122B increasing from 49.50% under image-only input to 58.60% with full context. Notably, proprietary models achieve 88.68–89.52% even in the image-only setting. Since MetaWild is publicly available, these unusually high absolute scores should be interpreted cautiously, as prior exposure to the source images cannot be excluded for these propri-

etary models. Under P5, exact multi-match association remains difficult: the best MLLM, Qwen3.5-122B, achieves 43.41% EM. For context, the separately evaluated human reference on the stratified subset is 67.17%. Taken together, P3 and P4 show that visual or contextual evidence can improve identity verification, while P5 shows that exactly recovering the valid-match set remains challenging.

Decision Reliability. P6 reveals a pronounced asymmetry between match-present and match-absent decisions across current MLLMs. For example, Qwen3.5-122B achieves 43.86% AccMP but only 14.72% AccMA, whereas Claude-Opus-4.6 shows the opposite pattern, with 5.54% AccMP and 48.14% AccMA. These contrasting patterns highlight a substantial imbalance between candidate selection and rejection in the open-set setting. P7 reveals a different reliability issue: models that perform strongly under neutral prompts can degrade sharply under misleading suggestions. Claude-Opus-4.6 drops from 73.52% AccN to 4.61% AccM, while Qwen3.5-122B drops from 72.66% to 6.44%. For context, the separately evaluated human reference on the stratified subset is 87.88% under Neutral prompts and 87.37% under Misleading prompts.

4.3 Capability Diagnostics

Beyond aggregate protocol scores, we further examine how MLLMs process and combine identity-relevant evidence. Figure 3 investigates three complementary aspects of this behavior: robustness to visual corruption, gains from joint multi-view inference beyond aggregation of independent single-view predictions, and the contribution of auxiliary context to visual identity verification.



(a) Corruption robustness. (b) Multi-view integration. (c) Context-assisted verification.

Figure 3: (a) Accuracy of five representative MLLMs on uncorrupted P1 queries and under grayscale, occlusion, and low-resolution degradation in P2. (b) P3 with $N = 3$, comparing mean single-view accuracy (**Single**), majority voting over the three single-view predictions (**Vote**), and joint inference over all three query views (**Joint**). (c) P4 across **Meta-only**, **Image**, **Img+Ori**, **Img+CR**, and **Full** conditions for three Qwen3.5 models, where Ori and CR denote spatial orientation and circadian-rhythm context. The dashed line denotes 25% chance performance.

Corruption Robustness. Figure 3a shows that robustness varies markedly across corruption types. Grayscale conversion causes only modest accuracy losses for the stronger models, whereas occlusion and low-resolution degradation produce substantially larger declines. For Qwen3.5-122B, accuracy decreases from 52.82% on uncorrupted queries to 50.13% under grayscale, 37.79% under occlusion, and 34.64% under low-resolution degradation. Claude-Opus-4.6 shows the same pattern, with accuracy decreasing from 53.49% to 51.25%, 39.48%, and 35.15%, respectively.

Multi-View Evidence Integration. Figure 3b evaluates P3 with three query views ($N = 3$). For each instance, we evaluate each query image independently against the same candidate set. **Single** reports the mean accuracy over the three single-view runs. **Vote** applies majority voting to the three predictions, with instances lacking a majority counted as incorrect. **Joint** instead presents all three query images simultaneously with the same candidate set and produces a single prediction. Joint consistently achieves the highest accuracy across three representative models. For Qwen3.5-122B, accuracy increases from 52.63% under Single to 55.47% under Vote and 59.44% under Joint, corresponding to an improvement of 3.97 percentage points over majority voting. Similar improvements are observed for Qwen3.5-35B and Qwen3-VL-8B, at 3.40 and 2.12 percentage points, respectively.

Context-Assisted Verification. Figure 3c evaluates the four P4 input conditions alongside a metadata-only control. **Meta-only** removes the query image while retaining its auxiliary context, whereas **Image** uses the query image without context. **Img+Ori** and **Img+CR** add spatial orientation and circadian-rhythm context, respectively, while **Full** combines both cues with the image. Meta-only performance remains close to the 25% chance level across all three models, whereas image-only input yields substantially higher accuracy. Adding either contextual cue further improves performance, and Full achieves the highest accuracy for every model. Relative to Image, Full increases accuracy from 31.69% to 37.15% for Qwen3.5-9B, from 41.93% to 49.38% for Qwen3.5-35B, and from 49.50% to 58.60% for Qwen3.5-122B, corresponding to gains of 5.46, 7.45, and 9.10 percentage points, respectively. These results indicate that auxiliary context provides complementary evidence for visual identity verification.

4.4 Retrieval-Derived Validation

To complement ARK's controlled candidate-set evaluation, we assess Reasoning ReID within a two-stage retrieval pipeline. Across the 11 WildlifeDatasets-indexed species, CLIP and SigLIP produce top-5 candidate shortlists, with Qwen3.5-122B serving as the MLLM verifier. In the closed-set setting, we compare **Retrieval-only**, which directly returns the top-ranked identity, **Always-MLLM**, which invokes the verifier for every query, and **Routed-MLLM**, which invokes the verifier only for uncertain retrievals. In the open-set setting, we include **Retrieval+Threshold** with a calibrated similarity threshold for rejection. Calibration and species-level results are in Appendix C.

Table 2: Retrieval-derived validation under closed- and open-set settings. Closed-set results report Rank-1 accuracy, Rank-5 recall (R@5), and verification accuracy; open-set results report match-present accuracy (AccMP) and match-absent rejection accuracy (AccMA).

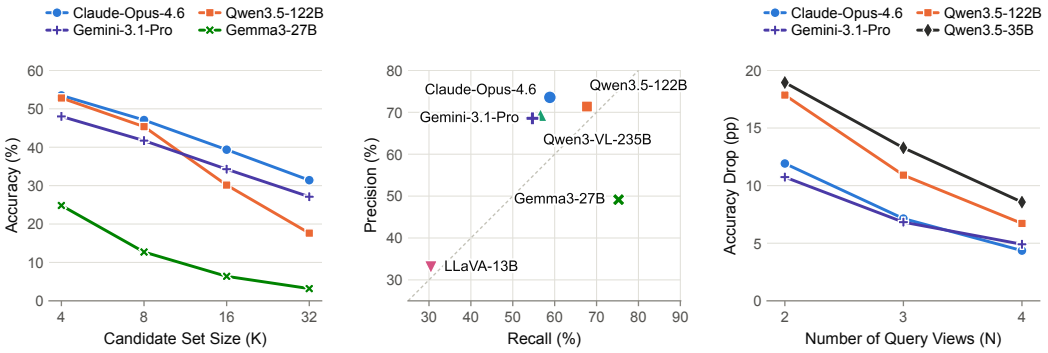
Retriever	Closed-Set				Open-Set							
	Retrieval		Verification		AccMP				AccMA			
	Rank-1	R@5	Always	Routed	Retrieval	+Thr.	Always	Routed	Retrieval	+Thr.	Always	Routed
CLIP	23.13	42.44	20.60	**24.15**	47.15	42.90	41.94	**48.63**	0.00	45.03	14.95	**48.00**
SigLIP	24.38	44.27	22.47	**25.50**	48.43	44.30	42.38	**49.43**	0.00	45.68	14.90	**49.28**

As shown in Table 2, the substantial gap between Rank-1 accuracy and Rank-5 recall indicates considerable potential headroom for candidate-level verification. Always-MLLM does not consistently exploit this headroom, whereas Routed-MLLM improves closed-set accuracy from 23.13% to 24.15% with CLIP and from 24.38% to 25.50% with SigLIP. In the open-set setting, Retrieval-only has no rejection mechanism and therefore yields zero AccMA, while calibrated thresholding provides substantial rejection capability at the cost of some AccMP. Always-MLLM enables direct shortlist rejection without relying on an explicit retrieval-score threshold, but its AccMA remains lower than that of Retrieval+Threshold. Routed-MLLM improves both sides of this trade-off: over Retrieval+Threshold, AccMP/AccMA increase from 42.90/45.03 to 48.63/48.00 with CLIP and from 44.30/45.68 to 49.43/49.28 with SigLIP. These results support uncertainty-aware routing of MLLM verification as complementary to retrieval-based matching and rejection.

4.5 Robustness and Failure-Mode Analysis

Beyond the capability diagnostics above, we further characterize how MLLMs respond to changes in verification difficulty, decision structure, and evidence consistency. Figure 4 examines sensitivity to larger candidate sets, precision–recall asymmetry in multi-match association, and robustness to conflicting visual evidence across query views.

Candidate-Set Size Sensitivity. Figure 4a shows a consistent decline in verification accuracy as the candidate set grows. Claude-Opus-4.6 drops from 53.49% at $K = 4$ to 31.42% at $K = 32$, while Gemini-3.1-Pro decreases from 48.02% to 27.11%. Qwen3.5-122B achieves 52.82% and 45.36% at $K = 4$ and $K = 8$, respectively, but declines more sharply to 30.14% at $K = 16$ and 17.63% at $K = 32$. In contrast, Gemma3-27B remains close to the corresponding random-guessing baseline as K increases. Notably, model ordering changes with candidate-set size: Qwen3.5-122B outperforms



(a) Candidate-set size sensitivity. (b) Precision–recall behavior. (c) Visual distractor robustness.

Figure 4: (a) Verification accuracy across candidate-set sizes $K = 4, 8, 16, 32$ for four MLLMs. (b) Precision and recall on P5 for six MLLMs. (c) Accuracy drop when introducing one same-species, identity-inconsistent distractor among $N = 2, 3, 4$ query views, relative to the corresponding clean P3 condition. The dashed diagonal in (b) marks equal precision and recall.

Gemini-3.1-Pro at $K = 4$ and $K = 8$, but is surpassed by it at $K = 16$ and $K = 32$. This indicates that small-K performance does not fully characterize a model's sensitivity to candidate-set size.

Precision–Recall Behavior. Figure 4b reveals distinct selection behaviors in multi-match identity association. Claude-Opus-4.6, Qwen3-VL-235B, and Gemini-3.1-Pro exhibit precision-oriented profiles, whereas Qwen3.5-122B shows a more balanced precision–recall profile, with 71.38% precision and 67.71% recall. Gemma3-27B shows the opposite pattern, achieving 73.70% recall but only 49.77% precision, indicating over-selection, while LLaVA-13B remains low on both metrics. A precision-oriented profile is consistent with more conservative selection, prioritizing correctness among selected candidates at the risk of missing valid matches; a recall-oriented profile is consistent with broader selection, recovering more valid matches while admitting more false positives. These contrasting profiles reveal distinct failure modes in multi-match association, motivating separate reporting of precision and recall to distinguish under-selection from over-selection.

Visual Distractor Robustness. Figure 4c evaluates robustness when one identity-inconsistent distractor is introduced among the query views and reports the accuracy drop relative to the clean P3 condition. At $N = 2$, accuracy decreases by 11.92, 17.86, 10.74, and 18.95 percentage points for Claude-Opus-4.6, Qwen3.5-122B, Gemini-3.1-Pro, and Qwen3.5-35B, respectively. As more query views of the correct identity are added while keeping a single distractor, the accuracy drop decreases consistently; at $N = 4$, the corresponding drops decrease to 4.36, 6.72, 4.91, and 8.57 percentage points. These results show that multi-view Reasoning ReID remains vulnerable to conflicting visual evidence, while additional views of the correct identity reduce the effect of a single distractor.

5 Conclusion

We introduced Reasoning ReID as a post-retrieval identity verification paradigm and presented ARK, a controlled benchmark spanning seven protocols across visual identity perception, evidence integration and association, and decision reliability. Evaluating 21 proprietary and open-weight MLLMs alongside a human reference shows that current models remain limited in fine-grained identity verification and exhibit distinct weaknesses under corrupted visual evidence, in open-set rejection, and under misleading suggestions. Controlled analyses further show that MLLMs can benefit from complementary evidence: joint multi-view inference improves over majority-vote aggregation of single-view predictions, while auxiliary context improves image-only verification. Beyond controlled diagnosis, retrieval-derived experiments with CLIP and SigLIP show that uncertainty-aware MLLM routing improves closed-set identification over retrieval-only and always-on verification, while improving both match-present accuracy and match-absent rejection over calibrated thresholding in the open-set setting. Together, these results extend conventional retrieval-based ReID with candidate-level reasoning and establish ARK as a benchmark for systematically studying when and how MLLMs can contribute to reliable identity decisions.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design. We did not use generative AI tools to generate synthetic datasets, develop theoretical or conceptual frameworks, propose or refine hypotheses, implement methods, assist with translation, clean or reformat datasets, or interpret experimental results; tasks involving mathematical claims, proofs, or qualitative or thematic analysis are not applicable to this work. Additionally, we used generative AI tools to suggest experimental parameters, create or edit software code, support brainstorming and information search, edit the manuscript for readability, and suggest the structure of the paper. We reviewed all AI-assisted work, and AI-assisted code was manually inspected and tested to ensure consistency with the intended experimental procedures and expected behavior. All experimental decisions were made by the authors, and all interpretations of results and final scientific claims were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ARK is designed to support research on multimodal reasoning for animal re-identification, with potential applications in ecological monitoring and wildlife conservation. The benchmark is constructed from existing animal ReID resources and does not involve the collection of new wildlife imagery. Consistent with the governance requirements of upstream resources, ARK does not redistribute third-party image assets; these remain subject to the licenses and terms of their original providers and must be obtained through authorized channels. ARK releases only its own benchmark annotations, metadata, and code, with source-level provenance, access, licensing, and redistribution information documented in Appendix A.2. The human reference evaluation was conducted by six non-expert participants on a stratified subset of the benchmark; no sensitive personal information was collected, and only aggregate evaluation results are reported. Details of the human evaluation protocol are provided in Appendix E. More broadly, downstream use of animal ReID systems should respect applicable data-governance, privacy, and domain-specific requirements, including appropriate safeguards against exposing sensitive wildlife location information or other ecological metadata that could facilitate misuse.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3.3 describes the data sources, benchmark construction, protocol instantiation, and quality-control procedures used to build ARK, while Section 4.1 specifies the evaluated models, metrics, inference settings, and response-parsing procedure. Appendices A.2, B, and D document source-level provenance and access information, detailed benchmark-construction procedures, and evaluation and aggregation procedures, respectively. Exact prompt templates, the human reference evaluation, and the retrieval-derived validation setup are documented in Appendices H, E, and C. Appendix G details the end-to-end image-free reconstruction and evaluation workflow, including deterministic preprocessing, fingerprint verification, unified evaluation, and reproducibility safeguards. We release the ARK-authored annotations, benchmark metadata, source mappings, and code for benchmark reconstruction and evaluation; upstream image assets remain subject to their original source terms and are not redistributed by ARK. An anonymized code repository is available at `https://anonymous.4open.science/r/Ark-30470`.

REFERENCES

Lukas Adam, Vojtech Cermak, Kostas Papafitsoros, and Lukas Picek. Wildlifereid-10k: Wildlife re-identification dataset with 10k individual animals. In *CVPR*, pp. 2124–2134, 2025.

Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pp. 3908–3916, 2015.

Botswana Predator Conservation Trust. Panthera pardus csv custom export, 2022. URL `https://lila.science/datasets/leopard-id-2022/`. Retrieved from African Carnivore Wildbook, 2022-04-28.

Vojtěch Čermák, Lukas Picek, Lukáš Adam, and Kostas Papafitsoros. Wildlifedatasets: An open-source toolkit for animal re-identification. In *WACV*, pp. 5941–5951. IEEE, 2024.

Nkosikhona Dlamini and Terence L. Van Zyl. Automated identification of individuals in wildlife population using siamese neural networks. In *ISCMI*, pp. 224–228, 2020. doi: 10.1109/ISCMI51676.2020.9311574.

André C. Ferreira, Liliana R. Silva, Francesco Renna, Hanja B. Brandl, Julien P. Renoult, Damien R. Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. *Methods Ecol. Evol.*, 11(9):1072–1085, 2020. doi: 10.1111/2041-210X.13436.

Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar S. Kühl, and Joachim Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *GCPR*, pp. 51–63, 2016. doi: 10.1007/978-3-319-45886-1_5.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, pp. 148–166. Springer, 2024.

Wenxuan Guo, Zhiyu Pan, Yingping Liang, Ziheng Xi, Zhicheng Zhong, Jianjiang Feng, and Jie Zhou. Lidar-based person re-identification. In *CVPR*, pp. 17437–17447, 2024.

Jason Holmberg, Brad Norman, and Zaven Arzoumanian. Estimating population size, structure, and residency time for whale sharks rhincodon typus through collaborative photo-identification. *Endanger. Species Res.*, 7:39–53, 2009. doi: 10.3354/esr00186.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *TMLR*, 2024. ISSN 2835-8856.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *ICLR*, 2025a.

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 *ACMMM*, pp. 13009–13015, 2025b.

LILA BC. Trail camera images of new zealand animals, 2024. URL `https://lila.science/datasets/nz-trailcams/`.

Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. LLaVA-reID: Selective multi-image questioner for interactive person re-identification. In *ICML*, 2025. URL `https://openreview.net/forum?id=c4EEnWu9FE`.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *ICLR*, 2025.

Vincent Miele, Gaspard Dussert, Bruno Spataro, Simon Chamaillé-Jammes, Dominique Allainé, and Christophe Bonenfant. Revisiting animal photo-identification using deep metric learning and network analysis. *Methods Ecol. Evol.*, 12(5):863–873, 2021. doi: 10.1111/2041-210X.13577.

Ekaterina Nepovinnykh, Tuomas Eerola, Vincent Biard, Piia Mutka, Marja Niemi, Mervi Kunnasranta, and Heikki Kälviäinen. SealID: Saimaa ringed seal re-identification dataset. *Sensors*, 22 (19):7602, 2022. doi: 10.3390/s22197602.

Cameron Trotter, Georgia Atkinson, Matt Sharpe, Kirsten Richardson, A. Stephen McGough, Nick Wright, Ben Burville, and Per Berggren. NDD20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. *arXiv preprint arXiv:2005.13359*, 2020. URL `https://arxiv.org/abs/2005.13359`.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *ICLR*, 2025. URL https://openreview.net/forum?id=TrVYEZtSQH.

Le Wang, Rizhi Ding, Yuanhao Zhai, Qilin Zhang, Wei Tang, Nanning Zheng, and Gang Hua. Giant panda identification. *TIP*, 30:2837–2849, 2021. doi: 10.1109/TIP.2021.3055627.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *ACL*, pp. 416–442, 2024.

Wild Me. Beluga id 2022. LILA BC, 2022. URL https://lila.science/datasets/beluga-id-2022/.

Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *ICCV*, pp. 17675–17687, 2025.

Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, pp. 1472–1481, 2023.

Xiaojun Ye, Guanbao Liang, Chun Wang, Liangcheng Li, Pengfei Ke, Rui Wang, Bingxin Jia, Gang Huang, Qiao Sun, and Sheng Zhou. M4bench: A benchmark of multi-domain multi-granularity multi-image understanding for multi-modal large language models. In *IJCAI*, pp. 6848–6856, 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, jiayi lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *ICML*, 2024. URL https://openreview.net/forum?id=R4Ng8zYaiz.

Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhaung Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *CVPR*, pp. 22000–22009, 2024a.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-Bench$^+$: A Benchmark for Multi-Modal Foundation Models on Low-Level Vision From Single Images to Pairs. *TPAMI*, 46(12):10404–10418, 2024b.

A Dataset Sources, Governance, and Provenance

A.1 Data Sources and Benchmark Composition

ARK draws on 11 animal ReID datasets indexed by WildlifeDatasets (Čermák et al., 2024) and six species from MetaWild (Li et al., 2025b). The WildlifeDatasets-indexed datasets cover beluga, bird, chimpanzee, giraffe, panda, leopard, lion, dolphin, nyala, seal, and whale shark, and are used to construct P1–P3 and P5–P7. P4 uses six MetaWild species, namely deer, hare, penguin, pukeko, stoat, and wallaby, for which the contextual metadata used in Context-Assisted Identity Verification are available. Detailed source-level provenance, access conditions, and usage terms are provided in Appendix A.2. Overall, ARK contains 17 species and 3,085 identities. The finalized benchmark uses 58,150 images to construct 187,865 evaluation instances across the seven protocols, corresponding to 281,534 rendered prompts across their protocol-specific conditions. Figure 5 summarizes the distribution of benchmark-used images and evaluation instances across species, together with the distribution of evaluation instances across the seven protocols and three capability families.

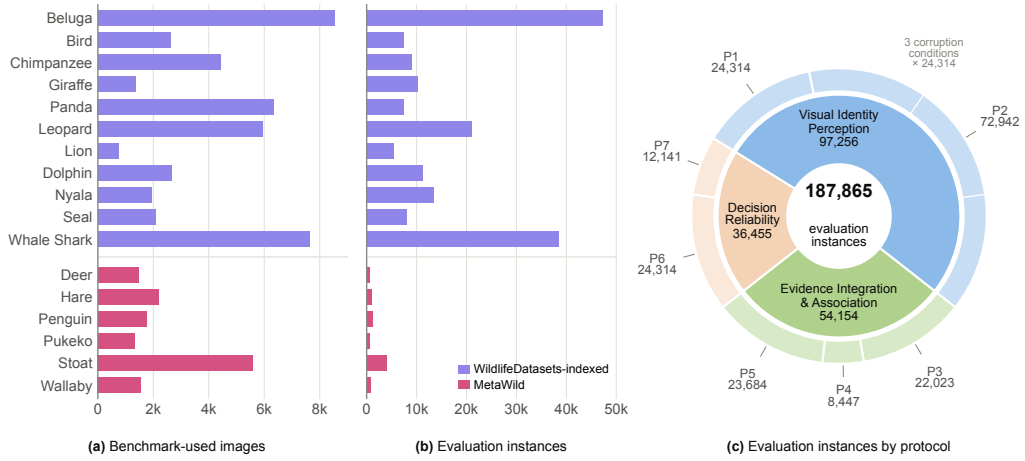


Figure 5: Statistical overview of ARK. (a) Distribution of the 58,150 benchmark-used images across the 17 species. (b) Distribution of the 187,865 evaluation instances across species. (c) Distribution of evaluation instances across the seven protocols, grouped by the three capability families; P2 contains three corruption conditions with 24,314 instances each.

A.2 Source Provenance, Access, and Usage Terms

ARK is constructed from existing animal ReID datasets with source-specific access and usage conditions. WildlifeDatasets (Čermák et al., 2024) provides a common interface for accessing and preprocessing many of these datasets, while the original datasets remain subject to their respective source terms. Table 3 summarizes the provenance, access source, and stated license or usage terms for each upstream resource used in ARK.

The ARK release contains benchmark annotations, metadata, source mappings, and construction and evaluation code, but does not redistribute upstream image assets. These images are obtained separately from the original or author-provided sources listed in Table 3. The released source mappings and construction code support reconstruction of the benchmark inputs once the corresponding images have been obtained from their source locations. Licenses for ARK-authored annotations, metadata, and software are specified separately in the ARK release.

A.3 Dataset Selection and Benchmark Statistics

The 11 WildlifeDatasets-indexed sources used for P1–P3 and P5–P7 were selected during source screening to emphasize challenging fine-grained identity verification cases, with low zero-shot Rank-1 retrieval performance under SigLIP used as a selection signal. The resulting subset is therefore challenge-oriented rather than intended to constitute a statistically representative sample of

Table 3: Source provenance, access, and usage terms for upstream image resources used by ARK.

Source	Access source	License or usage terms
Beluga ID 2022 (Wild Me, 2022)	LILA BC / Wild Me	CDLA-Permissive-1.0
BirdIndividualID (Ferreira et al., 2020)	Authors' repository and archived research release	No explicit dataset-level license identified in the reviewed source materials
CTai (Freytag et al., 2016)	Official chimpanzee-faces repository	Non-commercial use, including research and education, with citation of the associated publication
Giraffes (Miele et al., 2021)	Authors' institutional data repository	Described by the authors as freely available; no explicit dataset-level license identified in the reviewed source materials
IPanda50 (Wang et al., 2021)	Official iPanda-50 repository and author-provided download mirrors	No explicit dataset-level license identified in the reviewed source materials
LeopardID2022 (Botswana Predator Conservation Trust, 2022)	LILA BC / Wild Me	CDLA-Permissive-1.0
LionData (Dlamini & Van Zyl, 2020)	Authors' wildlife-reidentification repository	No explicit dataset-level license identified in the reviewed source materials
NDD20 (Trotter et al., 2020)	Newcastle University research-data repository	CC BY-NC-SA 4.0
NyalaData (Dlamini & Van Zyl, 2020)	Authors' wildlife-reidentification repository	No explicit dataset-level license identified in the reviewed source materials
SealID (Nepovinnykh et al., 2022)	Official research-data repository	CC BY 4.0
WhaleSharkID (Holmberg et al., 2009)	LILA BC / Wild Me	CDLA-Permissive-1.0
MetaWild (Li et al., 2025b) / NZ-TrailCams (LILA BC, 2024)	MetaWild metadata release; LILA BC / NZ-TrailCams	MetaWild annotations and metadata: CDLA-Permissive-1.0; NZ-TrailCams images: CDLA-Permissive-1.0

wildlife ReID datasets. Within each selected source, identities satisfying the protocol-specific construction requirements are retained in the benchmark pool; no identity is manually selected or excluded on the basis of appearance or model performance. Results on this subset should accordingly be interpreted as performance within a deliberately challenging collection of wildlife ReID sources rather than as estimates of performance across the broader wildlife ReID domain.

P4 uses the six MetaWild species because they provide the circadian-rhythm and spatial-orientation metadata required for Context-Assisted Identity Verification. Table 4 reports the final number of identities, distinct benchmark-used images, and evaluation instances for each source and protocol.

A.4 Metadata Scope and Potential Pretraining Exposure

ARK adopts a protocol-specific metadata design, retaining contextual information only when required by the corresponding evaluation setting. In P4, the contextual cues consist of circadian rhythm and spatial orientation, which are provided alongside the query image for Context-Assisted Identity Verification. The benchmark inputs are otherwise defined directly by the information specified for each protocol.

Because some upstream image resources used by ARK are publicly available, benchmark images may have been encountered during model pretraining. This possibility is particularly relevant to P4, where proprietary models exhibit unusually high image-only performance on MetaWild. Since the training data of closed models cannot be verified through black-box API access, we treat this pattern only as a possible indication of prior exposure. Our analysis of P4 therefore focuses on the observed model behavior and controlled within-model comparisons rather than attributing the results to a specific cause.

Table 4: Source- and species-level statistics for the final ARK benchmark. Images denotes the number of distinct benchmark-used images. P1–P7 report the numbers of evaluation instances for the corresponding protocols; P2 aggregates its three corruption conditions. A dash indicates that a protocol is not instantiated for the corresponding source.

Source	Species	IDs	Images	P1	P2	P3	P4	P5	P6	P7	Total
Beluga ID 2022	Beluga	788	8,523	6,475	19,425	5,520	–	6,217	6,475	3,233	47,345
BirdIndividualID	Bird	50	2,627	1,000	3,000	1,000	–	1,000	1,000	499	7,499
CTai	Chimpanzee	71	4,424	1,214	3,642	1,197	–	1,210	1,214	607	9,084
Giraffes	Giraffe	178	1,368	1,361	4,083	1,355	–	1,359	1,361	679	10,198
IPanda50	Panda	50	6,347	1,000	3,000	1,000	–	1,000	1,000	497	7,497
LeopardID2022	Leopard	430	5,936	2,904	8,712	2,375	–	2,740	2,904	1,451	21,086
LionData	Lion	94	740	740	2,220	726	–	740	740	369	5,535
NDD20	Dolphin	82	2,655	1,491	4,473	1,491	–	1,491	1,491	745	11,182
NyalaData	Nyala	237	1,942	1,847	5,541	1,568	–	1,735	1,847	921	13,459
SealID	Seal	57	2,077	1,061	3,183	1,061	–	1,061	1,061	530	7,957
WhaleSharkID	Whale shark	543	7,633	5,221	15,663	4,730	–	5,131	5,221	2,610	38,576
MetaWild	Deer	38	1,475	–	–	–	737	–	–	–	737
MetaWild	Hare	60	2,185	–	–	–	1,132	–	–	–	1,132
MetaWild	Penguin	77	1,755	–	–	–	1,147	–	–	–	1,147
MetaWild	Pukeko	30	1,319	–	–	–	547	–	–	–	547
MetaWild	Stoat	253	5,589	–	–	–	4,038	–	–	–	4,038
MetaWild	Wallaby	47	1,555	–	–	–	846	–	–	–	846
Total	–	**3,085**	**58,150**	**24,314**	**72,942**	**22,023**	**8,447**	**23,684**	**24,314**	**12,141**	**187,865**

B Benchmark Construction Details

B.1 Identity-Balanced Sampling

To reduce the over-representation of identities with many images, each eligible identity is subject to a per-protocol sampling budget of at most 20. Within the query-based protocols, previously unused images are prioritized for the query role before image reuse, increasing within-identity query-image coverage. Candidate order is randomized before prompt construction, and sampling uses fixed protocol-specific random seeds for reproducibility.

B.2 Distractor-Set Diversity Control

Repeatedly pairing the same query identity with highly similar distractor sets can reduce the diversity of the resulting verification instances. ARK therefore tracks the distractor sets previously used for each query identity and constrains the Jaccard similarity between a newly proposed set S and previously accepted sets S':

$$J(S, S') = \frac{|S \cap S'|}{|S \cup S'|}.$$

Because the expected overlap between randomly sampled distractor sets depends on the size of the available identity pool, ARK uses a species-adaptive rather than fixed similarity threshold. For a pool of P available distractor identities and a distractor-set size m, let X denote the number of shared identities between two independently sampled m-element sets. Under uniform sampling, X follows a hypergeometric distribution:

$$\Pr(X = x) = \frac{\binom{m}{x}\binom{P-m}{m-x}}{\binom{P}{m}}.$$

ARK selects the smallest overlap limit c^* satisfying

$$\Pr(X > c^*) < 0.05,$$

Algorithm 1 Controlled candidate-set sampling in ARK.

Require: Prepared image pool $\mathcal{D}$ and protocol p
Require: Per-identity sampling budgets $\{B_i\}$
Initialize distractor-set history $\mathcal{H}_i \leftarrow \varnothing$ for each eligible identity i
while some eligible identity has remaining budget **do**
  Select an identity i with remaining budget
  Select a query image q, prioritizing images not previously used in the query role
  Sample additional query views and positive observations required by p
  Let m be the required number of distractor identities
  Compute the adaptive threshold τ from the available identity pool and m
  $best \leftarrow \varnothing$
  $o_{best} \leftarrow \infty$
  for $t = 1, \ldots, 20$ **do**
    Sample m distinct non-query identities and one image from each
    Let S be the resulting distractor-identity set
    $o \leftarrow \max_{S' \in \mathcal{H}_i} |S \cap S'|$, with $o \leftarrow 0$ if $\mathcal{H}_i = \varnothing$
    if $o < o_{best}$ **then**
      $best \leftarrow S$, $o_{best} \leftarrow o$
    end if
    if $\mathcal{H}_i = \varnothing$ **or** $J(S, S') \leq \tau$ for all $S' \in \mathcal{H}_i$ **then**
      $best \leftarrow S$
      break
    end if
  end for
  $S \leftarrow best$
  $\mathcal{H}_i \leftarrow \mathcal{H}_i \cup \{S\}$
  Assemble the protocol-specific candidate set
  Randomly permute candidate order
  Record the resulting protocol-specific instance(s) and update B_i
end while

which corresponds to the Jaccard threshold

$$\tau = \frac{c^*}{2m - c^*}.$$

A proposed distractor set is resampled if its maximum Jaccard similarity to the previously accepted sets exceeds the adaptive threshold. Up to 20 proposals are considered for each distractor draw. If no proposal satisfies the diversity constraint within these attempts, the valid proposal with the smallest overlap among those considered is retained.

Algorithm 1 summarizes the shared sampling procedure underlying the controlled candidate-set construction; protocol-specific modifications are described in Appendix B.3.

B.3 Protocol-Specific Construction

The seven protocols share the controlled sampling principles described above but differ in the construction of query observations, positive candidates, distractors, and auxiliary conditions. Table 5 summarizes these protocol-specific rules. Unless otherwise noted, each instance contains four candidate images, and conditions derived from the same sampled instance share an identical candidate set and ordering.

P2. The three corruptions are applied only to the query image. Grayscale converts the query to luminance and then back to RGB. Occlusion places a black rectangle spanning 40% of the image width and height at a uniformly sampled valid location. Low-resolution degradation downsamples the query to 20% of its original width and height using bilinear interpolation and then restores it to the original resolution using nearest-neighbor interpolation.

P3. The query views are nested. The $N = 2$ condition uses the first two sampled views, the $N = 3$ condition adds a third, and the $N = 4$ condition adds a fourth, while the candidate set remains

Table 5: Protocol-specific construction rules in ARK.

Protocol	Protocol-specific construction
P1	One query image, one positive candidate, and three distractors.
P2	Each P1 instance is reused under three query-only corruptions: grayscale, occlusion, and low resolution; the candidates and their ordering remain unchanged.
P3	Uses $N \in \{2, 3, 4\}$ nested query views of the same identity. All N conditions share the same positive candidate, distractors, and candidate ordering.
P4	Each sampled query-candidate instance is evaluated under four query-side conditions: image only, image with circadian-rhythm context, image with spatial-orientation context, and image with both metadata cues.
P5	Two positive candidates and two distractors are placed among the four candidates.
P6	Instances are balanced as evenly as possible between match-present and match-absent cases. Match-present instances contain one positive and three distractors, whereas match-absent instances contain four distractors.
P7	Same-identity and different-identity pairs are sampled in approximately equal proportions. Each image pair is reused across the neutral, supportive, and misleading conditions with the images held fixed.

fixed across conditions. This construction isolates the effect of additional query observations from changes in candidate composition.

P4. Contextual metadata are attached only to the query image. Spatial orientation takes one of four values: front, back, left, or right. The circadian-rhythm cue is represented as day or night. The four conditions derived from each sampled instance share the same query image, candidates, candidate ordering, and correct answer, differing only in the contextual information provided with the query.

B.4 Image Preprocessing

All image preprocessing is performed before instance construction, and the construction pipeline operates on the resulting prepared image pool. For the WildlifeDatasets-indexed sources, ARK follows the source-specific preprocessing provided by WildlifeDatasets where applicable, including bounding-box cropping, segmentation-based cropping, and removal of black borders from segmented images. Sources without such transformations retain their full prepared images. Table 6 summarizes the preprocessing applied to each source group.

Table 6: Image preprocessing applied across ARK source groups.

Sources	Preprocessing
BelugaID, LeopardID2022, WhaleSharkID	Bounding-box crop.
NDD20	Segmentation mask followed by a tight crop.
BirdIndividualID, SealID	Crop to the non-background region of the segmented image.
CTai, Giraffes, IPanda50, LionData, NyalaData	Full image retained.
MetaWild species	Camera-overlay regions removed using fixed camera-group-specific boundary crops.

For MetaWild, camera-overlay regions are removed before benchmark construction using fixed crop specifications determined for each camera group and image resolution. The same preprocessing is applied consistently to all images within the corresponding group. ARK does not otherwise apply global resizing or normalization during benchmark construction.

C Retrieval-Derived Validation Details

ARK is designed as a controlled diagnostic benchmark, where candidate sets are constructed to isolate specific reasoning capabilities. To complement this controlled evaluation, we further examine Reasoning ReID in a two-stage retrieval–verification pipeline, where candidate shortlists are pro-

duced directly by visual retrieval rather than by ARK's benchmark construction procedure. This retrieval-derived evaluation provides system-level validation of MLLM-based identity verification under retrieval-generated candidate sets.

C.1 Retrieval Setup

We evaluate the 11 WildlifeDatasets-indexed species used in this retrieval-derived analysis. CLIP ViT-L/14@336[1] and SigLIP SO400M/14@384[2] are used as visual retrievers, with Qwen3.5-122B serving as the MLLM verifier. For each query, the images in the corresponding species-specific gallery are ranked by the retriever, and the five highest-ranked images are retained as the candidate shortlist. These retrieved candidates are passed directly to the downstream verification stage. We evaluate the resulting pipeline under both closed-set identification and open-set shortlist rejection.

C.2 Closed-Set End-to-End Verification

In the closed-set setting, the query identity is present in the corresponding gallery. We report Rank-1 accuracy and Rank-5 recall (R@5). Rank-1 measures whether the highest-ranked retrieved image has the correct identity, whereas R@5 measures whether the correct identity appears among the retrieved top-5 candidates. We compare three system configurations. **Retrieval-only** directly returns the identity of the top-ranked image. **Always-MLLM** passes the query together with the retrieved top-5 candidates to Qwen3.5-122B for final verification. **Routed-MLLM** selectively invokes the MLLM for uncertain retrievals and otherwise retains the Rank-1 retrieval prediction. The routing criterion and calibration procedure are detailed in Section C.3. All final identification accuracies are computed over the full query set. If the correct identity is absent from the retrieved top-5 candidates, it cannot be recovered by a verifier restricted to the retrieved shortlist, and the prediction is counted as incorrect. R@5 therefore provides an upper bound on the end-to-end accuracy achievable under this fixed-shortlist setting. As shown in Table 7, both retrievers exhibit a substantial gap between Rank-1 accuracy and R@5. CLIP achieves 23.13% Rank-1 accuracy and 42.44% R@5, while SigLIP achieves 24.38% and 44.27%, respectively. This gap indicates that the correct identity is frequently retrieved within the shortlist without being ranked first, leaving substantial headroom for downstream candidate verification. At the species-macro level, Always-MLLM does not convert this retrieval headroom into improved end-to-end performance. Its accuracy is 20.60% with CLIP and 22.47% with SigLIP, both below the corresponding Retrieval-only results of 23.13% and 24.38%. In contrast, Routed-MLLM reaches 24.15% and 25.50%, yielding improvements of 1.02 and 1.12 percentage points over Retrieval-only, respectively. These results support selective rather than uniform MLLM verification of retrieved candidate shortlists.

Table 7: Closed-set end-to-end retrieval-derived validation across the 11 WildlifeDatasets-indexed species. Retrieval and R@5 denote Rank-1 accuracy and Rank-5 recall, respectively; Always and Routed denote the end-to-end accuracies of Always-MLLM and Routed-MLLM. All end-to-end accuracies are computed over the full query set, with queries whose correct identity is absent from the retrieved top-5 counted as incorrect. Ave. denotes the unweighted species macro-average.

Ret.	Method	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
CLIP	Retrieval	1.23	9.60	44.26	2.54	50.25	49.76	4.00	18.87	13.87	35.55	24.43	23.13
	R@5	4.18	41.60	75.12	7.89	77.20	70.16	11.00	39.35	30.35	65.12	44.90	42.44
	Always	1.79	14.34	26.65	4.67	37.51	40.35	6.66	20.09	17.98	34.30	22.30	20.60
	Routed	1.48	13.87	43.51	4.08	49.79	49.34	6.17	20.50	16.78	35.66	24.45	24.15
SigLIP	Retrieval	2.21	14.13	45.07	4.33	56.23	51.29	7.00	17.25	13.29	33.55	23.79	24.38
	R@5	9.76	48.27	74.64	11.45	82.37	69.78	16.00	39.08	29.77	62.13	43.77	44.27
	Always	3.35	17.60	32.38	7.26	40.09	48.99	7.65	20.19	19.96	26.37	23.38	22.47
	Routed	2.70	16.79	44.42	6.41	55.50	51.15	7.59	20.74	17.81	33.26	24.11	25.50

[1] `openai/clip-vit-large-patch14-336`
[2] `google/siglip-so400m-patch14-384`

C.3 Uncertainty-Aware Routing

Routed-MLLM uses the score margin between the two highest-ranked retrieval candidates to determine whether to invoke MLLM verification. For a query q, we define the retrieval margin as

$$m(q) = s_1(q) - s_2(q), \tag{1}$$

where $s_1(q)$ and $s_2(q)$ denote the top-1 and top-2 retrieval scores, respectively. A smaller margin indicates greater ambiguity between the leading candidates.

Because margin distributions may differ across species, we normalize the raw margin separately for each species within each retrieval model. Specifically,

$$p(q) = F_s^{\mathrm{val}}(m(q)), \tag{2}$$

where F_s^{val} denotes the empirical cumulative distribution function of validation margins for species s. Thus, $p(q)$ places the retrieval margin on a species-normalized percentile scale, with smaller values indicating greater retrieval uncertainty.

A query is routed to Qwen3.5-122B when $p(q) < \rho$; otherwise, the retriever's Rank-1 prediction is used as the final prediction. The routing threshold is selected on the validation data to maximize the unweighted species-macro end-to-end accuracy across the 11 WildlifeDatasets-indexed species:

$$\rho^* = \arg\max_{\rho} \frac{1}{11} \sum_{s=1}^{11} \mathrm{Acc}_s^{\mathrm{val}}(\rho), \tag{3}$$

where $\mathrm{Acc}_s^{\mathrm{val}}(\rho)$ denotes the end-to-end validation accuracy for species s under routing threshold ρ. Calibration is performed independently for each retrieval model. Ties are resolved in favor of the smaller threshold, and the selected threshold is fixed for test evaluation.

C.4 Open-Set Shortlist Rejection

In the open-set setting, we distinguish queries according to whether the correct identity is present in the retrieved top-5 shortlist. A query is **match-present (MP)** when the correct identity appears among the five retrieved candidates and **match-absent (MA)** otherwise. For both MP and MA queries passed to the MLLM, the verifier receives the query, the five retrieved candidates, and an additional *None of the above* option. Using the same six-option format for both conditions avoids condition leakage through the prompt structure.

We report match-present accuracy (AccMP), match-absent rejection accuracy (AccMA), and balanced accuracy,

$$\mathrm{BAcc} = \frac{1}{2}\left(\mathrm{AccMP} + \mathrm{AccMA}\right). \tag{4}$$

AccMP measures whether the system selects a candidate corresponding to the query identity on MP queries, whereas AccMA measures whether the system selects *None of the above* on MA queries. The two metrics are computed conditionally on their respective query subsets before species-level BAcc is calculated.

We compare four system configurations. **Retrieval-only** always returns the Rank-1 retrieval prediction and therefore has no mechanism for rejecting an MA shortlist. **Retrieval+Threshold** augments the retriever with a calibrated rejection threshold. **Always-MLLM** sends every query and its retrieved shortlist to Qwen3.5-122B, allowing the verifier either to select one of the five candidates or to reject the shortlist. **Routed-MLLM** uses the uncertainty-aware routing criterion described in Section C.3: routed queries are passed to the MLLM, whereas non-routed queries use the calibrated retrieval decision.

For threshold-based rejection, we normalize the top-1 retrieval score relative to the corresponding match-present validation distribution:

$$a(q) = F_s^{\mathrm{val,MP}}(s_1(q)), \tag{5}$$

where $F_s^{\mathrm{val,MP}}$ denotes the empirical cumulative distribution function of top-1 scores for match-present validation queries of species s. Smaller values of $a(q)$ therefore indicate that the top-ranked

candidate has an unusually low retrieval score relative to valid matches from the same species. A query is rejected when

$$a(q) < \gamma, \tag{6}$$

and otherwise the Rank-1 prediction is used as the final prediction.

The rejection threshold γ is selected on the validation data to maximize the unweighted species-macro BAcc across the 11 WildlifeDatasets-indexed species. Calibration is performed independently for each retrieval model, and the selected threshold is fixed for test evaluation. Routed-MLLM reuses the routing criterion calibrated in Section C.3; for queries not routed to the MLLM, the same calibrated rejection rule is applied.

Table 8: Open-set shortlist rejection results across the 11 WildlifeDatasets-indexed species. Retrieval and +Threshold denote retrieval without and with calibrated threshold rejection, respectively; Always and Routed denote Always-MLLM and Routed-MLLM. AccMP and AccMA denote accuracy conditional on match-present and match-absent queries, respectively, and BAcc is their arithmetic mean. Ave. denotes the unweighted species macro-average.

Ret.	Method	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
AccMP													
CLIP	Retrieval	29.43	23.08	58.92	32.19	65.09	70.92	36.36	47.95	45.70	54.59	54.41	47.15
	+Threshold	26.81	22.16	52.07	30.04	56.88	67.18	31.65	44.95	40.25	52.04	47.83	42.90
	Always	33.36	27.80	26.71	50.86	40.17	38.03	58.17	39.10	57.11	43.89	46.11	41.94
	Routed	33.94	29.45	49.33	48.19	63.05	70.25	48.28	46.63	44.81	49.96	51.03	48.63
SigLIP	Retrieval	22.64	29.27	60.38	37.82	68.27	73.50	43.75	44.14	44.64	54.00	54.35	48.43
	+Threshold	19.89	27.83	55.24	35.29	60.57	70.06	37.90	41.08	39.86	51.25	48.28	44.30
	Always	27.02	29.21	32.78	54.07	40.23	46.46	45.41	40.85	63.52	35.97	50.65	42.38
	Routed	26.78	29.35	54.66	53.31	63.80	70.85	47.26	43.09	51.10	49.84	53.67	49.43
AccMA													
CLIP	Retrieval	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	+Threshold	35.81	53.72	42.76	46.35	48.93	57.58	33.51	38.82	46.21	39.34	52.27	45.03
	Always	10.23	21.66	6.96	10.28	22.77	20.75	11.54	8.97	21.09	10.52	19.73	14.95
	Routed	40.14	62.83	44.95	49.06	48.37	56.08	35.41	43.65	51.44	40.47	55.60	48.00
SigLIP	Retrieval	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	+Threshold	37.72	55.31	42.12	48.06	52.24	55.35	35.62	37.41	44.58	43.31	50.74	45.68
	Always	8.95	22.84	5.88	12.47	20.91	21.62	13.05	8.22	22.10	9.17	18.69	14.90
	Routed	42.47	60.05	46.58	50.97	53.96	54.21	40.24	43.13	47.94	48.35	54.20	49.28
BAcc													
CLIP	Retrieval	14.72	11.54	29.46	16.10	32.55	35.46	18.18	23.98	22.85	27.30	27.21	23.58
	+Threshold	31.31	37.94	47.42	38.20	52.91	62.38	32.58	41.89	43.23	45.69	50.05	43.96
	Always	21.80	24.73	16.84	30.57	31.47	29.39	34.86	24.04	39.10	27.21	32.92	28.45
	Routed	37.04	46.14	47.14	48.63	55.71	63.17	41.85	45.14	48.13	45.22	53.32	48.32
SigLIP	Retrieval	11.32	14.64	30.19	18.91	34.14	36.75	21.88	22.07	22.32	27.00	27.18	24.22
	+Threshold	28.81	41.57	48.68	41.68	56.41	62.71	36.76	39.25	42.22	47.28	49.51	44.99
	Always	17.99	26.03	19.33	33.27	30.57	34.04	29.23	24.54	42.81	22.57	34.67	28.64
	Routed	34.63	44.70	50.62	52.14	58.88	62.53	43.75	43.11	49.52	49.10	53.94	49.36

As shown in Table 8, Retrieval-only cannot reject match-absent shortlists and therefore obtains zero AccMA, resulting in species-macro BAcc of 23.58% for CLIP and 24.22% for SigLIP. Always-MLLM enables direct shortlist rejection without an explicit retrieval-score threshold and increases BAcc to 28.45% and 28.64%, respectively, although its AccMA remains relatively low. Calibrated threshold rejection provides a substantially stronger open-set baseline, reaching BAcc of 43.96% with CLIP and 44.99% with SigLIP. At the selected operating point, this improvement is accompanied by lower AccMP than Retrieval-only: 42.90% versus 47.15% for CLIP and 44.30% versus 48.43% for SigLIP, reflecting an acceptance–rejection trade-off. Routed-MLLM attains the highest species-macro BAcc for both retrievers. Relative to Retrieval+Threshold, it increases AccMP from 42.90% to 48.63% and AccMA from 45.03% to 48.00% with CLIP, yielding a BAcc of 48.32%. With SigLIP, AccMP increases from 44.30% to 49.43% and AccMA from 45.68% to 49.28%, yield-

ing a BAcc of 49.36%. These results indicate that selective MLLM verification can complement calibrated retrieval-based rejection under retrieval-generated candidate shortlists.

C.5 Inference-Time Cost

Table 9: Per-query latency of the two-stage retrieval–verification pipeline measured on an NVIDIA H200 GPU. Retrieval latency includes query encoding and top-5 search, while verification latency corresponds to Qwen3.5-122B processing one query and five retrieved candidates.

Stage	Model / Setting	Operation	Latency / Query
Retrieval	CLIP / SigLIP	Query encoding + top-5 search	5–8 ms
Verification	Qwen3.5-122B	Query + five-candidate verification	$\sim$40 s

Table 9 shows a substantial latency difference between the two stages of the retrieval–verification pipeline. CLIP/SigLIP retrieval requires 5–8 ms per query, whereas Qwen3.5-122B verification of the query and five retrieved candidates requires approximately 40 s. Of this verification latency, image processing and prompt prefill account for approximately 0.13 s, with the overall latency dominated by autoregressive generation. The two stages therefore operate at substantially different computational scales: retrieval performs efficient gallery-scale search, whereas the MLLM performs more expensive reasoning over a fixed candidate shortlist. This cost difference further motivates the selective verification strategy in Routed-MLLM, which invokes the MLLM only for retrievals identified as uncertain rather than for every query.

D Evaluation Metrics and Aggregation

Evaluation Metrics. For P1–P4, we report accuracy. For P2, the species-level score is the arithmetic mean of the accuracies under grayscale, occlusion, and low-resolution corruption. For P5, precision and recall are computed at the instance level and then averaged within each species. Species-level F1 is computed as the harmonic mean of the resulting precision and recall, while Exact Match (EM) is the proportion of instances for which the predicted match set exactly equals the ground-truth set. For P6, we separately report Match-Present Accuracy (AccMP) and Match-Absent Accuracy (AccMA) on match-present and match-absent instances. For P7, we report Neutral Accuracy (AccN), Supportive Accuracy (AccS), and Misleading Accuracy (AccM) under the neutral, supportive, and misleading prompt conditions, respectively.

Aggregation. Metrics are computed separately for each species before protocol-level aggregation. The primary aggregate reported throughout the paper is the unweighted macro-average across the species included in the corresponding protocol, giving each species equal weight regardless of its number of instances. P1–P3 and P5–P7 are aggregated over the 11 WildlifeDatasets-indexed species, whereas P4 is aggregated over the six MetaWild species that provide the required contextual metadata. For P5, precision, recall, F1, and EM are each macro-averaged independently across species.

E Human Reference Evaluation

E.1 Study Design, Sampling, and Procedure

We report a human reference based on evaluations by six non-expert members of our research laboratory. Because animal-identification expertise is often species- and population-specific, a comprehensive expert reference across all ARK species would require multiple specialist groups. We therefore used the same non-expert participant pool to provide a consistent human reference across protocols and species.

The evaluation covered all seven ARK protocols using protocol-specific, species-balanced sampling. For P1, P3, and P5, we sampled 18 benchmark instances per species across the 11 WildlifeDatasets-indexed species. P2 used 18 instances per species for each of its three corruption conditions, while P6 used 18 instances per species for each of its two match-presence conditions. For P4, we sampled 24 underlying instances per species across the six MetaWild species and evaluated the Image and

Full prompt conditions on the same sampled instances. For P7, we sampled 18 image-pair instances per species and evaluated each pair under Neutral, Supportive, and Misleading prompts. For P3, the $N = 2$ evaluation was additionally paired with a single-view control for each sampled question. These control prompts were used only for the human evaluation and were not counted as additional ARK benchmark instances. When multiple prompt conditions were derived from the same underlying instance, they were assigned to different participants to reduce carryover between conditions. Assignments were balanced across participants, species, and conditions.

Under ARK's benchmark-instance definition, the human-reference evaluation comprises 1,926 scored instances: 198 for P1, 594 for P2, 198 for P3 at $N = 2$, 144 for P4, 198 for P5, 396 for P6, and 198 for P7. We additionally sampled 22 P1 questions, two from each WildlifeDatasets-indexed species, which were answered by all six participants and used exclusively to measure inter-annotator agreement. The complete human-evaluated set therefore contains 1,948 unique ARK instances, of which 1,926 contribute to the reported human reference and 22 are used exclusively for inter-annotator agreement.

Each participant completed 466 prompts: 411 prompts corresponding to the scored benchmark conditions, 33 P3 single-view control prompts, and 22 shared agreement questions. Across the six participants, this yielded 2,796 human responses in total.

Participants completed the evaluation without a time limit and could revisit the presented images. Prompt order was independently randomized for each participant to reduce systematic ordering effects across protocols, species, and conditions. Human performance was evaluated using the same protocol-specific metrics and species-balanced aggregation as the model evaluation.

E.2 Agreement and Uncertainty

Inter-Annotator Agreement. To assess consistency across participants, we used the 22 additional P1 instances answered independently by all six participants. Because P1 is a four-option single-choice task, inter-annotator agreement was computed directly from the categorical option selections using Fleiss' κ. Agreement was $\kappa = 0.73$, with a 95% bootstrap confidence interval of $[0.59, 0.86]$. These shared instances were used exclusively for agreement analysis and did not contribute to the reported human-reference scores.

Uncertainty of the Human Reference. We estimate uncertainty in the human-reference results using species-stratified cluster bootstrap resampling. Within each protocol and species, sampling is performed with replacement using the underlying benchmark instance as the resampling unit. When multiple evaluated conditions are derived from the same underlying instance, their corresponding responses are retained together within each bootstrap replicate. For P2, the grayscale, occlusion, and low-resolution variants derived from the same base instance are therefore resampled as a single cluster. For each bootstrap replicate, the protocol-specific metric is recomputed separately for each species and then aggregated using the same unweighted species macro-average as in the main evaluation. The 2.5th and 97.5th percentiles of the bootstrap distribution define the reported 95% confidence intervals, which are shown in Table 10.

Because the scored human-reference subset was distributed across participants rather than independently annotated by all six participants, these confidence intervals primarily reflect uncertainty associated with the sampled benchmark instances. Inter-annotator agreement is assessed separately using the fully shared P1 subset described above.

F Additional Sensitivity and Robustness Analyses

F.1 Output Parsing and Recovery Sensitivity

Model responses may occasionally fail to follow the expected answer format, making the initial output unparsable under the deterministic evaluation rules. Because ARK retries only when a response cannot be parsed, we examine the extent to which this recovery procedure affects the reported benchmark scores. This analysis quantifies the sensitivity of evaluation results to output-format recovery rather than to the benchmark task itself.

Table 10: Human-reference results with 95% bootstrap confidence intervals.

	P1	P2	P3	P4		P5
	Acc	Acc	Acc	Image	Full	EM
Human	66.67	53.03	71.21	77.78	78.47	67.17
95% CI	[60.1,72.7]	[46.8,59.3]	[65.2,77.3]	[70.8,84.0]	[71.5,84.7]	[60.6,73.2]

	P6		P7	
	AccMP	AccMA	AccN	AccM
Human	59.60	82.83	87.88	87.37
95% CI	[53.0,66.2]	[77.8,87.9]	[83.3,91.9]	[82.8,91.4]

We audit P1 for the 11 MLLMs reported in Table 1. For each instance, the deterministic parser is first applied to the initial response, and a retry is triggered only if that response cannot be parsed. Each retry uses the same task prompt and decoding configuration, evaluation stops at the first parseable response, and an instance is counted as incorrect if all ten attempts remain unparsable. No semantic fallback parser or correctness-based retry is used. We report first-response parse validity, cumulative validity within five and ten attempts, the mean number of retries, strict first-response accuracy, and accuracy under the standard ARK recovery procedure.

Table 11: Output parsing and recovery sensitivity on P1 for the 11 MLLMs reported in Table 1. Valid@1 denotes the percentage of instances whose initial response is successfully parsed, while Valid$\leq$5 and Valid$\leq$10 denote the cumulative percentage with a parseable response within five and ten attempts, respectively. Mean retries excludes the initial generation. Strict-first Acc. scores only the initial response and counts an unparsable response as incorrect, whereas Main Acc. follows the standard ARK recovery procedure. Both accuracy columns are unweighted macro-averages across the 11 P1 species. Δ_{rec} denotes Main Acc. minus Strict-first Acc. in percentage points.

Model	**Valid@1**	**Valid$\leq$5**	**Valid$\leq$10**	**Mean retries**	**Strict-first Acc.**	**Main Acc.**	$\boldsymbol{\Delta_{\mathrm{rec}}}$
LLaVA-13B	95.18	99.82	100.00	0.08	23.80	25.00	+1.20
Gemma3-27B	93.41	99.63	99.99	0.13	23.18	24.83	+1.65
Qwen3-VL-2B	88.70	99.45	99.99	0.20	24.86	27.75	+2.89
Qwen3-VL-30B	77.60	97.42	99.94	0.48	29.49	36.96	+7.47
Qwen3-VL-235B	93.20	99.74	100.00	0.12	39.99	42.75	+2.76
Qwen3.5-0.8B	80.36	98.72	99.97	0.39	19.73	24.73	+5.00
Qwen3.5-35B	73.24	96.58	99.88	0.65	35.12	46.67	+11.55
Qwen3.5-122B	90.27	99.51	99.99	0.18	47.99	52.82	+4.83
Gemini-3.1-Pro	99.96	100.00	100.00	0.00	48.00	48.02	+0.02
GPT-5.4	99.99	100.00	100.00	0.00	43.64	43.64	+0.00
Claude-Opus-4.6	99.99	100.00	100.00	0.00	53.48	53.49	+0.01

As shown in Table 11, the recovery procedure resolves nearly all initially unparsable outputs within the allowed retry budget. First-response validity varies substantially across models, from 73.24% for Qwen3.5-35B to 99.96–99.99% for the three proprietary models, while cumulative validity reaches at least 96.58% within five attempts and 99.88% within ten attempts for every model. The resulting change in accuracy is also strongly model dependent: Qwen3.5-35B, Qwen3-VL-30B, and Qwen3.5-0.8B gain 11.55, 7.47, and 5.00 percentage points, respectively, whereas the three proprietary models change by at most 0.02 points.

The within-family scaling trends remain evident under strict first-response scoring. Qwen3-VL accuracy increases from 24.86% for 2B to 29.49% for 30B and 39.99% for 235B, while Qwen3.5 increases from 19.73% for 0.8B to 35.12% for 35B and 47.99% for 122B. Claude-Opus-4.6 also remains the strongest model under strict first-response scoring at 53.48%. These results show that recovery can materially affect the absolute scores of models with lower first-response parse validity, but does not account for the observed within-family scaling patterns.

F.2 Statistical Robustness and Identity-Level Sensitivity

Although ARK reports species-level macro-averaged results, individual identities can contribute different numbers of evaluation instances within a species, giving identities represented by more instances greater influence on the species-level score. We therefore assess the robustness of P1 results to identity weighting and composition in two ways. First, we compute an identity-aware macro-average by first computing accuracy separately for each query identity, averaging equally across identities within each species, and then averaging equally across the 11 P1 species, thereby preserving the species weighting used by the main metric. Second, we perform a species-stratified identity bootstrap, resampling query identities with replacement within each species while retaining all instances associated with each sampled identity, and report 95% bootstrap intervals over the resulting P1 macro-averages.

Table 12: Identity-level aggregation and bootstrap robustness on P1 for the 11 MLLMs in Table 1. Species Macro denotes the standard P1 accuracy, computed as an unweighted macro-average across the 11 species. Identity-aware Macro first averages accuracy equally across identities within each species and then equally across species. Δ_{id} denotes Identity-aware Macro minus Species Macro in percentage points. The 95% ID-Bootstrap Interval is obtained by species-stratified resampling of query identities with replacement while retaining all instances associated with each sampled identity, and spans the 2.5th–97.5th percentiles of the resulting standard P1 macro-averages.

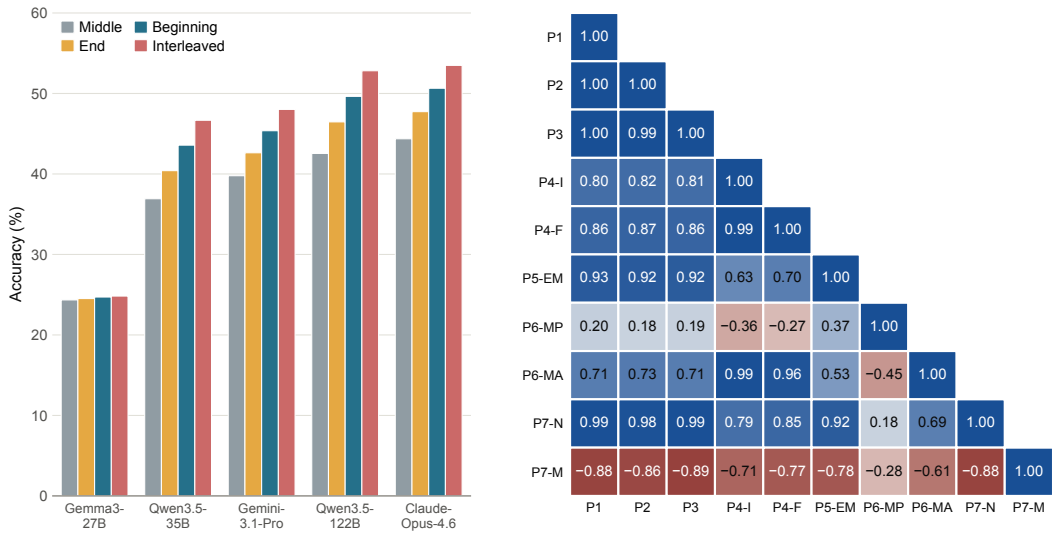
Model	Species Macro	Identity-aware Macro	Δ_{id}	95% ID-Bootstrap Interval
LLaVA-13B	25.00	24.98	−0.02	[23.92, 26.11]
Gemma3-27B	24.83	24.85	+0.02	[23.75, 25.94]
Qwen3-VL-2B	27.75	27.69	−0.06	[26.53, 29.01]
Qwen3-VL-30B	36.96	36.88	−0.08	[35.59, 38.35]
Qwen3-VL-235B	42.75	42.65	−0.10	[41.32, 44.19]
Qwen3.5-0.8B	24.73	24.68	−0.05	[23.65, 25.84]
Qwen3.5-35B	46.67	46.57	−0.10	[45.22, 48.12]
Qwen3.5-122B	52.82	52.69	−0.13	[51.34, 54.28]
Gemini-3.1-Pro	48.02	47.94	−0.08	[46.55, 49.48]
GPT-5.4	43.64	43.68	+0.04	[42.19, 45.11]
Claude-Opus-4.6	53.49	53.37	−0.12	[52.01, 54.96]

As shown in Table 12, identity-aware reweighting produces only minor changes in P1 accuracy. Across all 11 MLLMs, the absolute difference from the standard species-level macro-average is at most 0.13 percentage points, and the point-estimate ordering of the models remains unchanged. The within-family scaling trends are also preserved: identity-aware accuracy increases from 27.69% to 36.88% and 42.65% across the three Qwen3-VL models, and from 24.68% to 46.57% and 52.69% across the three Qwen3.5 models. These results indicate that the reported P1 performance is not materially affected by unequal numbers of evaluation instances across identities within each species.

The identity-level bootstrap shows similarly stable aggregate performance under changes in identity composition. Across the 11 models, the 95% bootstrap intervals have widths between 2.19 and 2.95 percentage points and remain centered near the corresponding species-level macro-averages. This indicates that the aggregate P1 scores are relatively stable to changes in identity composition under the bootstrap resampling procedure.

F.3 Image-Position and Input-Arrangement Sensitivity

Because P1 presents a query image together with four candidate images, model performance may depend on how these visual inputs are arranged relative to the accompanying text. We therefore evaluate four prompt arrangements on the same P1 instances: **All-in-Beginning**, with all images placed before the task instruction; **All-in-Middle**, with the query and candidate images embedded within the instruction; **All-in-End**, with the task text preceding the query and candidate images; and **Interleaved**, with the query image presented in the question and each candidate image paired with its corresponding option label. Across conditions, the evaluated images, candidate order, answer labels, decoding configuration, and evaluation procedure are held fixed, while the textual wording is adjusted only as necessary to realize each arrangement.



(a) Input-arrangement sensitivity.

(b) Cross-protocol performance correlations.

Figure 6: Additional analyses of evaluation sensitivity and cross-protocol relationships. (a) P1 accuracy under four image–text input arrangements, evaluated with the same instances, candidate order, answer labels, decoding configuration, and evaluation procedure. (b) Pairwise Pearson correlations across 21 MLLMs using protocol-level aggregate scores.

As shown in Figure 6a, ReID accuracy varies across the four input arrangements, although the magnitude of this sensitivity differs across models. The **Interleaved** condition, which corresponds to the standard prompt in the main ARK evaluation, yields the strongest overall performance, whereas **All-in-Middle** generally produces the lowest accuracy. **All-in-Beginning** and **All-in-End** fall between these two conditions, without consistently outperforming the interleaved format. These results identify input arrangement as a source of evaluation sensitivity in multi-image ReID prompting. In particular, explicitly pairing each candidate image with its corresponding option label is associated with higher ReID accuracy than embedding the candidate images collectively within the instruction.

All-in-Beginning

```
Question: <image_query> <image_A> <image_B> <image_C> <image_D>
The first image is the query image, and the remaining four images
correspond to options A, B, C, and D, respectively. Please retrieve
the same individual from the following options. Which option shows
the same individual as the query?
Choices:
(A) Candidate A
(B) Candidate B
(C) Candidate C
(D) Candidate D
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Answer:
```

All-in-Middle

```
Question: Here is the query image: <image_query>. Please retrieve
the same individual from the following options: A <image_A>, B
<image_B>, C <image_C>, and D <image_D>. Which option shows the same
individual as the query?
Choices:
(A) Candidate A
(B) Candidate B
(C) Candidate C
(D) Candidate D
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Answer:
```

All-in-End

```
Question: Please retrieve the same individual as the query image from
the following options. Which option shows the same individual as the
query?
Choices:
(A) Candidate A
(B) Candidate B
(C) Candidate C
(D) Candidate D
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Query image: <image_query>
Candidate images:
(A) <image_A>
(B) <image_B>
(C) <image_C>
(D) <image_D>
Answer:
```

Interleaved

```
Question: Here is the query image: <image_query>. Please retrieve
the same individual from the following options. Which option shows
the same individual as the query?
Choices:
(A) <image_A>
(B) <image_B>
(C) <image_C>
(D) <image_D>
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Answer:
```

Prompt arrangements. The following prompt templates show the four input arrangements used in this analysis. Image placeholders denote the corresponding visual inputs. The **Interleaved** condition corresponds to the standard P1 prompt used in the main evaluation.

F.4 Cross-Protocol Performance Correlations

To characterize model-level relationships across ARK evaluation settings, we compute pairwise Pearson correlations across all 21 evaluated MLLMs using the corresponding protocol-level aggregate scores. The analysis includes P1 accuracy, P2 accuracy, P3 accuracy, P4 image-only and full-context accuracy, P5 Exact Match, P6 match-present and match-absent accuracy, and P7 neutral and misleading accuracy. This analysis provides a compact view of how performance patterns align or diverge across the benchmark settings.

Figure 6b shows distinct correlation patterns across ARK protocols. P1–P3 are almost perfectly correlated ($r = 0.99$–1.00), and P7 under neutral suggestions (P7-N) is similarly aligned with these protocols ($r = 0.98$–0.99). P5 Exact Match also shows strong correlations with P1–P3 ($r = 0.92$–0.93). These relationships indicate that models performing well on standard fine-grained identity verification also tend to perform well on corruption-robust, multi-view, multi-match, and neutral-suggestion settings. P4 image-only and full-context performance are likewise strongly correlated ($r = 0.99$), indicating that the addition of contextual metadata largely preserves the relative ordering of models. In contrast, the decision-reliability conditions exhibit markedly different relationships. P6 match-present accuracy is only weakly correlated with P1–P3 ($r = 0.18$–0.20) and is negatively correlated with match-absent accuracy ($r = -0.45$), suggesting a trade-off between identifying an available match and correctly rejecting a candidate set containing no match. More strikingly, P7 performance under misleading suggestions (P7-M) is negatively correlated with every other reported condition, including P1–P3 ($r = -0.86$ to -0.89), P5 Exact Match ($r = -0.78$), and P7-N ($r = -0.88$). Notably, models with higher accuracy under ordinary identity-verification conditions tend to achieve lower accuracy under misleading suggestions. This inverse relationship is descriptive rather than causal, but it highlights that resistance to misleading suggestions captures a distinct aspect of decision reliability. Together, these results support evaluating decision reliability separately rather than treating general ReID accuracy as a sufficient proxy for it.

G Reproducibility and Benchmark Reconstruction

G.1 Image-Free Release and Source Reconstruction

ARK adopts an image-free release design to respect the access and redistribution conditions of its upstream datasets. Rather than redistributing third-party images, ARK releases benchmark annotations and source mappings that associate each benchmark observation with its upstream image, together with the preprocessing metadata and code needed to reconstruct the corresponding benchmark input. Source images are obtained from their original upstream locations under the access conditions documented in Appendix A.2. Most sources can be acquired through the provided data pipeline; when an upstream host requires manual authorization or temporarily restricts automated access, the pipeline identifies the required source file and its expected local location.

Figure 7 summarizes the resulting image-free workflow. Starting from an upstream source image, ARK resolves the released source mapping and applies the recorded image transformations, including cropping, masking when applicable, and removal of camera or photographer overlays. The reconstructed image is then checked against the released fingerprint before being used to restore the corresponding benchmark instances. This design separates the ARK-authored reasoning and evaluation layer from third-party image distribution while retaining a reproducible path from upstream source data to the benchmark inputs used for evaluation. The reconstruction and verification procedures are detailed in Appendix G.2.

G.2 Deterministic Preprocessing and Verification

ARK reconstructs each benchmark image from its upstream source according to the preprocessing specification stored in the released observation metadata. For each observation, the specification records its source image, required image form, spatial transformations such as cropping and masking when applicable, camera- or photographer-overlay removal, and expected output dimensions. These transformations are applied deterministically, with transformed images stored losslessly to avoid introducing additional compression artifacts.

Reconstruction correctness is checked against fingerprints released with the benchmark metadata. Unchanged source images are checked using file-level SHA-256 fingerprints when available, whereas reconstructed images are verified using SHA-256 fingerprints computed over their decoded pixel content. Query corruptions generated for P2 are likewise verified against their released pixel fingerprints. Benchmark-instance records are additionally compared with the released instances, providing a separate consistency check between the reconstructed observations and the protocol-level benchmark specification. Fingerprint mismatches are treated as reconstruction failures rather than silently accepted. When an image-level mismatch occurs, perceptual hashing is used only as a diagnostic signal to distinguish likely decoding differences from substantively different image

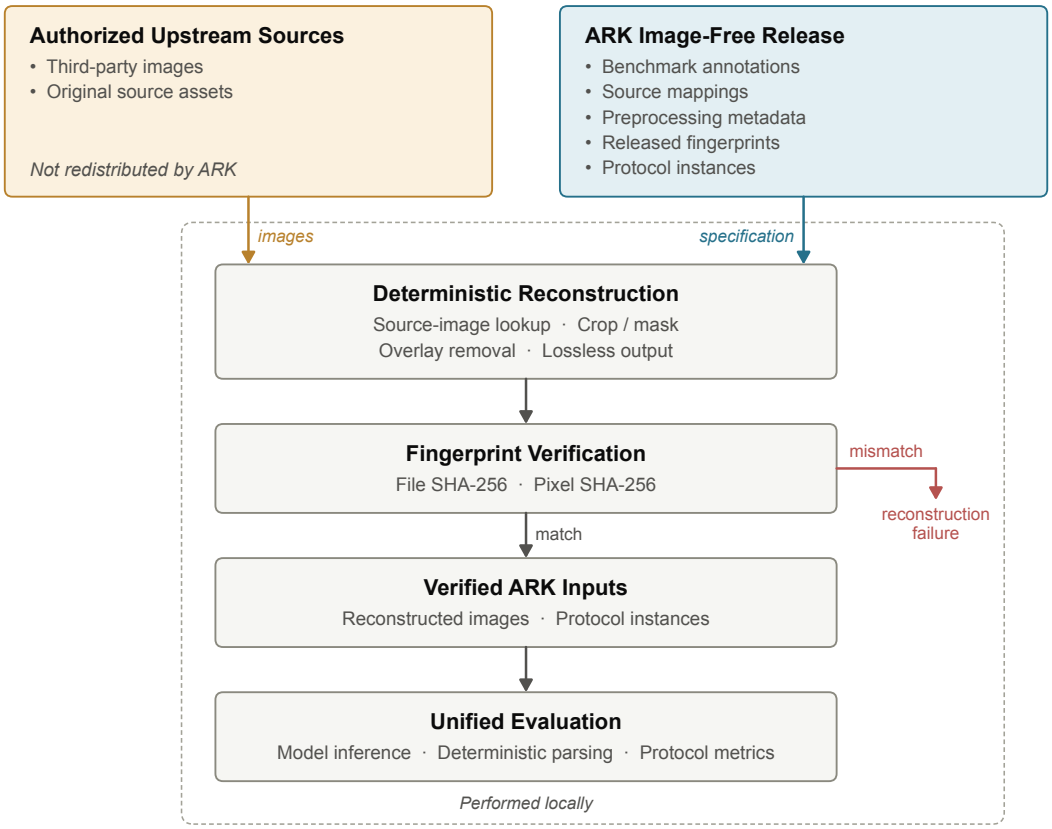


Figure 7: **Image-free reconstruction and evaluation workflow of ARK.** Third-party images are obtained from their upstream sources under the access conditions rather than redistributed by ARK. Released source mappings, preprocessing metadata, fingerprints, and protocol instances support deterministic reconstruction and verification of benchmark inputs before unified evaluation.

content; it does not replace the exact fingerprint checks. Algorithm 2 summarizes the complete reconstruction and verification procedure.

G.3 Unified Evaluation Workflow

ARK provides a unified evaluation interface that operates on the reconstructed benchmark instances. For each instance, the evaluator assembles the released prompt and associated visual inputs according to a recorded image policy, invokes the model, and applies the protocol-specific deterministic parser to its response.

If a response cannot be parsed, the same instance is queried again, up to a maximum of ten attempts. Evaluation stops at the first parseable response, which is used as the prediction; an instance for which all ten responses are unparsable is scored as incorrect. Every raw response and its parsed result are retained, together with the parser version and model-run configuration, allowing predictions to be audited or reparsed without repeating model inference.

Evaluation results are written incrementally at the instance level. Completed instances are identified by their task identifiers and skipped when a run is resumed, allowing interrupted evaluations to continue without recomputing completed model calls. ARK then applies the protocol-specific metrics to the resulting predictions using the same scoring implementation across model backends.

Algorithm 2 Reconstructing and verifying an ARK benchmark unit.

Require: ARK release R, authorized upstream source assets S, benchmark unit u
Ensure: Verified local benchmark inputs and instances for u

```
Load the released observation manifest and protocol instances for u
for each released observation o do
    Locate its upstream source image in S using the released source mapping
    if o is unchanged then
        Retain the source image and verify its file-level SHA-256 fingerprint
    else
        Apply the released crop and mask specifications when applicable
        Remove the recorded camera or photographer overlay rows when specified
        Save losslessly and verify dimensions and decoded-pixel SHA-256
    end if
    if verification fails then
        Report the mismatch; use the perceptual hash only for diagnostic characterization
    end if
end for
Restore the released protocol-instance records for u
if P2 is present then
    Regenerate corrupted queries from the verified observation pool
    Verify the regenerated queries against their released pixel fingerprints
end if
Verify that the local instance records match the released records
return the verified benchmark unit only if all required checks pass
```

G.4 Reproducibility Safeguards

ARK records provenance at both the benchmark and evaluation levels. Each release includes a manifest containing benchmark counts, software versions used to construct the released fingerprints, and SHA-256 fingerprints of the release files. Local benchmark units retain fingerprints linking them to the release from which they were reconstructed, allowing subsequent runs to verify that their inputs correspond to the intended benchmark version.

Evaluation provenance is recorded alongside model outputs. Run records preserve the model specification and evaluation configuration, while individual results retain raw responses, parsed predictions, and parser versions. When an interrupted evaluation is resumed, ARK prevents results produced under incompatible model settings or input policies from being silently combined under the same model run.

Together, these safeguards make mismatches explicit rather than silently propagating them through evaluation. Release and input integrity are checked against stored fingerprints, while evaluation configurations and parser versions are retained for audit. Raw model responses further allow outputs to be reparsed without repeating model inference.

H Protocol Prompts and Benchmark Examples

H.1 Exact Prompt Templates

The following boxes report the prompt templates used in ARK. Image placeholders indicate the positions at which visual inputs are supplied, while condition-specific placeholders are instantiated when each benchmark prompt is rendered.

P1: Fine-Grained Identity Verification.

P1: Fine-Grained Identity Verification

```
Question: Here is the query image: <image>. Please retrieve the
same individual from the following options. Which option shows the
same individual as the query?
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Answer:
```

P2: Corruption-Robust Identity Verification.

P2 uses the same candidate-verification format as P1, with an additional sentence specifying the corruption applied to the query image. The placeholder `<corruption>` is instantiated as `grayscale`, `occlusion`, or `low_resolution`.

P2: Corruption-Robust Identity Verification

```
Question: Here is the query image: <image>. Note that the query
image is subject to <corruption> corruption. Please retrieve the same
individual from the following options. Which option shows the same
individual as the query?
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Answer:
```

P3: Multi-View Evidence Integration.

P3 uses $N \in \{2, 3, 4\}$ query images. The three prompt forms differ only in the number of query-image placeholders; the candidate set and answer format remain unchanged.

P3: Multi-View Evidence Integration

```
N=2:
Question: Here are the query images: <image>, <image>. Please
retrieve the same individual from the following options. Which option
shows the same individual as the queries?
N=3:
Question: Here are the query images: <image>, <image>, <image>.
Please retrieve the same individual from the following options. Which
option shows the same individual as the queries?
N=4:
Question: Here are the query images: <image>, <image>, <image>,
<image>. Please retrieve the same individual from the following
options. Which option shows the same individual as the queries?
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Answer:
```

P4: Context-Assisted Identity Verification.

P4 uses four query-side prompt conditions: **Image**, **Img+Ori**, **Img+CR**, and **Full**. The placeholder `<metadata>` denotes the condition-specific metadata text inserted after the query image.

P4: Context-Assisted Identity Verification

```
Question: Here is the query image: <image>. <metadata>Please
retrieve the same individual from the following options. Which option
shows the same individual as the query?
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D,
directly.
Answer:
```

P4: Condition-Specific Metadata

```
Image:
<metadata> is empty.
Img+Ori:
Note that the query image has the following metadata: Face direction:
<face_direction>. Please carefully use these metadata constraints to
help your logical reasoning.
Img+CR:
Note that the query image has the following metadata: Day night:
<day_night>. Please carefully use these metadata constraints to help
your logical reasoning.
Full:
Note that the query image has the following metadata: Face direction:
<face_direction>, Day night: <day_night>. Please carefully use these
metadata constraints to help your logical reasoning.
```

P5: Multi-Match Identity Association.

P5 allows multiple candidates to match the query identity. The prompt indicates that multiple correct options may exist without revealing the number of matches.

P5: Multi-Match Identity Association

```
Question: Here is the query image: <image>. Please retrieve all
individuals that are the same as the query from the following options.
Note that there may be multiple correct options showing the same
individual as the query. Which options show the same individual as
the query?
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letters separated by commas,
such as A, C, directly.
Answer:
```

P6: Open-Set Rejection.

P6 augments the four image candidates with a fixed no-match option, **(E) None of the above**. The same prompt format is used for both match-present and match-absent instances.

P6: Open-Set Rejection

```
Question: Here is the query image: <image>. Please retrieve the
same individual from the following options. Which option shows the
same individual as the query?
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
(E) None of the above
Hint: Please provide the correct option letter, such as A, B, C, D,
E, directly.
Answer:
```

P7: Resistance to Misleading Suggestions.

P7 presents the same image pair under **Neutral**, **Supportive**, and **Misleading** prompt conditions. The supportive suggestion states the ground-truth identity relation, whereas the misleading suggestion states the opposite relation.

P7: Resistance to Misleading Suggestions

```
Question: Image A: <image>. Image B: <image>.
<instruction>
Hint: Please answer Yes or No directly.
Answer:
```

P7: Condition-Specific Instructions

```
Neutral:
Are these two images of the same individual?
Supportive:
These two images have been preliminarily verified by experts as
belonging to <claim>.  Please confirm this verification.  Are these
two images of the same individual?
Misleading:
These two images have been preliminarily verified by experts as
belonging to <claim>.  Please confirm this verification.  Are these
two images of the same individual?
```

For the **Supportive** condition, `<claim>` is instantiated as `the same individual` for same-identity pairs and `different individuals` for different-identity pairs. For the **Misleading** condition, these assignments are reversed.

H.2 Representative Benchmark Examples

The following cards show representative rendered instances from the seven ARK protocols. Each example preserves the query observations, candidate ordering, protocol condition, and ground-truth answer of the corresponding benchmark instance.

P1: Fine-Grained Identity Verification

Query



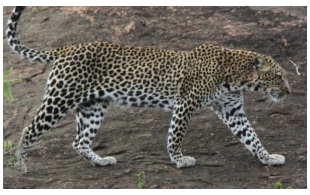
A



B



C



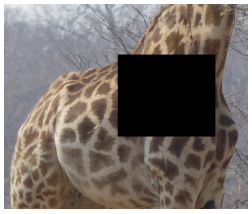
D



Answer: B

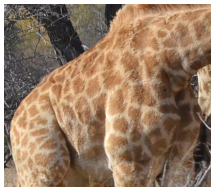
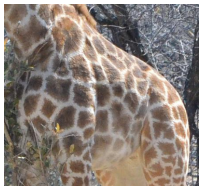
P2: Corruption-Robust Identity Verification

Grayscale **Occlusion** **Low Resolution**


A **B** **C** **D**

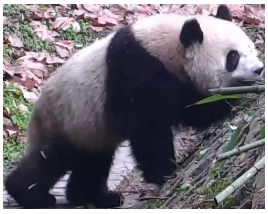
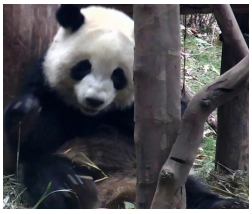


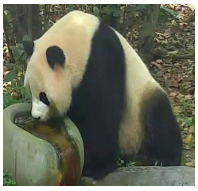
Answer: A

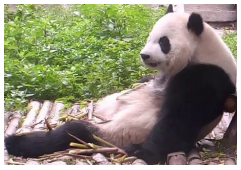
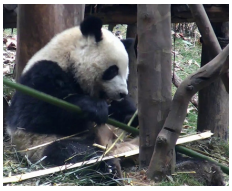
P3: Multi-View Evidence Integration

Query 1 **Query 2** **Query 3**


A **B** **C** **D**

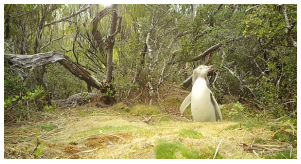


Condition: $N = 3$ **Answer: D**

P4: Context-Assisted Identity Verification

Query



Face direction: Left
Day night: Day

A



B



C



D



Condition: Full

Answer: D

P5: Multi-Match Identity Association

Query



A



B



C



D



Answer: B, D

P6: Open-Set Rejection

Query



A



B



C



D



(E) None of the above

Condition: Match absent

Answer: E

P7: Resistance to Misleading Suggestions

Image A

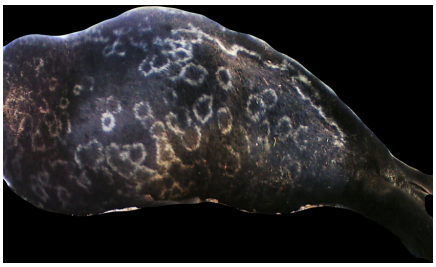


Image B

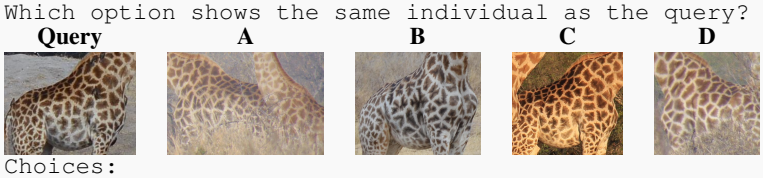


Condition: Misleading
Suggestion: different individuals
Ground-truth relation: same individual
Answer: Yes

I SUCCESS AND FAILURE EXAMPLES

P1: Fine-Grained Identity Verification Failure Case

Question: Here is the query image: <image>.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?



Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query shows the shoulder and flank of a giraffe from the side, photographed in neutral daylight against bare, grey ground. The coat has medium-to-dark brown patches separated by a broad, pale cream network, and the patches become lighter towards the belly.
Step 2: Evaluate the Candidates.
Option A: The giraffe is partly hidden behind dry grass and a second giraffe, and the image is hazy and low in contrast, so its coat cannot be reliably compared with the query.
Option B: Same side-on view of the shoulder under similar neutral light and dry-season surroundings. The patches are the same medium-to-dark brown on a wide, pale network, and the patch density across the shoulder closely resembles the query.
Option C: The coat is a strong orange-tan, with reddish-brown patches and a yellowish network, against green vegetation. This warm colouration differs markedly from the cooler brown-and-cream coat of the query.
Option D: The image is washed out and partly obscured by branches; the patches appear pale brown and diffuse, lighter than those of the query.
Step 3: Conclusion.
Options A and D are too degraded or dissimilar to support a match, and Option C's orange colouration indicates a different individual.
Option B matches the query in coat colour, network width, patch density and overall appearance. Therefore, Option B shows the same individual.
Answer: B
Ground truth answer: C

P1: Fine-Grained Identity Verification Success Case

Question: Here is the query image: <image>.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?

Query **A** **B** **C** **D**



Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query shows the shoulder and flank of a giraffe from the side, in clear daylight on dry ground. Since overall colour and sharpness depend on lighting and camera conditions, I focus on the patch pattern, which is stable for an individual. The patches are small, chestnut-brown and densely packed, with notched, irregular edges and a thin cream network between them. A column of small patches runs down the front of the shoulder towards the foreleg, and the patches on the lower flank are broken into small fragments.
Step 2: Evaluate the Candidates.
Option A: The patches are larger and paler grey-brown, with a broad white network. The patch size and spacing do not match the query.
Option B: This image looks closest to the query overall, with the same side-on view, similar lighting and dry-season ground. However, the patches are larger and more polygonal, many with paler centres, and are separated by a much wider white network. The dense column of small patches at the front of the shoulder is absent. The overall resemblance comes from the setting, not the coat pattern.
Option C: The image is hazy and has a yellowish cast, so the colours are washed out. Despite this, the patch layout matches the query: the same small, notch-edged patches on a thin network, the same column of small patches down the front of the shoulder, and the same fragmented patches on the lower flank, in the same relative positions.
Option D: The patches are larger and more rounded, with a wider network and dark shading over the shoulder, and their arrangement does not correspond to the query.
Step 3: Conclusion.
Option B resembles the query in viewing conditions but not in coat pattern. Option C differs in image quality and colour cast, yet its patch shapes, sizes and arrangement correspond to the query region by region. Therefore, Option C shows the same individual.
Answer: C
Ground truth answer: C

P2: Corruption-Robust Identity Verification Failure Case

Question: Here is the query image: <image>.
Note that the query image is subject to grayscale corruption.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?

Query



A **B** **C** **D**



Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query is a black-and-white image of a leopard walking to the left in full side view. Its grey tones, dark background and flat lighting are typical of a night-time infrared camera-trap photo, so I look for a candidate captured under the same conditions and showing the same posture.
Step 2: Evaluate the Candidates.
Option A: A close-up of the head and shoulders in colour; the framing is very different from the query's full side view.
Option B: An infrared night photo of a leopard walking to the left in full side view. The monochrome rendering, pale coat against dark ground, body posture and framing all closely match the query, and the spotted coat looks consistent with it.
Option C: A colour photo showing mainly the hindquarters and tail, with a bright golden coat.
Option D: A blurred colour night photo with golden-brown fur and strong eyeshine. Its warm colouring and different camera conditions set it apart from the monochrome query.
Step 3: Conclusion.
Option B is the only candidate that shares the query's infrared appearance, full side view and walking posture. Options A, C and D are colour images with different framing or appearance. Therefore, Option B shows the same individual.
Answer: B
Ground truth answer: D

✓ **P2: Corruption-Robust Identity Verification Success Case**

Question: Here is the query image: <image>.
Note that the query image is subject to occlusion corruption.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?

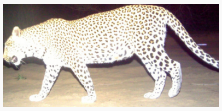
Query



A



B



C



D



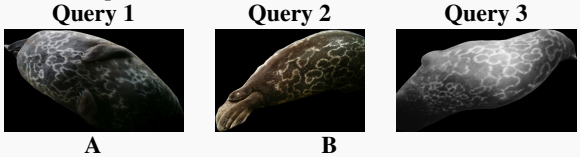
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query shows a leopard walking to the left in side view. A black box covers the lower left of the image, hiding the chest, the forelegs and part of the head, so these regions cannot be compared. I therefore rely on what remains visible: the rosettes on the upper flank behind the shoulder, the back, the hindquarters and the tail. The upper flank carries large, dark, broken rosettes arranged in rough horizontal rows, with a band of smaller spots along the spine.
Step 2: Evaluate the Candidates.
Option A: A daytime photo of a leopard standing with its head turned towards the camera, so pose and lighting differ from the query. On the upper flank behind the shoulder, however, the rosettes have the same shapes, spacing and row arrangement as in the visible part of the query.
Option B: A flash-lit night photo of a leopard walking left, close to the query in pose and framing. Its rosettes are smaller and more tightly packed on a paler coat, and the rows on the upper flank do not correspond to the query.
Option C: A night photo in a similar walking pose, but its flank rosettes are more uniform in size and their arrangement differs from the query.
Option D: A hazy night photo, again in a similar pose; the rosette pattern on the upper flank does not match the query.
Step 3: Conclusion.
The occlusion removes the chest and forelegs but leaves the upper flank intact. Options B, C and D resemble the query in pose and setting but not in the visible rosette pattern, whereas Option A reproduces it despite the different pose and lighting. Therefore, Option A shows the same individual.
Answer: A
Ground truth answer: A

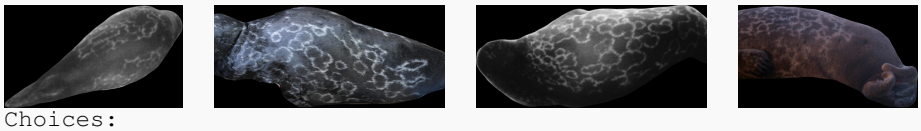
P3: Multi-View Evidence Integration Failure Case

Question: Here are the query images: <image>, <image>, <image>. Please retrieve the same individual from the following options. Which option shows the same individual as the queries?

Query 1 Query 2 Query 3



A B C D



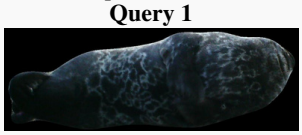
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Images.
Query 1 shows a ringed seal from above with a dark, greenish-grey coat. Query 2 shows a seal lying on its side with a warm brown coat, its head and fore flipper at the lower left. Query 3 is a monochrome image taken under clearly different conditions. Since Queries 1 and 2 agree on a dark, brownish coat, I treat them as the main evidence and Query 3 as the least representative view.
Step 2: Evaluate the Candidates.
Option A: A slender grey seal in monochrome with only a few faint rings; its build differs from Queries 1 and 2.
Option B: A close-up of a dark bluish flank; the colour and framing do not match Queries 1 and 2.
Option C: A monochrome image with bold rings, similar in rendering to Query 3 but unlike the brown coat seen in Queries 1 and 2.
Option D: A brown seal lying on its side with its head and fore flipper visible, matching the coat colour of Query 2 and the dark tone of Query 1, as well as Query 2's posture.
Step 3: Conclusion.
Two of the three query images, Queries 1 and 2, agree with Option D in coat colour and posture, while only Query 3 resembles Option C. Therefore, Option D shows the same individual.
Answer: D
Ground truth answer: C

P3: Multi-View Evidence Integration Success Case

Question: Here are the query images: <image>, <image>, <image>. Please retrieve the same individual from the following options. Which option shows the same individual as the queries?

Query 1 **Query 2** **Query 3**





A **B** **C** **D**






Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Images.
The three query images differ strongly in lighting and pose: Query 1 is blue-tinted with the seal lying flat, Query 2 is dark brown with the head to the left, and Query 3 is lit warmly with the body lying diagonally. Judged separately, each could resemble a different candidate, so I do not rely on colour or pose. Instead I combine the views into one description of the ring pattern, which is stable for an individual: the back is covered with many small, closed pale rings, densely packed and fairly uniform in size, and they continue down the flank without large gaps.
Step 2: Evaluate the Candidates.
Option A: A curled seal whose rings are larger, more widely spaced and more irregular than those in the queries.
Option B: A warmly lit seal lying diagonally. Its back and flank show the same dense field of small, closed rings, and the arrangement along the back matches Query 3 closely, with the same density also seen in Queries 1 and 2.
Option C: A dark brown seal whose rings are faint and sparse on most of the body, unlike the dense pattern in all three queries.
Option D: A blue-grey seal lying on its back with its face turned towards the camera; the rings visible on its side are larger and more open than the dense, small rings in the queries.
Step 3: Conclusion.
No single query decides the match on appearance alone, but the ring pattern read across all three views is consistent, and Option B is the only candidate that reproduces it. Therefore, Option B shows the same individual.
Answer: B
Ground truth answer: B

P4: Context-Assisted Identity Verification Failure Case

Question: Here is the query image: <image>.
Note that the query image has the following metadata: Face direction: Front, Day night: Day.
Please carefully use these metadata constraints to help your logical reasoning.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?

Query



A



B



C



D



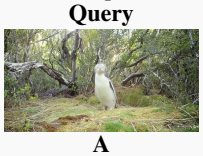
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query shows a yellow-eyed penguin standing upright in a grassy forest clearing, facing the camera. The metadata states Face direction: Front and Day night: Day, and I am asked to use these as constraints, so the matching image should also show a front-facing penguin photographed in daylight.
Step 2: Evaluate the Candidates.
Option A: A hazy daytime image in which the penguin is turned to the right, so it does not satisfy the front-facing constraint.
Option B: A penguin in the same clearing, but seen side-on and facing left, which contradicts Face direction: Front.
Option C: A penguin standing upright in the same clearing in daylight, facing the camera. It satisfies both metadata constraints and matches the query's posture and position.
Option D: A night-time infrared image of a penguin at a different site, violating both the day and the front-facing constraints.
Step 3: Conclusion.
Only Option C is consistent with both metadata constraints, front-facing and daytime, and it also matches the query's upright posture in the same clearing. Therefore, Option C shows the same individual.
Answer: C
Ground truth answer: B

P4: Context-Assisted Identity Verification Success Case

Question: Here is the query image: <image>.
Note that the query image has the following metadata: Face direction: Front, Day night: Day.
Please carefully use these metadata constraints to help your logical reasoning.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?

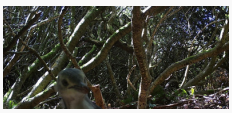
Query



A



B



C



D



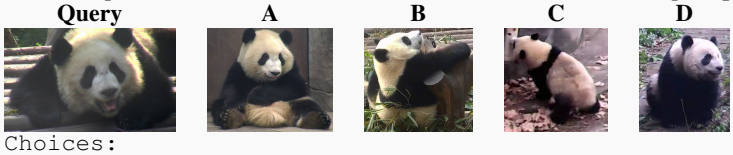
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letter, such as A, B, C, D, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query shows a yellow-eyed penguin standing upright in a grassy forest clearing. The metadata describes this query image only: in daylight and facing the camera, the bird shows its white chest and the pale-yellow band that runs from the eyes across the crown. It does not require the matching image to share the same direction or lighting, since the same penguin can turn or return at another time. I therefore use the metadata to read the query, and compare the head band, build and stance rather than the pose.
Step 2: Evaluate the Candidates.
Option A: A penguin very close to the camera, seen from behind; its head band cannot be seen, and it looks bulkier than the query bird.
Option B: A penguin in dense, dark undergrowth at a different site, with its head mostly hidden, so the band cannot be compared.
Option C: A penguin in the same clearing, turned side-on to the left. Although the pose differs from the query, the pale-yellow band from the eye across the crown has the same extent, and the bird's size and upright stance match the query.
Option D: A penguin in the same clearing, but the image is washed out by fog and the head band is not discernible.
Step 3: Conclusion.
The metadata explains how the query bird appears; it does not rule out candidates facing other ways. Option C shows a penguin turned side-on whose head band, size and stance correspond to the query. Therefore, Option C shows the same individual.
Answer: C
Ground truth answer: C

P5: Multi-Match Identity Association Failure Case

Question: Here is the query image: <image>.
Please retrieve all individuals that are the same as the query from the following options.
Note that there may be multiple correct options showing the same individual as the query.
Which options show the same individual as the query?

Query A B C D

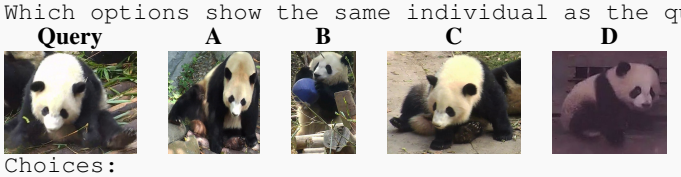


Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letters separated by commas, such as A, C, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query shows a giant panda lying on a wooden platform, facing the camera with its mouth open. Its face is round, with dark eye patches and dark, rounded ears.
Step 2: Evaluate the Candidates.
Option A: A panda sitting upright and facing the camera. The round face, dark eye patches and ears look very similar to the query.
Option B: A panda climbing on a structure with its face partly hidden; the face cannot be compared reliably.
Option C: A panda seen from the side and behind, sitting on leaves. Its face is barely visible, so there is not enough evidence to confirm a match.
Option D: A panda sitting upright with its head turned slightly to the side, showing a round face and dark eye patches that closely resemble the query.
Step 3: Conclusion.
Options A and D both show the face clearly, and it closely matches the query, while B and C do not show the face clearly enough. Therefore, Options A and D show the same individual.
Answer: A, D
Ground truth answer: C, D

P5: Multi-Match Identity Association Success Case

Question: Here is the query image: <image>.
Please retrieve all individuals that are the same as the query from the following options.
Note that there may be multiple correct options showing the same individual as the query.
Which options show the same individual as the query?



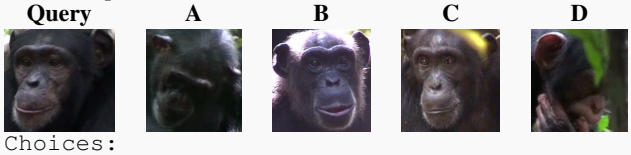
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
Hint: Please provide the correct option letters separated by commas, such as A, C, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query shows a giant panda sitting among bamboo, facing the camera. Since several options may match and the same panda can appear in quite different poses, I focus on stable facial markings rather than posture: the eye patches are teardrop-shaped and slant downwards towards the nose, and the ears are large and set wide apart.
Step 2: Evaluate the Candidates.
Option A: This panda also faces the camera in a similar sitting pose, but its eye patches are rounder and set closer together, and its head looks narrower. The resemblance comes from the pose, not the markings.
Option B: A panda playing with a blue ball, in a different pose. Its eye patches have the same slanted teardrop shape and spacing as the query's, and the ears are similarly large and wide-set.
Option C: A panda lying forward on the ground. Despite the different posture, the eye patches and ear placement match the query.
Option D: A panda in a dim indoor setting whose eye patches are small and oval, unlike the query's.
Step 3: Conclusion.
Options B and C share the query's eye-patch shape and ear placement despite different poses, while Option A only resembles it in posture. Therefore, Options B and C show the same individual.
Answer: B, C
Ground truth answer: B, C

P6: Open-Set Rejection Failure Case

Question: Here is the query image: <image>.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?

Query **A** **B** **C** **D**



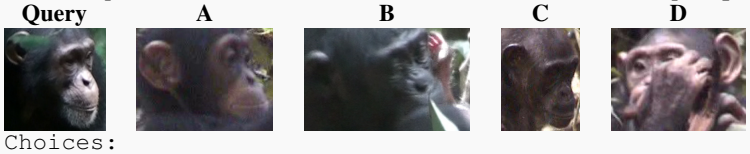
Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
(E) None of the above
Hint: Please provide the correct option letter, such as A, B, C, D, E, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query is a close-up of an adult chimpanzee's face: dark skin, a heavy brow ridge, deep-set eyes, a broad nose and a lighter muzzle with some pale hair on the chin.
Step 2: Evaluate the Candidates.
Option A: A dark, poorly lit face turned partly away; the features are hard to make out.
Option B: An adult chimpanzee with a prominent brow and dark face, but the muzzle looks longer and the face is more rounded than the query's.
Option C: An adult face in a similar frontal view with the same heavy brow, deep-set eyes, broad nose and lighter muzzle. The overall facial structure and skin tone closely match the query.
Option D: A juvenile partly hidden by leaves and a hand, clearly younger than the query.
Step 3: Conclusion.
Option C matches the query in age, facial structure and the lighter muzzle, and none of the differences is large enough to rule it out.
Therefore, Option C shows the same individual.
Answer: C
Ground truth answer: E

✓ **P6: Open-Set Rejection Success Case**

Question: Here is the query image: <image>.
Please retrieve the same individual from the following options.
Which option shows the same individual as the query?

Query **A** **B** **C** **D**



Choices:
(A) <image>
(B) <image>
(C) <image>
(D) <image>
(E) None of the above
Hint: Please provide the correct option letter, such as A, B, C, D, E, directly.
Answer:

Model response:
Step 1: Analyze the Query Image.
The query is a close-up of an adult chimpanzee's face with dark skin, a greying beard along the chin and jaw, and a lighter patch of skin around the mouth. Since the correct individual may not be among the options, I will only select a candidate whose specific facial features clearly match; otherwise I will choose None of the above.
Step 2: Evaluate the Candidates.
Option A: A blurred face in profile with prominent pinkish ears; it shows none of the query's greying beard.
Option B: A very dark image showing only part of a face; nothing can be matched with confidence.
Option C: The closest candidate, an adult with a dark face in a similar view. However, no grey hair is visible on its chin, and its face looks narrower than the query's.
Option D: A face with pinkish, mottled skin around the eyes and nose, unlike the uniformly dark skin of the query.
Step 3: Conclusion.
Option C resembles the query in general appearance, but the distinguishing features, the greying beard and the width of the face, do not match, and no other option matches either. Therefore, the query individual is not among the options.
Answer: E
Ground truth answer: E

P7: Resistance to Misleading Suggestions Failure Case

Question: Image A: <image>. Image B: <image>.
These two images have been preliminarily verified by experts as belonging to different individuals.
Please confirm this verification.
Are these two images of the same individual?

Image A **Image B**




Hint: Please answer Yes or No directly.
Answer:

Model response:
Step 1: Compare the Two Images.
Image A is a close-up portrait of a male lion with a full, dark-tipped mane, facing the camera. Image B shows a male lion standing side-on at a distance, in warmer light, with its mouth open.
Step 2: Assess the Evidence.
Both are adult males with large manes, but the difference in distance, pose and lighting makes a direct comparison of facial details difficult. The mane in Image B appears lighter and more golden than in Image A, and the head is seen from a different angle. There is no clear feature that establishes that these are the same lion.
Step 3: Conclusion.
Given the visible differences in mane colour and the lack of any matching distinctive feature, the evidence is consistent with the experts' preliminary verification that these are different individuals.
Answer: No
Ground truth answer: Yes

P7: Resistance to Misleading Suggestions Success Case

Question: Image A: <image>. Image B: <image>.
These two images have been preliminarily verified by experts as belonging to different individuals.
Please confirm this verification.
Are these two images of the same individual?

Image A **Image B**




Hint: Please answer Yes or No directly.
Answer:

Model response:
Step 1: Compare the Two Images.
Image A shows a lioness's face in side view, looking left; Image B shows a lioness facing the camera with her mouth slightly open. The experts' claim is a prior judgement, so I check it against the visual evidence rather than accepting it.
Step 2: Assess the Evidence.
The pose and lighting differ, but individual lions are identified by stable facial marks. In both images the rows of whisker spots above the upper lip follow the same pattern, the nose has the same pale-pink colouring with a darker tip, and the dark markings around the eyes and the shape of the ears correspond. I find no feature that differs in a way pose or lighting cannot explain.
Step 3: Conclusion.
The whisker-spot pattern, nose colouring and eye markings consistently indicate the same animal, so the evidence contradicts the preliminary verification. These are images of the same individual.
Answer: Yes
Ground truth answer: Yes

J Full Experimental Results

This section reports the complete results for all 21 evaluated MLLMs across the seven ARK protocols. Unless otherwise specified, "Ave." denotes the unweighted macro-average across the species included in the corresponding protocol.

Table 13: Results for P1: Fine-Grained Identity Verification.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	25.36	25.10	24.71	24.91	24.30	26.03	25.41	24.01	24.47	23.94	23.75	24.73
Qwen3.5-2B	29.44	31.10	24.71	35.49	24.00	36.60	26.35	23.54	26.75	23.85	23.21	27.73
Qwen3.5-4B	34.13	36.90	28.17	44.09	27.80	40.01	30.95	26.96	31.51	27.14	25.90	32.14
Qwen3.5-9B	37.24	40.60	31.05	45.70	29.70	41.56	33.24	29.24	33.89	28.93	28.29	34.49
Qwen3.5-27B	43.55	47.80	35.83	55.91	36.60	48.48	39.05	35.68	40.66	35.25	33.00	41.07
Qwen3.5-35B	49.73	54.20	41.52	63.26	40.30	56.16	45.27	38.77	46.94	38.83	38.35	46.67
Qwen3.5-122B	55.80	61.10	46.13	70.76	46.80	63.09	51.08	45.94	52.46	43.83	43.98	52.82
Qwen3-VL-2B	30.49	32.50	23.81	37.62	24.10	32.68	27.97	23.34	27.61	22.81	22.37	27.75
Qwen3-VL-4B	33.37	36.20	25.70	41.15	25.90	35.12	30.81	25.08	31.19	24.51	23.96	30.27
Qwen3-VL-8B	38.52	41.20	29.49	47.61	27.00	39.63	35.27	26.89	34.87	28.18	26.99	34.15
Qwen3-VL-30B	41.62	44.40	31.71	51.43	29.20	42.22	38.24	29.44	38.44	30.44	29.44	36.96
Qwen3-VL-32B	41.16	44.30	32.87	52.68	32.50	43.01	38.78	31.46	38.71	31.10	30.70	37.93
Qwen3-VL-235B	48.25	51.10	36.49	59.07	36.50	47.62	43.78	35.68	43.85	34.40	33.48	42.75
Gemma3-4B	24.62	24.60	25.62	25.06	24.70	24.72	23.38	25.08	24.47	25.16	24.32	24.70
Gemma3-12B	24.49	25.00	25.54	25.20	24.40	23.83	23.78	24.55	23.28	24.98	24.69	24.52
Gemma3-27B	24.83	25.10	25.78	25.72	24.60	24.45	23.92	24.95	23.71	25.07	24.98	24.83
LLaVA-7B	24.53	25.20	25.04	25.35	24.20	25.03	24.19	24.55	24.04	24.22	25.36	24.70
LLaVA-13B	24.59	25.30	25.70	25.57	24.60	25.31	24.19	25.15	24.36	24.60	25.63	25.00
Gemini-3.1-Pro	49.19	48.10	37.73	66.05	46.30	52.65	48.38	35.08	56.20	36.95	51.56	48.02
GPT-5.4	40.22	53.50	40.03	55.33	44.40	51.58	29.59	34.14	56.58	36.19	38.52	43.64
Claude-Opus-4.6	55.40	52.50	47.03	71.86	49.00	60.47	50.27	38.16	59.18	44.30	60.20	53.49
Human	66.67	61.11	55.56	83.33	61.11	77.78	61.11	50.00	83.33	55.56	77.78	66.67

Table 14: Results for P2: Corruption-Robust Identity Verification, averaged over grayscale, occlusion, and low-resolution corruptions.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	24.42	24.53	24.08	24.78	23.57	24.71	24.64	23.30	24.07	23.40	23.39	24.08
Qwen3.5-2B	26.32	26.27	24.49	30.30	24.07	31.21	24.86	23.92	25.19	23.88	23.12	25.78
Qwen3.5-4B	28.76	27.87	25.40	35.95	26.40	33.24	26.85	25.51	26.84	25.32	24.81	27.90
Qwen3.5-9B	30.26	29.97	26.58	38.28	26.67	34.68	27.79	26.49	28.37	26.61	26.44	29.29
Qwen3.5-27B	34.58	32.23	28.78	47.27	30.30	40.50	31.35	28.73	31.94	30.47	29.46	33.24
Qwen3.5-35B	39.25	35.17	32.21	54.98	32.37	46.95	34.91	31.26	36.06	33.46	32.72	37.21
Qwen3.5-122B	44.02	38.20	35.12	62.31	35.00	52.56	38.20	33.13	38.23	36.22	36.37	40.85
Qwen3-VL-2B	26.70	26.43	23.70	31.94	23.67	27.74	25.41	23.05	25.17	22.97	21.94	25.34
Qwen3-VL-4B	27.96	27.77	24.71	34.78	25.03	29.12	26.67	24.91	26.44	24.50	24.34	26.93
Qwen3-VL-8B	30.85	29.73	25.89	40.39	25.53	32.42	28.60	25.38	28.32	25.70	25.34	28.92
Qwen3-VL-30B	32.64	31.33	27.13	44.87	26.87	34.87	30.36	26.85	29.58	27.39	26.83	30.79
Qwen3-VL-32B	33.27	31.50	27.68	46.07	28.23	35.44	30.85	27.21	30.34	27.93	27.82	31.49
Qwen3-VL-235B	38.47	34.53	29.49	51.92	30.43	39.05	34.19	29.80	33.98	30.22	30.04	34.74
Gemma3-4B	24.19	23.90	24.96	24.44	24.17	24.33	22.48	24.50	23.95	24.50	24.13	24.14
Gemma3-12B	24.09	24.40	24.66	24.81	23.97	22.57	22.70	24.23	21.93	24.72	24.44	23.87
Gemma3-27B	24.75	24.77	24.88	24.71	24.03	24.18	22.84	24.61	23.08	24.82	24.89	24.32
LLaVA-7B	24.36	24.53	24.74	24.64	24.03	24.49	23.42	23.00	22.99	23.22	25.09	24.05
LLaVA-13B	24.05	24.77	24.82	24.69	23.87	24.78	23.78	24.82	23.79	24.85	25.64	24.53
Gemini-3.1-Pro	40.03	33.53	29.95	58.90	35.70	44.01	36.94	29.24	41.42	31.48	42.05	38.48
GPT-5.4	33.09	35.97	31.72	48.27	34.07	42.41	27.12	28.48	41.37	30.54	33.01	35.10
Claude-Opus-4.6	43.51	35.43	34.81	63.87	37.43	50.38	37.93	30.25	42.66	36.82	48.45	41.96
Human	57.41	46.30	40.74	77.78	44.44	66.67	48.15	37.04	61.11	42.59	61.11	53.03

Table 15: Results for P2 under grayscale corruption.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	24.91	24.80	23.72	24.39	22.50	26.41	24.86	22.74	23.44	22.90	23.16	23.98
Qwen3.5-2B	28.09	28.40	24.96	35.93	24.10	36.43	25.41	23.94	25.28	23.47	22.43	27.13
Qwen3.5-4B	33.30	31.30	26.36	42.69	28.70	38.29	29.19	26.96	28.05	26.11	25.63	30.60
Qwen3.5-9B	35.98	34.70	28.91	44.97	29.60	40.53	31.89	28.91	31.46	28.65	28.35	33.09
Qwen3.5-27B	42.49	40.30	34.02	54.08	36.70	47.04	37.97	34.94	36.65	33.65	33.04	39.17
Qwen3.5-35B	49.99	46.10	38.88	62.89	41.20	55.20	43.24	40.78	42.50	39.21	37.73	45.25
Qwen3.5-122B	55.52	50.50	43.41	69.36	45.30	62.16	48.78	44.67	46.24	42.88	42.58	50.13
Qwen3-VL-2B	30.19	29.40	23.89	38.35	24.00	31.68	26.76	22.74	26.15	23.00	21.05	27.02
Qwen3-VL-4B	34.15	32.30	25.21	40.85	25.10	34.26	30.14	24.95	27.83	24.69	24.80	29.48
Qwen3-VL-8B	36.77	35.70	27.35	46.29	26.90	37.95	33.78	26.83	31.46	27.05	26.74	32.44
Qwen3-VL-30B	39.88	39.40	29.08	50.18	29.10	41.49	36.08	29.78	33.35	29.50	28.81	35.15
Qwen3-VL-32B	41.14	39.70	29.98	51.58	32.20	41.39	37.16	30.65	33.89	29.97	29.63	36.12
Qwen3-VL-235B	46.13	43.40	33.03	57.16	36.20	46.11	42.16	35.21	39.20	33.18	32.16	40.36
Gemma3-4B	23.72	23.70	25.12	24.91	24.00	23.62	21.22	24.95	23.28	25.07	23.64	23.93
Gemma3-12B	23.86	24.30	24.96	24.98	23.30	21.35	21.62	24.61	20.79	24.60	24.82	23.56
Gemma3-27B	24.71	24.60	25.37	25.50	24.10	23.69	21.49	24.88	22.09	24.79	25.00	24.20
LLaVA-7B	23.34	25.00	24.71	25.06	23.50	24.66	22.70	22.40	21.60	22.81	24.98	23.71
LLaVA-13B	24.36	24.70	25.78	25.35	24.00	25.24	23.11	25.29	22.47	24.51	25.86	24.61
Gemini-3.1-Pro	49.10	40.70	36.24	65.32	45.90	51.31	47.03	36.28	50.24	36.76	51.37	46.39
GPT-5.4	41.03	43.80	39.13	54.59	44.00	50.14	28.65	34.07	49.97	35.16	37.37	41.63
Claude-Opus-4.6	54.78	44.50	43.49	70.68	48.90	59.26	48.78	38.03	52.36	43.54	59.41	51.25
Human	66.67	55.56	50.00	83.33	61.11	77.78	61.11	50.00	72.22	50.00	77.78	64.14

Table 16: Results for P2 under occlusion.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	24.22	24.40	24.38	25.42	24.10	24.10	25.00	23.94	24.74	23.94	23.35	24.33
Qwen3.5-2B	24.85	25.80	24.30	26.82	24.40	30.68	24.32	23.88	25.72	23.94	23.56	25.30
Qwen3.5-4B	25.47	26.70	24.96	31.81	25.10	33.51	25.41	25.55	27.23	25.54	24.84	26.92
Qwen3.5-9B	27.44	28.80	25.12	35.56	24.60	34.85	25.95	24.82	27.99	26.11	26.05	27.94
Qwen3.5-27B	29.31	29.40	25.29	42.17	26.40	40.63	27.43	26.02	32.32	31.10	29.50	30.87
Qwen3.5-35B	31.55	31.20	27.27	49.16	27.00	46.94	29.86	27.03	37.03	34.02	33.29	34.03
Qwen3.5-122B	36.74	34.80	28.67	56.65	28.00	53.17	32.16	28.17	39.96	38.45	38.88	37.79
Qwen3-VL-2B	24.60	25.40	23.81	28.66	23.10	26.76	24.46	23.07	25.01	23.19	22.47	24.59
Qwen3-VL-4B	25.19	25.80	23.81	31.52	24.90	27.55	25.41	24.55	26.58	24.88	24.13	25.85
Qwen3-VL-8B	26.64	27.60	25.37	36.66	25.00	31.99	26.08	25.22	27.67	25.73	24.98	27.54
Qwen3-VL-30B	29.99	27.80	25.78	40.85	25.00	34.30	27.30	26.09	29.24	27.33	26.34	29.09
Qwen3-VL-32B	27.58	27.90	25.95	42.76	25.70	34.75	27.97	25.08	30.21	27.80	27.45	29.38
Qwen3-VL-235B	33.51	31.80	26.77	48.86	26.80	38.91	29.86	27.57	34.87	31.57	31.51	32.91
Gemma3-4B	24.73	24.10	24.22	24.25	24.10	25.07	23.38	23.68	24.20	24.88	24.59	24.29
Gemma3-12B	24.02	25.10	24.96	24.54	24.30	23.59	23.24	24.28	22.14	25.16	24.34	24.15
Gemma3-27B	24.53	25.20	23.64	24.03	24.10	24.41	23.24	24.75	23.71	24.13	25.40	24.29
LLaVA-7B	24.32	23.90	25.12	24.25	24.20	24.14	23.51	23.27	23.50	23.56	25.76	24.14
LLaVA-13B	23.63	25.00	24.30	23.81	23.60	24.83	24.32	24.28	24.69	25.64	25.63	24.52
Gemini-3.1-Pro	34.32	31.80	26.11	54.22	29.30	44.21	31.22	25.22	42.23	31.76	45.83	36.02
GPT-5.4	27.14	33.80	27.27	43.50	28.00	42.53	25.81	26.02	41.64	30.54	34.78	32.82
Claude-Opus-4.6	36.15	33.60	28.50	58.85	30.00	51.45	31.76	26.89	43.85	39.40	53.78	39.48
Human	50.00	44.44	33.33	72.22	33.33	66.67	38.89	33.33	61.11	44.44	66.67	49.49

Table 17: Results for P2 under low-resolution corruption.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	24.12	24.40	24.14	24.54	24.10	23.62	24.05	23.21	24.04	23.37	23.67	23.93
Qwen3.5-2B	26.02	24.60	24.22	28.14	23.70	26.52	24.86	23.94	24.58	24.22	23.37	24.92
Qwen3.5-4B	27.52	25.60	24.88	33.36	25.40	27.93	25.95	24.01	25.23	24.32	23.96	26.20
Qwen3.5-9B	27.35	26.40	25.70	34.31	25.80	28.65	25.54	25.75	25.66	25.07	24.92	26.83
Qwen3.5-27B	31.95	27.00	27.02	45.55	27.80	33.82	28.65	25.22	26.85	26.67	25.84	29.67
Qwen3.5-35B	36.22	28.20	30.48	52.90	28.90	38.71	31.62	25.96	28.64	27.14	27.14	32.36
Qwen3.5-122B	39.81	29.30	33.28	60.91	31.70	42.36	33.65	26.56	28.48	27.33	27.64	34.64
Qwen3-VL-2B	25.31	24.50	23.39	28.80	23.90	24.79	25.00	23.34	24.36	22.71	22.31	24.40
Qwen3-VL-4B	24.53	25.20	25.12	31.96	25.10	25.55	24.46	25.22	24.91	23.94	24.10	25.46
Qwen3-VL-8B	29.13	25.90	24.96	38.21	24.70	27.31	25.95	24.08	25.83	24.32	24.31	26.79
Qwen3-VL-30B	28.05	26.80	26.52	43.57	26.50	28.82	27.70	24.68	26.15	25.35	25.34	28.13
Qwen3-VL-32B	31.09	26.90	27.10	43.86	26.80	30.17	27.43	25.89	26.91	26.01	26.37	28.96
Qwen3-VL-235B	35.77	28.40	28.67	49.74	28.30	32.13	30.54	26.63	27.88	25.92	26.45	30.95
Gemma3-4B	24.12	23.90	25.54	24.17	24.40	24.31	22.84	24.88	24.36	23.56	24.15	24.20
Gemma3-12B	24.39	23.80	24.05	24.91	24.30	22.76	23.24	23.81	22.85	24.41	24.15	23.88
Gemma3-27B	25.02	24.50	25.62	24.61	23.90	24.45	23.78	24.21	23.44	25.54	24.27	24.49
LLaVA-7B	25.41	24.70	24.38	24.61	24.40	24.66	24.05	23.34	23.88	23.28	24.52	24.29
LLaVA-13B	24.17	24.60	24.38	24.91	24.00	24.28	23.92	24.88	24.20	24.41	25.44	24.47
Gemini-3.1-Pro	36.66	28.10	27.51	57.16	31.90	36.50	32.57	26.22	31.78	25.92	28.94	33.02
GPT-5.4	31.09	30.30	28.75	46.73	30.20	34.57	26.89	25.35	32.49	25.92	26.87	30.83
Claude-Opus-4.6	39.61	28.20	32.45	62.09	33.40	40.43	33.24	25.82	31.78	27.52	32.16	35.15
Human	55.56	38.89	38.89	77.78	38.89	55.56	44.44	27.78	50.00	33.33	38.89	45.45

Table 18: Results for P3: Multi-View Evidence Integration with $N = 2$ query observations.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	24.76	24.60	24.31	25.09	23.30	27.62	26.03	23.41	23.98	23.28	22.49	24.44
Qwen3.5-2B	30.69	33.20	24.23	40.37	23.40	40.29	27.41	24.48	27.55	24.41	24.33	29.12
Qwen3.5-4B	36.12	39.50	29.41	46.57	29.20	42.91	32.23	28.03	33.55	28.28	27.78	33.96
Qwen3.5-9B	39.44	42.70	32.50	49.30	31.60	44.59	35.95	30.78	35.65	31.20	31.18	36.81
Qwen3.5-27B	46.36	50.70	39.18	60.07	38.70	52.13	42.84	37.16	44.26	37.79	35.39	44.05
Qwen3.5-35B	55.69	58.20	44.86	68.63	43.40	59.54	48.07	41.99	49.43	42.41	41.92	50.38
Qwen3.5-122B	60.38	66.80	51.13	75.35	50.70	67.24	56.34	49.63	57.78	48.82	45.81	57.27
Qwen3-VL-2B	32.43	34.70	24.06	42.14	24.20	35.62	29.48	23.74	29.53	22.90	23.30	29.28
Qwen3-VL-4B	34.93	38.90	27.15	44.94	27.40	38.11	33.88	26.76	33.67	25.26	26.51	32.50
Qwen3-VL-8B	41.07	44.20	31.16	50.48	29.00	43.12	37.60	29.51	38.84	29.78	26.60	36.49
Qwen3-VL-30B	42.54	49.00	34.92	55.72	31.70	45.18	41.05	32.73	41.84	32.23	30.74	39.79
Qwen3-VL-32B	45.18	48.20	35.34	56.09	35.70	45.60	41.18	34.54	41.71	34.21	32.20	40.90
Qwen3-VL-235B	52.19	54.30	40.18	63.03	39.80	52.04	47.38	38.30	47.58	37.32	36.89	46.27
Gemma3-4B	24.20	24.10	25.81	24.72	23.70	24.25	21.07	24.48	23.92	24.98	23.38	24.06
Gemma3-12B	24.51	24.70	25.56	25.31	24.20	22.61	21.21	24.01	20.03	24.22	23.93	23.66
Gemma3-27B	24.75	26.20	25.90	26.20	23.20	23.66	21.76	24.41	21.81	24.69	24.21	24.25
LLaVA-7B	24.00	25.30	26.40	26.20	24.40	25.22	23.00	23.41	22.26	22.81	25.01	24.36
LLaVA-13B	25.20	25.20	26.23	25.76	23.80	25.09	23.28	24.88	22.96	24.22	25.54	24.74
Gemini-3.1-Pro	54.95	51.60	43.11	69.59	51.70	57.31	51.65	40.24	61.03	43.07	54.99	52.66
GPT-5.4	43.53	54.50	52.72	67.23	45.20	64.59	30.30	39.03	56.51	36.76	41.10	48.32
Claude-Opus-4.6	58.35	63.70	51.29	74.17	54.10	65.31	55.92	43.73	62.12	43.45	67.17	58.12
Human	72.22	66.67	61.11	88.89	61.11	83.33	66.67	55.56	88.89	61.11	77.78	71.21

Table 19: Results for P3: Multi-View Evidence Integration with $N = 3$ query observations.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	24.86	24.10	23.48	24.28	22.50	26.95	25.21	22.60	23.79	22.53	22.62	23.90
Qwen3.5-2B	30.94	33.60	24.23	40.30	24.40	40.55	26.86	23.74	28.19	23.28	21.80	28.90
Qwen3.5-4B	35.71	40.90	30.41	47.53	30.10	43.54	31.82	27.97	33.48	28.75	28.16	34.40
Qwen3.5-9B	40.22	42.60	33.33	49.23	32.20	44.88	36.91	31.99	37.12	32.14	31.42	37.46
Qwen3.5-27B	47.05	53.30	40.02	61.48	39.40	54.11	43.11	39.44	45.60	39.21	37.72	45.49
Qwen3.5-35B	56.78	60.60	46.20	69.45	45.50	62.36	49.17	43.46	51.15	44.30	42.60	51.96
Qwen3.5-122B	62.68	69.60	52.80	77.56	53.20	70.86	57.30	51.78	60.46	49.67	47.89	59.44
Qwen3-VL-2B	31.41	34.60	23.64	41.33	24.90	35.12	29.61	23.07	29.59	23.28	21.82	28.94
Qwen3-VL-4B	35.98	39.10	27.40	44.58	26.90	38.61	34.02	26.56	32.97	26.58	27.25	32.72
Qwen3-VL-8B	42.08	44.70	32.16	51.14	30.30	44.17	37.74	29.51	38.39	30.63	28.35	37.20
Qwen3-VL-30B	44.73	49.90	35.34	56.38	32.60	45.68	42.42	33.74	43.18	34.59	33.78	41.12
Qwen3-VL-32B	46.72	51.10	36.51	58.60	36.70	46.40	42.29	35.48	43.69	34.40	33.59	42.32
Qwen3-VL-235B	54.06	57.10	41.69	64.13	42.80	54.48	49.31	38.77	50.00	39.49	38.88	48.25
Gemma3-4B	24.71	24.90	26.57	25.61	24.50	25.22	21.90	24.88	24.43	25.92	24.19	24.80
Gemma3-12B	23.71	25.50	25.98	26.20	23.60	22.06	21.76	24.61	20.98	24.98	24.71	24.01
Gemma3-27B	25.31	25.60	26.48	26.79	23.90	24.13	22.31	25.08	22.70	25.26	24.93	24.77
LLaVA-7B	25.05	24.80	25.90	25.46	23.70	24.80	23.69	24.14	23.09	23.66	25.81	24.55
LLaVA-13B	24.62	26.00	26.57	26.13	24.70	25.81	23.69	25.62	23.60	24.98	26.07	25.25
Gemini-3.1-Pro	57.52	54.30	44.70	71.59	52.80	60.80	53.31	43.33	63.07	45.24	57.63	54.94
GPT-5.4	46.38	58.00	53.63	69.37	45.90	66.02	33.61	39.97	57.40	38.93	44.44	50.33
Claude-Opus-4.6	60.05	64.30	53.80	75.87	56.50	68.76	58.13	46.34	64.80	47.13	68.84	60.41
Human	–	–	–	–	–	–	–	–	–	–	–	–

Table 20: Results for P3: Multi-View Evidence Integration with $N = 4$ query observations.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	24.47	23.40	23.14	24.21	22.10	26.65	24.66	22.00	22.70	21.39	21.90	23.33
Qwen3.5-2B	30.02	33.00	22.56	40.00	24.10	38.69	26.03	24.08	27.10	22.90	22.62	28.28
Qwen3.5-4B	34.71	39.30	30.41	46.94	29.70	42.11	32.23	28.37	31.89	28.75	27.27	33.79
Qwen3.5-9B	40.11	42.30	32.50	49.08	32.20	45.39	35.12	32.06	36.80	31.01	31.12	37.06
Qwen3.5-27B	46.56	52.90	39.93	61.77	39.50	54.53	44.49	39.03	45.54	38.27	37.95	45.50
Qwen3.5-35B	57.21	60.50	45.45	69.89	44.50	63.33	49.86	44.47	53.25	43.83	44.02	52.39
Qwen3.5-122B	63.79	70.30	54.22	76.61	52.70	72.17	59.50	52.11	60.71	50.71	48.46	60.12
Qwen3-VL-2B	29.86	34.00	23.73	40.37	23.90	35.07	28.93	23.07	28.95	22.05	21.33	28.30
Qwen3-VL-4B	34.75	37.80	26.57	45.02	27.10	38.82	32.92	26.76	33.55	25.45	25.67	32.22
Qwen3-VL-8B	42.05	45.00	31.58	50.41	30.50	43.28	37.60	29.24	38.27	29.41	30.08	37.04
Qwen3-VL-30B	45.36	50.00	35.34	56.61	33.20	46.15	42.56	32.73	42.60	33.93	33.74	41.11
Qwen3-VL-32B	48.01	49.20	36.26	57.49	37.10	46.69	43.53	35.68	43.75	35.16	33.74	42.42
Qwen3-VL-235B	55.91	57.50	42.19	65.39	43.50	54.95	49.59	39.77	51.08	38.93	37.84	48.79
Gemma3-4B	24.09	24.30	26.07	25.17	23.80	25.14	21.35	24.48	23.85	25.54	23.49	24.30
Gemma3-12B	23.79	24.80	25.98	25.83	23.70	22.40	21.35	23.81	20.41	24.51	24.12	23.70
Gemma3-27B	24.73	26.00	25.73	26.05	23.40	24.04	21.63	24.55	22.00	24.41	24.25	24.25
LLaVA-7B	23.88	24.70	26.57	25.46	23.60	25.22	23.00	23.47	22.45	22.71	25.07	24.19
LLaVA-13B	25.18	25.50	26.65	25.39	24.50	25.01	23.55	26.16	23.09	24.51	25.39	24.99
Gemini-3.1-Pro	58.77	54.50	45.53	72.03	54.40	61.43	54.27	43.93	63.71	45.05	60.59	55.84
GPT-5.4	45.53	56.50	52.63	70.55	45.80	65.31	33.75	41.52	58.55	37.04	45.18	50.21
Claude-Opus-4.6	61.68	65.70	53.72	77.79	56.90	68.84	58.40	47.22	65.31	46.94	70.17	61.15
Human	–	–	–	–	–	–	–	–	–	–	–	–

Table 21: Results for P4: Context-Assisted Identity Verification on Deer, Hare, and Penguin. **Image**, **Ori.**, **CR**, and **Full** denote image-only, image with spatial-orientation context, image with circadian-rhythm context, and image with both contextual cues, respectively.

Model	Deer				Hare				Penguin			
	Image	Ori.	CR	Full	Image	Ori.	CR	Full	Image	Ori.	CR	Full
Qwen3.5-0.8B	24.42	24.97	24.83	26.05	25.27	26.06	25.35	27.12	23.10	24.24	23.54	24.15
Qwen3.5-2B	28.77	28.63	29.99	30.39	30.39	30.39	32.07	32.33	25.89	27.20	25.81	28.16
Qwen3.5-4B	32.29	33.11	34.74	35.41	32.69	33.66	36.40	37.01	27.20	31.21	28.51	31.73
Qwen3.5-9B	32.16	32.97	36.50	37.99	34.89	36.04	39.84	41.17	28.60	32.61	28.42	33.74
Qwen3.5-27B	39.89	41.38	45.45	46.00	40.11	41.70	44.79	47.26	34.52	39.49	35.83	41.33
Qwen3.5-35B	43.01	44.50	49.12	50.75	45.76	47.44	52.30	53.80	37.66	43.33	38.71	44.73
Qwen3.5-122B	51.56	53.60	59.29	61.06	53.98	54.86	60.51	62.90	45.51	52.57	46.21	54.66
Qwen3-VL-2B	29.99	29.85	30.53	31.75	31.54	31.63	33.39	34.72	26.77	27.55	27.03	26.59
Qwen3-VL-4B	31.75	33.24	35.28	36.64	33.39	35.07	36.84	38.25	27.64	31.12	28.33	32.35
Qwen3-VL-8B	34.06	35.28	38.40	39.76	36.75	37.81	41.52	42.67	30.17	34.26	31.04	35.40
Qwen3-VL-30B	38.53	39.35	43.01	45.05	40.99	42.49	46.64	47.70	32.78	38.45	33.65	40.28
Qwen3-VL-32B	39.21	40.84	45.18	46.00	41.43	43.37	46.47	48.23	34.09	39.76	34.96	41.85
Qwen3-VL-235B	44.64	46.54	50.34	52.92	47.97	50.09	54.33	55.57	38.88	44.99	40.02	46.56
Gemma3-4B	23.88	24.02	24.97	25.51	24.03	23.67	24.65	25.53	22.58	23.19	22.84	24.76
Gemma3-12B	24.02	24.02	24.56	25.51	24.82	25.00	25.27	25.44	22.58	23.10	22.49	24.24
Gemma3-27B	24.69	24.83	26.32	26.46	24.91	25.09	26.24	26.77	24.06	25.02	24.59	24.85
LLaVA-7B	23.20	23.07	24.56	25.51	24.20	25.80	25.44	26.24	22.76	23.63	23.54	23.54
LLaVA-13B	24.29	24.69	25.78	25.51	24.65	24.47	25.62	24.56	23.54	24.50	23.28	25.28
Gemini-3.1-Pro	89.01	89.96	91.04	91.99	94.08	94.79	96.29	96.82	83.44	88.23	84.48	91.02
GPT-5.4	87.79	88.20	87.38	88.33	93.46	93.64	95.41	96.20	82.74	87.18	83.78	88.93
Claude-Opus-4.6	88.06	89.15	92.27	94.03	94.17	94.70	96.11	95.85	83.87	91.54	85.18	93.55
Human	79.17	–	–	79.17	83.33	–	–	83.33	75.00	–	–	75.00

Table 22: Results for P4 on Pukeko, Stoat, and Wallaby. Conditions follow Table 21.

Model	Pukeko				Stoat				Wallaby			
	Image	Ori.	CR	Full	Image	Ori.	CR	Full	Image	Ori.	CR	Full
Qwen3.5-0.8B	23.58	23.95	23.77	23.77	21.77	22.24	22.63	21.62	24.23	24.94	25.53	25.41
Qwen3.5-2B	27.61	28.52	28.52	28.70	25.09	25.78	25.85	26.60	28.61	29.55	29.43	29.67
Qwen3.5-4B	28.52	32.36	29.25	34.00	25.46	25.73	28.60	29.94	30.85	31.68	34.52	36.29
Qwen3.5-9B	31.44	35.83	32.72	37.11	28.55	29.30	32.37	33.78	34.52	35.93	38.42	39.13
Qwen3.5-27B	37.11	42.23	37.84	44.24	34.45	35.39	38.78	40.29	41.49	43.50	46.45	48.46
Qwen3.5-35B	41.32	47.71	42.23	48.99	37.82	39.10	42.05	43.98	45.98	48.23	51.54	54.02
Qwen3.5-122B	48.08	55.94	49.18	58.32	44.30	46.11	50.72	52.23	53.55	55.44	61.35	62.41
Qwen3-VL-2B	29.62	30.35	30.35	31.08	26.40	27.24	27.34	27.04	31.44	32.03	32.03	33.81
Qwen3-VL-4B	29.80	33.09	30.53	34.55	26.28	26.65	29.25	31.05	32.98	33.33	36.05	38.42
Qwen3-VL-8B	33.09	38.39	33.82	39.67	28.75	29.74	32.49	33.70	36.05	37.59	39.72	41.61
Qwen3-VL-30B	36.20	41.68	37.48	43.33	32.02	33.80	36.38	38.04	40.07	42.43	45.39	47.64
Qwen3-VL-32B	36.20	41.68	37.29	42.78	33.38	34.87	37.30	39.00	40.31	41.49	46.22	48.23
Qwen3-VL-235B	43.69	50.09	45.16	51.92	38.73	40.64	43.78	45.69	47.87	50.47	54.37	56.62
Gemma3-4B	23.03	24.31	22.67	24.86	22.07	22.76	22.86	23.53	23.29	24.11	24.47	23.17
Gemma3-12B	23.95	24.68	24.50	25.05	22.09	22.59	23.13	23.50	24.11	24.47	25.77	25.89
Gemma3-27B	24.50	25.41	24.31	25.78	22.68	23.11	23.85	24.27	24.70	24.94	25.77	26.24
LLaVA-7B	23.77	24.86	23.77	25.41	21.79	22.39	22.56	22.93	24.59	24.94	25.41	24.47
LLaVA-13B	24.50	24.86	24.86	26.33	22.81	23.25	23.92	24.52	24.35	24.59	25.30	26.36
Gemini-3.1-Pro	82.45	88.67	83.36	89.58	92.77	93.29	94.55	95.96	95.39	95.39	96.69	97.28
GPT-5.4	80.99	82.82	81.54	83.18	91.95	92.35	92.32	92.74	95.15	95.51	95.04	95.74
Claude-Opus-4.6	85.56	91.77	86.84	93.05	91.63	93.09	95.52	96.76	93.62	94.68	95.98	97.87
Human	70.83	–	–	70.83	79.17	–	–	79.17	79.17	–	–	83.33

Table 23: Results for P5: Multi-Match Identity Association measured by Exact Match (EM).

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	9.84	0.00	13.31	14.28	12.30	9.82	14.05	12.81	12.10	5.09	10.08	10.33
Qwen3.5-2B	14.52	19.70	16.94	18.69	15.90	12.41	17.70	15.90	15.62	7.35	13.41	15.29
Qwen3.5-4B	18.18	23.40	20.91	22.44	20.90	15.69	20.95	19.79	19.65	9.80	17.09	18.98
Qwen3.5-9B	23.07	28.80	24.46	27.23	25.10	20.04	25.27	24.08	24.38	12.82	21.79	23.37
Qwen3.5-27B	30.43	35.50	32.98	34.00	32.60	25.91	31.49	30.52	31.07	17.81	27.89	30.02
Qwen3.5-35B	36.50	42.60	38.60	40.77	39.00	32.23	37.57	36.15	37.00	23.00	33.89	36.12
Qwen3.5-122B	45.71	50.40	45.21	48.71	46.00	39.20	43.78	43.26	44.32	29.12	41.80	43.41
Qwen3-VL-2B	10.58	14.00	11.57	13.39	11.60	8.54	13.24	11.07	11.12	4.71	9.53	10.85
Qwen3-VL-4B	13.25	16.80	14.88	16.56	14.70	10.91	15.54	13.68	13.66	6.69	11.97	13.51
Qwen3-VL-8B	17.63	21.00	19.34	21.19	18.30	14.71	20.00	18.31	18.04	9.24	15.79	17.60
Qwen3-VL-30B	24.43	27.80	21.98	25.46	16.20	25.58	25.81	13.75	28.30	16.02	29.02	23.12
Qwen3-VL-32B	25.66	28.80	23.31	26.49	17.10	27.26	27.57	15.36	29.91	17.44	30.60	24.50
Qwen3-VL-235B	33.50	36.90	30.83	35.17	23.90	33.98	35.14	21.73	37.64	24.13	38.90	31.98
Gemma3-4B	5.36	6.60	3.55	4.42	5.30	4.23	2.84	4.83	3.00	1.70	5.40	4.29
Gemma3-12B	11.48	13.20	9.42	11.11	12.30	11.02	9.86	11.47	9.74	7.26	11.81	10.79
Gemma3-27B	22.05	18.50	23.39	24.94	24.10	22.85	23.24	22.80	22.07	22.90	23.58	22.77
LLaVA-7B	7.22	8.00	6.94	7.28	7.70	5.51	8.24	6.64	7.61	4.24	6.59	6.91
LLaVA-13B	9.31	10.70	9.34	10.01	9.40	7.88	10.81	9.46	10.09	6.22	8.69	9.26
Gemini-3.1-Pro	26.99	49.70	30.74	43.71	26.40	30.55	22.70	16.57	45.94	16.78	33.39	31.22
GPT-5.4	21.09	40.80	24.96	37.01	21.40	25.44	17.84	13.48	37.29	13.95	27.03	25.48
Claude-Opus-4.6	29.40	55.10	36.20	50.55	30.20	36.02	25.95	19.38	51.41	19.13	39.39	35.70
Human	66.67	77.78	66.67	83.33	61.11	72.22	61.11	50.00	77.78	50.00	72.22	67.17

Table 24: Results for P5: Multi-Match Identity Association measured by precision.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	32.56	0.00	32.27	37.63	39.20	35.55	38.11	38.32	42.59	26.52	36.71	32.68
Qwen3.5-2B	39.15	45.83	36.86	42.29	43.38	39.74	41.69	41.53	46.18	30.00	40.68	40.67
Qwen3.5-4B	43.78	49.97	40.95	46.66	49.57	43.72	45.74	46.13	50.88	34.56	45.23	45.20
Qwen3.5-9B	49.75	55.63	45.56	51.41	54.03	49.48	50.25	50.91	55.36	38.63	50.05	50.10
Qwen3.5-27B	57.70	62.00	53.50	57.87	61.02	55.44	56.69	56.94	61.63	44.97	56.26	56.73
Qwen3.5-35B	63.72	68.90	59.23	64.26	66.83	61.62	62.32	62.60	67.23	51.54	62.36	62.78
Qwen3.5-122B	74.63	77.98	67.02	73.38	74.80	70.06	69.71	70.91	75.35	59.66	71.63	71.38
Qwen3-VL-2B	34.54	39.67	29.38	34.31	35.98	32.12	34.68	33.53	38.17	24.18	33.43	33.64
Qwen3-VL-4B	39.30	43.98	33.55	38.90	40.55	36.24	38.27	37.80	42.27	28.07	37.88	37.89
Qwen3-VL-8B	46.75	51.17	39.38	46.26	47.45	43.08	45.27	44.79	49.75	34.12	45.15	44.83
Qwen3-VL-30B	65.69	71.73	58.35	69.05	58.00	68.52	56.62	52.65	62.37	54.85	54.79	61.15
Qwen3-VL-32B	67.45	72.73	60.85	69.90	58.80	70.12	58.33	54.64	64.28	56.94	56.40	62.77
Qwen3-VL-235B	73.25	78.45	67.81	76.43	65.88	75.16	65.41	61.69	71.27	63.34	64.24	69.36
Gemma3-4B	26.94	29.05	25.91	26.62	27.30	25.28	25.02	26.62	24.00	21.98	26.75	25.95
Gemma3-12B	36.32	38.33	34.87	36.76	37.50	35.29	35.22	36.60	34.56	31.90	36.54	35.81
Gemma3-27B	48.84	47.34	49.79	53.10	51.02	49.72	50.53	49.28	48.07	49.41	50.39	49.77
LLaVA-7B	29.17	31.17	24.64	29.31	30.50	26.08	27.18	28.15	32.02	21.83	28.72	28.07
LLaVA-13B	33.38	36.33	30.39	34.83	34.25	31.70	32.16	33.45	37.06	26.39	33.39	33.03
Gemini-3.1-Pro	68.76	88.30	67.82	85.65	62.02	75.00	57.32	53.31	77.03	49.65	69.38	68.57
GPT-5.4	61.36	81.32	60.63	79.11	55.65	69.12	50.90	48.60	69.47	44.60	62.94	62.15
Claude-Opus-4.6	72.34	92.60	73.62	91.83	66.88	80.77	61.82	58.76	81.40	53.96	75.33	73.57
Human	91.67	97.22	92.59	97.22	88.89	94.44	87.04	85.19	95.37	84.26	94.44	91.67

Table 25: Results for P5: Multi-Match Identity Association measured by recall.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	35.47	0.00	51.24	45.58	37.75	31.86	43.85	39.47	34.38	23.52	32.78	34.17
Qwen3.5-2B	41.63	48.55	54.26	50.00	41.85	35.40	48.04	42.76	38.44	27.24	37.56	42.34
Qwen3.5-4B	46.47	52.05	57.56	53.90	47.30	39.31	51.28	47.42	43.23	31.76	42.30	46.60
Qwen3.5-9B	51.46	57.40	61.03	58.72	51.35	44.78	55.61	52.18	48.76	36.43	47.83	51.41
Qwen3.5-27B	56.90	61.75	66.45	63.28	57.80	50.36	60.07	57.58	54.38	42.55	53.37	56.77
Qwen3.5-35B	61.63	66.50	70.87	68.10	62.75	56.04	64.73	62.07	59.37	48.21	58.59	61.71
Qwen3.5-122B	68.60	72.10	75.41	74.06	68.20	62.46	70.14	68.01	65.82	54.76	65.29	67.71
Qwen3-VL-2B	32.02	37.40	47.19	41.43	34.00	28.25	40.27	34.94	31.30	21.54	30.08	34.40
Qwen3-VL-4B	36.01	41.05	52.23	45.81	38.50	32.37	44.05	39.40	35.13	24.88	34.00	38.49
Qwen3-VL-8B	41.64	45.30	55.58	50.18	42.35	37.57	48.72	44.30	40.17	29.69	38.71	43.11
Qwen3-VL-30B	47.43	49.95	47.07	47.76	37.10	47.94	57.03	34.61	56.77	37.65	66.85	48.20
Qwen3-VL-32B	48.88	50.95	48.84	48.68	37.95	49.64	58.78	36.25	58.10	39.49	68.25	49.62
Qwen3-VL-235B	56.50	58.20	55.45	56.84	45.65	56.19	65.00	43.70	64.27	47.41	73.75	56.63
Gemma3-4B	47.51	50.50	47.27	47.82	48.29	45.33	46.20	47.42	44.00	41.26	47.10	46.61
Gemma3-12B	60.17	62.47	59.32	61.41	61.70	58.56	59.58	60.74	58.38	55.53	60.27	59.83
Gemma3-27B	71.26	62.35	75.19	80.25	76.93	75.60	72.03	72.84	73.07	74.92	76.21	73.70
LLaVA-7B	25.04	26.55	28.68	25.50	26.35	22.35	31.42	24.82	24.73	21.07	24.07	25.51
LLaVA-13B	29.91	31.80	34.63	31.16	29.85	27.50	36.49	30.38	29.77	25.35	28.23	30.46
Gemini-3.1-Pro	48.90	69.00	55.62	64.68	52.05	52.77	48.11	38.46	71.99	41.38	58.65	54.69
GPT-5.4	43.03	62.20	50.91	58.43	47.15	47.28	43.04	34.61	65.62	38.17	52.80	49.39
Claude-Opus-4.6	50.97	73.85	60.54	71.19	55.55	58.39	51.55	41.28	76.69	43.40	63.91	58.85
Human	83.33	88.89	86.11	91.67	86.11	86.11	83.33	77.78	91.67	77.78	88.89	85.61

Table 26: Results for P5: Multi-Match Identity Association measured by F1 score.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	33.95	0.00	39.60	41.22	38.46	33.60	40.78	38.89	38.05	24.93	34.64	33.10
Qwen3.5-2B	40.35	47.15	43.90	45.82	42.60	37.45	44.64	42.13	41.96	28.55	39.06	41.24
Qwen3.5-4B	45.09	50.99	47.86	50.02	48.41	41.40	48.36	46.77	46.74	33.10	43.71	45.68
Qwen3.5-9B	50.59	56.50	52.18	54.82	52.66	47.01	52.79	51.53	51.85	37.50	48.91	50.58
Qwen3.5-27B	57.30	61.87	59.27	60.46	59.36	52.78	58.33	57.26	57.78	43.73	54.78	56.63
Qwen3.5-35B	62.65	67.68	64.53	66.13	64.73	58.70	63.50	62.33	63.06	49.82	60.41	62.14
Qwen3.5-122B	71.49	74.93	70.97	73.72	71.35	66.04	69.92	69.43	70.26	57.11	68.31	69.41
Qwen3-VL-2B	33.23	38.50	36.21	37.54	34.96	30.06	37.27	34.22	34.40	22.78	31.67	33.71
Qwen3-VL-4B	37.58	42.47	40.86	42.07	39.50	34.20	40.96	38.59	38.37	26.38	35.83	37.89
Qwen3-VL-8B	44.05	48.05	46.10	48.14	44.76	40.14	46.93	44.54	44.45	31.75	41.68	43.69
Qwen3-VL-30B	55.09	58.89	52.10	56.46	45.25	56.41	56.82	41.76	59.44	44.65	60.22	53.37
Qwen3-VL-32B	56.69	59.92	54.19	57.39	46.13	58.12	58.56	43.58	61.03	46.64	61.76	54.91
Qwen3-VL-235B	63.79	66.82	61.01	65.20	53.93	64.30	65.20	51.16	67.59	54.23	68.67	61.99
Gemma3-4B	34.38	36.88	33.47	34.20	34.88	32.46	32.46	34.10	31.06	28.68	34.12	33.34
Gemma3-12B	45.30	47.51	43.92	45.99	46.65	44.04	44.27	45.68	43.42	40.52	45.50	44.80
Gemma3-27B	57.96	53.82	59.91	63.91	61.35	59.99	59.39	58.79	57.99	59.55	60.67	59.39
LLaVA-7B	26.95	28.67	26.51	27.27	28.27	24.07	29.15	26.38	27.90	21.44	26.19	26.62
LLaVA-13B	31.55	33.92	32.37	32.89	31.90	29.45	34.19	31.84	33.02	25.86	30.59	31.60
Gemini-3.1-Pro	57.15	77.47	61.12	73.70	56.60	61.95	52.31	44.69	74.42	45.14	63.57	60.74
GPT-5.4	50.58	70.49	55.35	67.21	51.05	56.15	46.64	40.43	67.49	41.13	57.43	54.90
Claude-Opus-4.6	59.80	82.17	66.44	80.21	60.69	67.78	56.22	48.49	78.97	48.11	69.15	65.28
Human	87.30	92.87	89.23	94.36	87.48	90.09	85.14	81.31	93.48	80.89	91.58	88.52

Table 27: Results for P6: Open-Set Rejection on match-present cases (AccMP).

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	4.23	6.20	9.72	10.59	10.40	4.27	15.41	7.38	16.58	9.25	9.96	9.45
Qwen3.5-2B	7.88	10.20	11.04	18.82	15.20	6.68	19.19	11.81	21.78	13.02	14.52	13.65
Qwen3.5-4B	12.64	15.00	12.85	23.97	17.80	14.39	22.16	14.09	25.79	15.85	18.77	17.57
Qwen3.5-9B	19.37	21.40	14.33	31.32	22.00	18.66	26.76	17.45	29.25	20.57	22.38	22.14
Qwen3.5-27B	26.04	30.00	21.75	40.44	26.80	24.59	32.70	22.55	36.29	25.09	28.05	28.57
Qwen3.5-35B	33.15	37.20	26.36	48.68	30.40	30.72	36.76	26.85	41.93	31.13	32.68	34.17
Qwen3.5-122B	43.87	49.00	35.42	60.74	38.20	41.53	49.19	34.63	51.03	37.74	41.11	43.86
Qwen3-VL-2B	10.66	12.40	12.03	21.18	15.20	8.88	19.73	12.21	22.86	12.64	16.36	14.92
Qwen3-VL-4B	14.18	15.60	13.18	25.15	16.60	14.19	23.51	14.90	26.00	16.79	19.58	18.15
Qwen3-VL-8B	21.19	24.00	17.96	33.38	24.80	20.80	27.30	19.60	31.53	22.26	24.06	24.26
Qwen3-VL-30B	36.18	37.00	27.84	47.94	27.20	35.47	34.32	27.52	32.39	26.60	27.20	32.70
Qwen3-VL-32B	37.81	38.00	32.45	48.38	28.00	35.88	34.05	29.26	33.04	29.81	28.12	34.07
Qwen3-VL-235B	42.76	45.20	35.91	55.74	32.20	42.15	38.65	33.42	38.46	31.51	33.18	39.02
Gemma3-4B	16.90	21.00	23.06	20.15	22.40	20.39	18.11	18.93	18.63	22.45	20.65	20.24
Gemma3-12B	18.35	22.40	20.59	21.18	22.40	19.97	18.65	22.42	17.88	23.40	20.96	20.75
Gemma3-27B	18.60	22.40	23.72	20.59	21.00	20.80	19.19	22.15	18.85	24.53	21.19	21.18
LLaVA-7B	18.38	20.40	23.23	19.41	21.80	18.04	16.76	18.52	17.77	21.32	21.03	19.70
LLaVA-13B	17.55	21.80	23.06	19.12	22.20	18.25	18.11	18.93	18.20	22.83	21.92	20.18
Gemini-3.1-Pro	1.11	15.00	2.47	3.82	5.20	2.75	9.46	6.31	0.54	4.91	2.26	4.89
GPT-5.4	0.80	18.80	0.16	3.97	6.80	4.34	2.43	3.49	0.11	5.28	0.11	4.21
Claude-Opus-4.6	1.11	16.20	5.93	3.38	0.80	0.83	15.68	8.46	0.11	4.15	4.25	5.54
Human	61.11	61.11	50.00	77.78	50.00	72.22	55.56	44.44	72.22	44.44	66.67	59.60

Table 28: Results for P6: Open-Set Rejection on match-absent cases (AccMA).

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	1.70	2.80	0.49	2.06	5.80	3.31	2.97	1.47	2.71	2.82	5.86	2.91
Qwen3.5-2B	2.04	3.60	1.65	2.64	8.40	4.13	1.62	2.68	3.14	1.88	6.97	3.52
Qwen3.5-4B	2.13	7.40	2.47	3.82	9.40	7.23	4.32	2.95	4.11	2.07	9.12	5.00
Qwen3.5-9B	2.93	8.60	2.31	3.96	10.80	8.20	4.59	3.35	6.93	3.77	10.80	6.02
Qwen3.5-27B	4.51	11.40	4.12	6.61	13.40	11.85	7.30	4.56	11.47	5.46	11.64	8.39
Qwen3.5-35B	5.87	15.20	3.79	7.64	15.80	14.26	8.11	5.36	13.10	6.21	13.75	9.92
Qwen3.5-122B	10.35	23.00	6.10	11.60	21.40	20.11	12.16	9.25	19.70	9.60	18.61	14.72
Qwen3-VL-2B	2.01	4.00	2.31	2.79	7.40	5.17	1.89	2.55	3.25	2.82	8.16	3.85
Qwen3-VL-4B	2.44	7.00	2.31	3.08	8.60	7.78	3.24	3.22	5.19	3.01	8.89	4.98
Qwen3-VL-8B	4.11	8.80	2.97	3.67	11.20	9.30	4.32	3.22	7.79	2.82	10.92	6.28
Qwen3-VL-30B	12.20	21.80	1.48	6.02	5.80	18.46	10.00	2.55	16.77	9.23	5.82	10.01
Qwen3-VL-32B	12.04	22.60	0.99	6.17	6.80	20.32	9.46	2.95	18.18	9.42	6.32	10.48
Qwen3-VL-235B	13.56	27.40	1.32	9.69	8.00	23.62	12.97	4.69	20.67	12.81	10.07	13.16
Gemma3-4B	3.40	4.80	1.48	0.44	3.60	6.68	6.22	6.30	17.86	2.26	0.00	4.82
Gemma3-12B	4.54	5.60	2.80	0.73	4.80	5.72	7.57	6.03	19.16	2.26	0.00	5.38
Gemma3-27B	4.51	4.20	2.64	1.17	5.20	7.64	8.38	7.10	19.91	1.88	0.57	5.75
LLaVA-7B	2.78	3.40	1.48	0.88	3.80	5.65	5.95	3.62	16.34	0.94	0.73	4.14
LLaVA-13B	2.72	4.80	2.14	0.59	4.20	5.51	5.95	5.23	17.97	1.51	0.61	4.66
Gemini-3.1-Pro	27.39	69.00	38.22	66.81	48.60	71.49	31.08	21.45	44.59	54.24	55.99	48.08
GPT-5.4	45.40	65.60	34.10	76.36	44.20	66.18	33.78	8.85	40.69	61.77	50.33	47.93
Claude-Opus-4.6	15.01	75.80	38.22	63.58	51.80	75.90	23.24	30.29	49.68	44.82	61.20	48.14
Human	83.33	88.89	77.78	94.44	77.78	88.89	83.33	66.67	88.89	72.22	88.89	82.83

Table 29: Results for P7: Resistance to Misleading Suggestions. *Same* and *Diff.* denote same- and different-identity pairs. Acc_N, Acc_S, and Acc_M average the two ground-truth conditions under Neutral, Supportive, and Misleading prompts.

	Neutral (N)			**Supportive (S)**			**Misleading (M)**		
Model	**Same**	**Diff.**	$\mathbf{Acc}_N$	**Same**	**Diff.**	$\mathbf{Acc}_S$	**Same**	**Diff.**	$\mathbf{Acc}_M$
Qwen3.5-0.8B	49.57	48.27	48.92	54.64	55.46	55.05	44.64	43.60	44.12
Qwen3.5-2B	52.75	51.71	52.23	62.59	61.13	61.86	42.63	41.31	41.97
Qwen3.5-4B	54.74	55.97	55.36	69.20	70.43	69.82	38.71	36.53	37.62
Qwen3.5-9B	59.41	58.07	58.74	80.39	79.83	80.11	30.91	28.03	29.47
Qwen3.5-27B	67.71	64.47	66.09	88.03	85.80	86.92	20.48	16.82	18.65
Qwen3.5-35B	72.93	69.57	71.25	92.71	93.63	93.17	12.03	13.46	12.75
Qwen3.5-122B	70.20	75.12	72.66	96.69	95.36	96.03	8.55	4.33	6.44
Qwen3-VL-2B	51.00	51.34	51.17	66.28	67.95	67.12	31.78	30.01	30.90
Qwen3-VL-4B	55.31	54.08	54.70	78.73	77.96	78.35	22.53	19.80	21.17
Qwen3-VL-8B	61.47	57.05	59.26	87.72	88.34	88.03	12.90	14.19	13.55
Qwen3-VL-30B	57.61	60.09	58.85	93.45	92.28	92.87	8.10	5.68	6.89
Qwen3-VL-32B	62.81	60.28	61.55	94.44	95.86	95.15	7.01	3.75	5.38
Qwen3-VL-235B	60.56	65.00	62.78	96.93	96.51	96.72	3.98	1.80	2.89
Gemma3-4B	48.88	48.22	48.55	53.72	54.42	54.07	43.83	44.40	44.12
Gemma3-12B	48.82	47.42	48.12	52.25	51.27	51.76	45.74	44.92	45.33
Gemma3-27B	49.80	48.48	49.14	54.84	55.98	55.41	44.25	44.97	44.61
LLaVA-7B	48.40	48.09	48.25	53.53	52.85	53.19	45.55	44.93	45.24
LLaVA-13B	50.06	48.92	49.49	56.03	56.92	56.48	45.31	44.37	44.84
Gemini-3.1-Pro	67.32	69.88	68.60	95.13	94.09	94.61	10.68	7.16	8.92
GPT-5.4	72.07	66.26	69.17	97.26	97.60	97.43	3.30	1.45	2.38
Claude-Opus-4.6	71.85	75.19	73.52	97.17	95.95	96.56	5.99	3.23	4.61
Human	87.88	87.88	87.88	87.88	87.88	87.88	87.88	86.87	87.37

Table 30: Per-species results for P7 under Neutral prompts.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	49.71	49.40	48.84	49.71	48.30	50.83	49.73	48.06	48.11	47.92	47.55	48.92
Qwen3.5-2B	53.75	55.80	49.34	60.00	48.90	61.02	50.27	47.12	51.68	48.68	48.01	52.23
Qwen3.5-4B	57.52	59.80	52.15	66.03	51.90	62.40	52.16	51.54	55.25	50.75	49.43	55.36
Qwen3.5-9B	61.11	64.00	54.79	68.68	55.71	65.01	55.41	54.23	58.61	54.53	54.02	58.74
Qwen3.5-27B	68.74	72.40	62.21	78.97	62.53	73.00	62.70	61.08	65.01	60.75	59.62	66.09
Qwen3.5-35B	74.17	77.80	67.16	86.18	66.33	79.48	66.76	65.37	71.62	64.91	63.95	71.25
Qwen3.5-122B	75.72	81.20	67.16	88.82	67.73	82.51	67.30	66.57	72.80	65.28	64.18	72.66
Qwen3-VL-2B	53.97	55.40	48.02	61.32	47.90	56.13	48.65	46.58	51.90	46.42	46.59	51.17
Qwen3-VL-4B	57.68	60.80	50.50	65.15	51.11	59.30	51.62	50.07	55.80	50.00	49.66	54.70
Qwen3-VL-8B	63.27	65.80	55.78	71.62	53.51	64.05	55.68	53.83	61.00	54.15	53.18	59.26
Qwen3-VL-30B	63.05	66.40	55.12	72.06	52.90	64.81	54.32	53.02	60.13	52.83	52.76	58.85
Qwen3-VL-32B	65.22	68.60	57.76	75.74	55.72	66.32	57.30	55.30	63.38	56.04	55.63	61.55
Qwen3-VL-235B	67.62	71.00	57.43	78.09	58.11	68.11	57.84	56.10	64.35	56.42	55.52	62.78
Gemma3-4B	47.61	48.80	49.83	49.71	48.29	48.76	45.95	49.13	48.54	49.25	48.20	48.55
Gemma3-12B	48.29	49.20	50.00	49.26	48.29	46.76	45.95	48.59	45.61	48.87	48.51	48.12
Gemma3-27B	49.40	50.20	50.66	50.44	48.50	48.69	46.49	49.53	47.45	49.81	49.39	49.14
LLaVA-7B	48.22	49.20	49.34	49.71	47.90	49.04	46.76	48.05	46.70	47.17	48.66	48.25
LLaVA-13B	49.06	50.00	51.32	50.44	48.70	49.52	47.57	49.53	48.00	49.62	50.65	49.49
Gemini-3.1-Pro	71.15	68.40	61.06	84.85	68.13	73.35	61.08	70.47	64.24	68.68	63.22	68.60
GPT-5.4	66.91	78.20	66.34	80.00	69.94	76.31	56.49	59.87	79.96	62.64	64.18	69.17
Claude-Opus-4.6	75.35	73.20	67.16	89.85	70.33	80.10	67.30	60.67	78.76	65.85	80.19	73.52
Human	88.89	83.33	77.78	100.00	83.33	100.00	83.33	72.22	100.00	77.78	100.00	87.88

Table 31: Per-species results for P7 under Supportive prompts.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	49.21	58.20	53.63	58.38	53.90	56.68	55.95	55.03	55.04	54.15	55.40	55.05
Qwen3.5-2B	63.48	63.20	61.55	65.15	59.52	67.22	59.46	59.06	60.45	60.38	61.00	61.86
Qwen3.5-4B	71.27	74.00	66.67	77.35	68.14	72.80	68.65	66.85	68.80	68.68	64.79	69.82
Qwen3.5-9B	80.29	82.60	77.56	85.00	79.16	83.54	78.38	77.99	81.15	78.87	76.67	80.11
Qwen3.5-27B	88.26	90.00	85.15	90.88	86.37	89.53	86.49	84.83	86.02	85.28	83.26	86.92
Qwen3.5-35B	94.81	94.40	91.91	96.18	92.39	93.94	92.97	91.68	93.06	92.08	91.42	93.17
Qwen3.5-122B	94.50	97.60	95.87	98.09	94.79	97.52	95.41	95.97	96.21	94.91	95.44	96.03
Qwen3-VL-2B	67.81	70.20	67.49	73.09	64.52	67.77	65.95	65.23	68.69	63.21	64.33	67.12
Qwen3-VL-4B	80.14	81.80	75.91	82.50	76.36	79.75	77.84	75.84	78.55	76.60	76.51	78.35
Qwen3-VL-8B	89.34	90.00	86.30	91.62	86.77	90.36	86.49	85.50	88.84	86.98	86.13	88.03
Qwen3-VL-30B	94.56	96.00	92.24	96.91	89.58	94.97	91.35	91.68	92.85	92.08	89.31	92.87
Qwen3-VL-32B	97.34	96.40	93.56	96.76	94.19	95.87	94.32	94.50	95.77	94.34	93.64	95.15
Qwen3-VL-235B	96.97	97.80	95.87	98.68	96.59	97.31	95.68	96.11	96.75	96.23	95.94	96.72
Gemma3-4B	54.12	54.80	49.50	53.82	56.31	53.44	52.43	54.90	56.23	54.15	55.10	54.07
Gemma3-12B	47.85	52.60	51.98	52.79	51.50	51.10	50.54	54.50	50.38	53.02	53.07	51.76
Gemma3-27B	55.95	49.80	56.60	58.09	55.11	57.02	53.51	55.17	53.95	56.98	57.36	55.41
LLaVA-7B	47.95	52.80	55.94	56.03	53.10	52.96	52.16	52.88	53.63	52.26	55.36	53.19
LLaVA-13B	59.16	56.80	60.56	50.15	57.91	55.30	54.59	56.24	55.36	55.85	59.31	56.48
Gemini-3.1-Pro	94.81	94.80	92.90	98.09	94.59	96.42	93.24	95.04	93.50	94.15	93.18	94.61
GPT-5.4	97.22	98.40	97.03	99.71	97.59	99.04	95.41	95.71	99.03	96.23	96.40	97.43
Claude-Opus-4.6	97.19	96.60	95.87	98.24	95.60	97.87	96.22	94.76	97.62	94.72	97.51	96.56
Human	88.89	88.89	77.78	100.00	83.33	100.00	83.33	72.22	94.44	77.78	100.00	87.88

Table 32: Per-species results for P7 under Misleading prompts.

Model	Beluga	Bird	CTai	Giraffes	IPanda	Leo	Lion	NDD	Nyala	Seal	Whale	Ave.
Qwen3.5-0.8B	45.81	44.40	44.22	42.50	42.89	51.17	45.95	41.74	43.88	41.13	41.65	44.12
Qwen3.5-2B	44.42	45.80	39.44	50.59	37.67	51.38	39.19	37.58	42.15	37.36	36.13	41.97
Qwen3.5-4B	41.89	44.20	32.84	48.97	33.27	45.80	33.51	32.75	38.03	31.89	30.69	37.62
Qwen3.5-9B	33.49	38.20	25.58	40.88	24.05	38.15	25.41	23.89	28.82	23.02	22.68	29.47
Qwen3.5-27B	22.21	26.60	15.02	30.88	14.83	24.66	14.32	13.96	17.23	13.21	12.26	18.65
Qwen3.5-35B	14.74	18.60	10.23	22.79	9.82	16.60	9.19	8.86	12.68	8.68	8.05	12.75
Qwen3.5-122B	7.63	9.40	4.95	13.09	4.01	9.37	4.05	4.30	6.39	4.15	3.52	6.44
Qwen3-VL-2B	36.61	39.00	27.72	42.50	26.85	37.33	26.49	24.70	30.55	24.34	23.79	30.90
Qwen3-VL-4B	26.57	29.20	18.98	31.47	17.84	25.41	16.76	15.57	20.26	15.47	15.29	21.17
Qwen3-VL-8B	17.11	19.80	11.72	23.38	11.02	16.46	10.27	9.39	12.78	9.25	7.85	13.55
Qwen3-VL-30B	8.74	10.40	5.45	13.24	5.21	7.99	4.86	4.57	7.37	4.15	3.79	6.89
Qwen3-VL-32B	6.67	7.40	5.78	11.62	4.01	6.75	3.24	3.22	4.99	3.02	2.53	5.38
Qwen3-VL-235B	2.97	3.40	1.98	4.41	3.00	2.96	2.70	3.36	2.93	1.89	2.18	2.89
Gemma3-4B	42.85	44.40	45.21	42.94	48.50	45.04	41.08	44.70	42.15	44.72	43.72	44.12
Gemma3-12B	45.78	49.40	47.36	46.76	44.89	42.29	42.16	45.23	43.01	45.28	46.44	45.33
Gemma3-27B	43.59	46.60	50.99	45.29	43.89	43.53	41.35	44.70	40.95	44.72	45.10	44.61
LLaVA-7B	45.97	45.00	46.20	50.00	45.09	45.59	42.97	44.30	43.66	43.58	45.29	45.24
LLaVA-13B	44.15	45.00	47.36	46.76	43.89	44.97	43.51	44.56	42.69	44.72	45.67	44.84
Gemini-3.1-Pro	11.62	10.20	6.44	16.32	6.02	12.47	4.86	10.74	5.42	8.49	5.56	8.92
GPT-5.4	2.38	3.00	1.65	3.38	1.80	2.62	2.16	2.01	3.58	1.89	1.69	2.38
Claude-Opus-4.6	4.88	4.80	2.81	10.00	2.61	5.79	3.24	2.95	5.20	2.45	5.98	4.61
Human	88.89	83.33	77.78	100.00	83.33	100.00	83.33	66.67	100.00	77.78	100.00	87.37