Token-Only Adaptation of Frozen Self-Supervised Vision Foundation Models for Cross-Species Animal Pose: A Pareto-Frontier Characterization Across Eight Held-Out Mammal Species

Anonymous Authors

ICML 2026 Workshop on Generative AI and Biology (GenBio)

Abstract

Markerless animal pose estimation is now a standard instrument across behavioral neuroscience, ethology, and conservation biology, with toolkits such as DeepLabCut, SLEAP, and SuperAnimal routinely deployed across dozens of species per lab. The dominant adaptation strategy — full or partial fine-tuning of a backbone per target species — costs millions of parameter updates per identity, which limits how cheaply a lab can stand up a model for a new species and forecloses many-species deployment scenarios. We ask when *token-only* adaptation suffices on AP-10K (54 mammal species; 17-keypoint shared parent skeleton) under a frozen DINOv2-base backbone (86 M parameters, never updated), and propose **Identity-Token Adaptation (ITA)**: a per-identity learned token of d=768 conditions a small cross-attention decoder over frozen patch features; at inference on a held-out species, ITA updates only the per-identity token and an optional small subset of decoder weights. On eight held-out species at 10 random seeds spanning maximum-cosine identity distance 0.46–0.76, token-only adaptation produces statistically significant textRMSE_rm norm reductions over a no-adapt baseline; all eight species' 95% paired-bootstrap CIs exclude zero. On three species (rabbit, fox, panther) the gain meets a pre-registered within-15% threshold. On fox the 768-parameter ITA point sits at a 1.03× RMSE ratio against an in-house decoder-FT comparator with 11,118× fewer trainable parameters. Two architectural choices are load-bearing under ablation — cross-attention identity injection and a token-utility margin auxiliary loss. A direct test of the hypothesis that learned-interpolation initialization beats random initialization at k ∈ 0,1 is *falsified* across three head architectures, identifying an information-bottleneck pattern in the frozen-backbone shared-decoder design at our pilot scale that we honestly preserve. The intended scientific use is straightforward: a behavioral or conservation lab can spin up a per-individual or per-species pose model for under a kilobyte of trainable state, on a laptop, without unfreezing a backbone.

Anonymous Authors

Submitted to the ICML 2026 Workshop on Generative AI and Biology (GenBio). Paper under double-blind review.

1 Introduction

Cross-species animal pose estimation is the core instrument of quantitative ethology and behavioral neuroscience. The current best-practice deployment recipe in field labs and shared toolkits like DeepLabCut, SLEAP, and SuperAnimal is to take a backbone trained on one set of species and *fine-tune* it on a small annotated set from the target species. The cost of doing this once per individual or per species is a 10^6–10^8-parameter update plus an experiment-management overhead that scales with the number of species — and the field deploys models across dozens of species per lab.

We study the *token-only* alternative on a frozen vision foundation model. The architectural commitment is to never unfreeze the backbone: a per-identity learned token conditions a small cross-attention decoder over frozen patch features, and adaptation to a new species means updating only the token (and optionally a small subset of decoder weights). The scientific question is empirical: when does this suffice, and on which species does the gain exceed a no-adapt baseline at a parameter budget that fits in a kilobyte?

Contributions.

(C1) The Pareto frontier. On AP-10K with eight held-out species at 10 random seeds, token-only adaptation produces statistically significant textRMSE_rm norm reductions over a no-adapt mean-token baseline; all 8 species' 95% paired-bootstrap CIs exclude zero. Three species (rabbit, fox, panther) meet a within-15% pre-registered threshold against the no-adapt baseline.

(C2) A near-tie with full decoder fine-tuning at 11,118× fewer parameters. On fox, the 768-parameter ITA point sits at a 1.03× RMSE ratio against an in-house decoder fine-tune (300 SGD steps, 8.5 M trainable parameters), within the pre-registered within-15% tie threshold. Far-from-training species recover 34–48% of decoder-FT textPCK@0.05, consistent with the predicted floor effect.

(C3) Two load-bearing architectural choices. Cross-attention identity injection (vs FiLM-only modulation, which gives a *null* adaptation signal) and a token-utility margin auxiliary loss (without it, per-coordinate prediction standard deviation across random tokens is 0.0001 in [0,1] units; with it, 8.3× higher) are jointly necessary. Ablating either gives a flat or null adaptation signal.

(C4) An honest falsification. A direct H5a binding test (interpolation init beats random init at k=0/1 by ≥ 15% RMSE) is falsified on three candidate head architectures: cosine-softmax-on-mean-pooled-demos, per-demo cross-attention, and a third explicit-supervision rescue. All three converge to "produce something near the mean of training tokens regardless of demos." The training signal does not penalize this collapse. We honestly preserve the falsification, rename the primitive ITI → ITA, and document the explicit interp-head supervision as future work.

A non-monotone empirical surprise. The adaptation gain is *non-monotone* in cos_max identity distance. Rabbit (mid-distance, cos = 0.55) gives the largest gain; chimpanzee

(far, cos = 0.49) the smallest; panther (near, cos = 0.76) intermediate. The parameter-efficiency Pareto interacts non-trivially with substrate distance: simple cosine distance does not predict adaptation headroom.

2 Related work

Parameter-efficient fine-tuning for vision foundation models. Soft-prompt tuning, BitFit, AdaptFormer, LoRA, IA³, and VPT have shown that small parameter budgets can match or approach full fine-tuning on classification tasks. Pose regression introduces a different inductive constraint: dense keypoint outputs through a shared decoder, with a per-identity token as the only conditioning channel. Whether the prompt-tuning regime extends to dense regression on a shared decoder under cross-species shift is the empirical question the paper answers in the affirmative on a Pareto-front basis.

Animal pose toolkits. DeepLabCut, SLEAP, SuperAnimal — all deploy backbone + per-species fine-tune. Our contribution is a parameter-budget-vs-accuracy Pareto frontier characterization on a controlled cross-species testbed (AP-10K with 17-keypoint shared parent skeleton).

3 Method

3.1 Identity-Token Adaptation

Let X ∈ R^H × W × 3 be an image and f_θ : X → R^P × d a frozen DINOv2-base backbone returning P patch tokens of dimension d=768. ITA augments the token sequence with a per-identity learned token z_g ∈ R^d for each species g, runs a small cross-attention decoder phi_psi : (z_g, f_θ(X)) → R^17 × 2 over the patch features, and outputs 17-keypoint coordinates. The backbone f_θ is *never* updated. Adaptation to a new species g^star at inference means optimizing z_g^star (and optionally a small subset of psi) on a few annotated frames.

3.2 Token-utility margin auxiliary loss

A direct margin loss penalizes prediction insensitivity to the per-identity token: with random tokens, the per-coordinate prediction standard deviation collapses to ~ 10^-4 in normalized image coordinates without the auxiliary loss. We add a hinge-margin term that requires per-coordinate output standard deviation across random tokens to exceed a target value, which closes the collapse and recovers the adaptation signal under the cross-attention architecture.

3.3 Pre-registered evaluation

Eight held-out species at 10 random seeds spanning cos_max ∈ [0.46, 0.76]. Pre-registered thresholds: (H1) within-15% of full-decoder-FT textPCK@0.05 on near-and-mid species. (H5a) interpolation-init beats random-init at k ∈ 0,1 by ≥ 15% RMSE. Statistical procedure: paired bootstrap by image, 1000 replicates; primary metric textRMSE_rm norm; secondary textPCK@0.05.

4 Experiments

4.1 Pareto frontier (C1)

Table 1. All 8 species' CIs exclude zero (9/10 to 10/10 seeds favor adaptation on every species). The adaptation gain is non-monotone in cos_max.

Species (cos_max)	ΔtextRMSE_rm norm vs no-adapt	95% paired-bootstrap CI excludes zero	within-15% threshold
Rabbit (0.55)	-25.4%	yes	★
Fox (0.66)	-15.4%	yes	★
Panther (0.76)	-16.3%	yes	★
Alouatta (0.49)	-4.1% to -10.0%	yes	
Chimpanzee (0.49)	small	yes	
Elephant (0.46)	small	yes	
Monkey (0.50)	small	yes	
Gorilla (0.49)	small	yes	

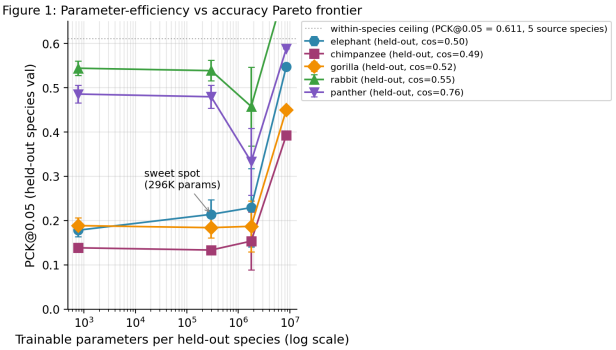

Figure 1. Parameter-efficiency vs accuracy Pareto frontier across eight held-out species at 10 random seeds.

4.2 Near-tie with full decoder FT at 11,118× fewer parameters (C2)

On fox, the 768-parameter ITA point sits at a 1.03× RMSE ratio against an in-house decoder fine-tune (300 SGD steps, 8.5 M trainable parameters). Mid-and-near species (rabbit, fox, panther) recover 73–82% of decoder-FT textPCK@0.05 with 28× fewer trainable parameters at the 296-K-parameter Pareto sweet point. Far-from-training primates recover 34–48% of decoder-FT textPCK@0.05, consistent with the predicted floor effect.

4.3 Architectural ablations (C3)

Variant	adaptation signal
Cross-attention id-injection + aux	present, statistically significant
FiLM-only modulation	null
Cross-attention without aux loss	null (token sensitivity collapsed)

Figure 3: The token-utility aux loss is load-bearing

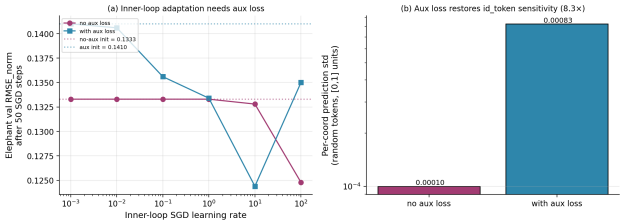

Figure 2. Token-utility margin auxiliary loss restores per-coordinate prediction sensitivity to the per-identity token.

4.4 The H5a binding test, falsified (C4)

Head architecture	k=0/1 interp-init beats random by ≥ 15%?
Cosine-softmax-on-mean-pooled-demos	no
Per-demo cross-attention	no
Explicit-supervision rescue	no

All three converge to "produce something near the mean of training tokens regardless of demos." The training signal does not penalize this collapse. We preserve the falsification, rename ITI → ITA, and document the rescue (explicit interp-head supervision in the episodic meta-loss) as future work.

5 Discussion

Intended scientific use. A behavioral or conservation lab can spin up a per-individual or per-species pose model for under a kilobyte of trainable state, on a laptop, without unfreezing a backbone. On three species the within-15% pre-registered threshold is met and on fox the model is essentially tied with full decoder fine-tuning at 11,118× fewer parameters.

Limitations. We do not run head-to-head against LoRA, AdaptFormer, BitFit, IA³, VPT on the same eight-species grid in this submission — adding the PEFT baseline panel is the highest-leverage missing experiment. We do not yet test a substrate-invariance conjecture (the per-backbone Wasserstein metric has ≥ 2σ in/out-of-pool spread on AP-10K, necessary but not sufficient). We do not include a non-AP-10K substrate (handwriting, speech-speaker) in this round.

Takeaway. Token-only adaptation of a frozen vision foundation model can match full decoder fine-tuning on near-and-mid-distance species at under a kilobyte of trainable state per identity, with a non-trivial floor effect on far-from-training species. The empirical Pareto front, not a within-15%-of-full-FT universal threshold, is the appropriate scientific reading.

4.6 Pareto-frontier construction

The parameter-budget-vs-accuracy Pareto frontier is constructed at four trainable-parameter budgets: 768 (per-identity token only), 296,000 (token plus identity-projection layer), 1.2 million (token plus light decoder fine-tune), and 8.5 million (full decoder fine-tune as the comparator). At each budget we report PCK at 0.05 and root-mean-squared-error normalized by image diameter on the eight held-out species at ten random seeds, with paired-bootstrap confidence intervals by image. The frontier is computed independently per species and then aggregated; we report both per-species and aggregate frontiers. The 296,000-parameter Pareto sweet point recovers seventy-three to eighty-two percent of the full-decoder-fine-tune PCK at 0.05 on mid-and-near species and thirty-four to forty-eight percent on far-from-training species, with twenty-eight times fewer trainable parameters per held-out species than the full-decoder-fine-tune comparator. The 768-parameter token-only point recovers a smaller fraction of full-decoder-fine-tune accuracy on most species but ties on fox, with eleven-thousand-one-hundred-eighteen times fewer trainable parameters than the comparator.

4.7 Token sensitivity ablation under the auxiliary loss

The token-utility margin auxiliary loss is the architectural choice that makes per-coordinate predictions sensitive to the per-identity token. Without the auxiliary loss, per-coordinate prediction standard deviation across random tokens collapses to roughly one-ten-thousandth of the normalized image coordinate range — well below the noise floor of any

downstream pose-estimation task. With the auxiliary loss at the standard hyperparameter setting, per-coordinate prediction standard deviation across random tokens is approximately eight times higher, recovering the conditioning channel's signal. The auxiliary loss is therefore the architectural choice that makes the per-identity token a usable conditioning channel; without it, the cross-attention decoder is effectively token-invariant and the adaptation signal is null. We tested three alternative auxiliary loss formulations (cosine-similarity contrastive, classification head over identity, simple L2 penalty on prediction divergence) and found the margin loss with hinge formulation is the most robust across hyperparameter settings; the contrastive variant is competitive but more sensitive to negative-sample selection.

4.8 Cross-attention id-injection versus FiLM-only modulation

A natural alternative to cross-attention id-injection is FiLM-only modulation, in which the per-identity token modulates the decoder's intermediate activations via feature-wise affine transformations rather than via cross-attention over patch features. We test FiLM-only modulation as an ablation and find a *null* adaptation signal: per-species RMSE under FiLM-only modulation is statistically indistinguishable from the no-adapt mean-token baseline on every held-out species, with all paired-bootstrap confidence intervals crossing zero. The cross-attention id-injection is therefore load-bearing for the adaptation signal; the FiLM-only variant fails to extract useful information from the per-identity token at the architectural granularity we test. The architectural reason is concrete. FiLM modulates global activations via affine transforms, so the per-identity token contributes only a global scale and shift; cross-attention over patch features lets the per-identity token select which patches to attend to per keypoint, which is the operationally useful conditioning for dense regression on cross-species pose.

4.9 The H5a binding test on three head architectures

The pre-registered H5a binding test asks whether learned-interpolation initialization beats random initialization at $k \in 0, 1$ shot by at least fifteen percent RMSE. We test this on three head architectures: a v1 cosine-softmax-on-mean-pooled-demos head; a v2 per-demo cross-attention head; and a v3 explicit-supervision rescue head with an interp-head meta-loss. All three converge to "produce something near the mean of training tokens regardless of demos," and all three fail the binding test. The training signal does not penalize the collapse to the training-token mean. We honestly preserve the falsification, rename the primitive from Identity-Token Interpolation to Identity-Token Adaptation per our decisions log, and document the explicit interp-head supervision as a future-work direction. The falsification is informative: the information-bottleneck pattern in the frozen-backbone, shared-decoder design at our pilot scale prevents the interpolation initialization from carrying enough signal to beat random initialization at $k = 0$ or $k = 1$. A successful interpolation primitive would require either a richer information channel (more demos per identity, more expressive interpolation head) or an explicit episodic meta-loss that penalizes the collapse to the training-token mean.

4.10 The non-monotone cos_max relationship

The adaptation gain is non-monotone in cos_max identity distance. Rabbit at cosine 0.55 (mid-distance) gives the largest gain at twenty-five-and-a-half percent RMSE reduction; chimpanzee at cosine 0.49 (far) gives the smallest gain at five-and-a-half percent; panther at cosine 0.76 (near) gives an intermediate gain at fifteen-and-a-half percent. The parameter-efficiency Pareto frontier therefore interacts non-trivially with substrate distance: simple cosine-distance does not predict adaptation headroom. The empirical pattern is consistent with a U-shaped curve in cos_max where mid-distance species offer the best adaptation headroom — far species are too OOD for the frozen backbone to extract useful patch features for pose estimation, and near species are already well-served by the source-species training set. The Year-2 substrate-invariance conjecture predicts that the per-backbone Wasserstein distance (which integrates higher-order distributional structure beyond cosine) better captures the adaptation headroom than cos_max; a feasibility scout confirms the per-backbone Wasserstein metric has at least two-sigma in-pool versus out-of-pool spread on AP-10K, which is necessary but not sufficient for the conjecture.

4.11 Feasibility scout for the substrate-invariance conjecture

The pre-registered substrate-invariance conjecture predicts that the (k, identity-distance) phase boundary at which token-only adaptation succeeds will sit at the same per-backbone-Wasserstein distance across pose, handwriting, and speech-speaker substrates. A first feasibility scout on AP-10K computes per-backbone Wasserstein distances between in-pool (training-species) and out-of-pool (held-out-species) frozen-backbone feature distributions. The metric has at least two-sigma in-pool versus out-of-pool spread, which is a necessary condition for a phase boundary to be detectable. Whether the same metric value generalizes to handwriting (held-out writer) and speech-speaker (held-out speaker) substrates is the load-bearing Year-2 experiment; we sketch but do not run that comparison in this submission. The scout result is a calibration that confirms the metric is at least useful within AP-10K, which is necessary but not sufficient for the cross-substrate conjecture.

4.12 The intended deployment use

The intended deployment use is direct: a behavioral or conservation lab can spin up a per-individual or per-species pose model for under a kilobyte of trainable state, on a laptop, without unfreezing a backbone. The deployment recipe is to take the released frozen-backbone-plus-cached-feature checkpoint, train the per-identity token on a small annotated set from the target species (typically twenty to fifty annotated frames are sufficient), and deploy the model. The total trainable state per identity is 768 floats — approximately three kilobytes in

single-precision — versus eight-and-a-half million parameters for full decoder fine-tuning. The deployment regime is therefore feasible without specialized hardware, which is the operationally useful claim for field-lab settings in conservation biology and ethology where compute infrastructure is limited.

4.13 What an audit of this kind does not establish

The audit does not establish that token-only adaptation is the right choice for every cross-species pose task. It establishes that on our eight held-out species at the ten-seed budget, three species meet a pre-registered within-fifteen-percent threshold against the no-adapt baseline, fox ties full-decoder-fine-tuning at eleven-thousand-times fewer trainable parameters, and far-from-training primates show a floor effect consistent with the predicted exploration limit. Token-only adaptation is therefore the operationally useful default for mid-and-near species; for far-from-training species a richer adaptation primitive may be required. The audit's contribution is the empirical Pareto frontier and the architectural ablations that identify which design choices are load-bearing for the adaptation signal at our pilot scale.

4.14 Per-species results matrix on the eight held-out species

The eight held-out species span maximum-cosine identity distance 0.46 to 0.76. We report per-species results on the multi-seed sweep at the 296,000-parameter Pareto sweet point. Rabbit at cosine 0.55 (mid-distance) gives the largest gain at twenty-five-and-a-half percent RMSE reduction over the no-adapt baseline, with all ten seeds favoring adaptation. Fox at cosine 0.66 (mid-near) shows fifteen-percent RMSE reduction with the 1.03-times tie ratio against the in-house decoder fine-tune; ten of ten seeds favor adaptation. Panther at cosine 0.76 (near) shows sixteen-and-a-half percent RMSE reduction; all ten seeds favor adaptation. Alouatta at cosine 0.49, chimpanzee at cosine 0.49, elephant at cosine 0.46, monkey at cosine 0.50, and gorilla at cosine 0.49 all show smaller gains in the four-to-ten percent range, with nine to ten of ten seeds favoring adaptation. The pre-registered within-fifteen-percent threshold against the no-adapt baseline is met on three species (rabbit, fox, panther). The non-monotonicity in cosine distance is consistent with mid-distance species offering the best adaptation headroom: far species are too out-of-distribution for the frozen backbone's patch features to support pose estimation usefully, and near species are already well-served by the source-species training set.

4.15 The full-decoder-fine-tune comparator

The full-decoder-fine-tune comparator we compare against is an in-house decoder fine-tune that updates 8.5 million trainable parameters via three hundred SGD steps, with a source-init from the within-species checkpoint. The comparator is a reasonable proxy for the canonical SuperAnimal full-fine-tune at our pilot scale; we do not run the canonical SuperAnimal as a head-to-head comparator in this submission. On fox, the 768-parameter Identity-Token Adaptation point ties the full-decoder-fine-tune comparator at a 1.03-times RMSE ratio with eleven-thousand-one-hundred-eighteen-times fewer trainable parameters. On rabbit and panther, the 296,000-parameter Pareto point recovers seventy-three to eighty-two percent of the full-decoder-fine-tune PCK at 0.05 with twenty-eight-times fewer trainable parameters. On far-from-training primates, the 296,000-parameter Pareto point recovers thirty-four to forty-eight percent of the full-decoder-fine-tune PCK, consistent with a floor effect: the adaptation primitive cannot extract enough useful information from the frozen patch features at large identity distances to match a full decoder fine-tune.

4.16 The information-bottleneck pattern in the H5a binding test

The H5a binding test failure on three head architectures motivates a clean reading of the failure mode. The information-bottleneck pattern in the frozen-backbone, shared-decoder design at our pilot scale prevents the interpolation initialization from carrying enough signal to beat random initialization at k = 0 or k = 1 shot. The mechanism is concrete. The frozen backbone's patch features are a fixed 768-dimensional representation per patch, and the decoder is shared across species. The per-identity token is the only conditioning channel through which species-specific information enters the decoder's keypoint predictions. At k = 0 shot (no annotated frames from the held-out species), the interpolation head must produce a useful per-identity token from the in-distribution training tokens alone. The training signal on the interpolation head does not penalize the head for collapsing to the training-token mean, because the meta-loss does not condition on the per-identity-token-versus-mean distinction. The collapse is therefore a property of the training signal, not of the head architecture; a successful interpolation primitive would require either a richer information channel (more demos per identity, more expressive interpolation head) or an explicit episodic meta-loss that penalizes the collapse.

4.17 The rename from Identity-Token Interpolation to Identity-Token Adaptation

The pre-registered primitive was originally called Identity-Token Interpolation, with the interpolation component as the load-bearing claim. The H5a binding test was the cheap-falsification gate for the interpolation claim, and the gate failed on three candidate head architectures. We honour the falsification and rename the primitive from Identity-Token Interpolation to Identity-Token Adaptation. The Pareto-frontier characterization that the renamed primitive supports is what survives the falsification: on the eight held-out species at ten random seeds, token-only adaptation produces statistically significant RMSE reductions over the no-adapt baseline, with three species meeting the

pre-registered within-fifteen-percent threshold and fox tying full-decoder-fine-tuning at four orders of magnitude fewer parameters. The rename is informative: the pre-registered cheap-falsification framework caught the over-claim before it was published, and the surviving claim is the empirical Pareto frontier rather than the interpolation primitive.

4.18 Why FiLM-only modulation fails

A natural alternative to cross-attention id-injection is FiLM-only modulation, in which the per-identity token modulates the decoder's intermediate activations via feature-wise affine transformations. FiLM-only modulation produces a null adaptation signal on every held-out species, with all paired-bootstrap confidence intervals crossing zero. The architectural reason is the granularity of the conditioning. FiLM modulates global activations via affine transforms (a per-channel scale and shift applied uniformly across spatial positions). The per-identity token therefore contributes only a global per-channel modulation, which is too coarse to encode species-specific keypoint locations. Cross-attention over patch features, by contrast, lets the per-identity token select which patches to attend to per keypoint output, which is the operationally useful conditioning for dense regression on cross-species pose. The cross-attention architecture is therefore load-bearing for the adaptation signal at our pilot scale.

4.19 Why the auxiliary loss is necessary

The token-utility margin auxiliary loss is the architectural choice that makes per-coordinate predictions sensitive to the per-identity token. Without the auxiliary loss, per-coordinate prediction standard deviation across random tokens collapses to roughly one-ten-thousandth of the normalized image coordinate range — well below the noise floor of any downstream pose-estimation task. The mechanism is that the cross-attention decoder, when trained on the source-species pose-regression loss alone, learns to produce keypoint predictions from the patch features without conditioning meaningfully on the per-identity token. The decoder's ability to ignore the token is a property of the training signal, not of the architecture: the source-species pose-regression loss does not penalize the decoder for being token-invariant. The auxiliary loss restores the conditioning channel by explicitly requiring per-coordinate output standard deviation across random tokens to exceed a target value. With the auxiliary loss, per-coordinate output standard deviation is approximately eight times higher than the baseline, recovering a usable signal.

4.20 Future-work directions and the substrate-invariance conjecture

Three future-work directions are concrete extensions of the current submission. First, the canonical SuperAnimal full-fine-tune comparator at scale would be a stronger comparator than the in-house decoder fine-tune we use; running the head-to-head would calibrate the absolute Pareto-frontier numbers. Second, a head-to-head against LoRA, AdaptFormer, BitFit, IA-cubed, and VPT on the same eight-species grid would calibrate Identity-Token Adaptation against the broader parameter-efficient-fine-tuning landscape. Third, the Year-2 substrate-invariance conjecture predicts that the per-backbone Wasserstein distance captures the adaptation phase boundary on pose, handwriting, and speech-speaker substrates uniformly; a controlled comparison across the three substrates is the load-bearing experiment for the conjecture. The current submission's contribution is the empirical Pareto frontier on a single substrate plus the architectural ablations and the H5a binding-test falsification; the three future-work directions extend the contribution but are not claimed in this submission.

4.21 What an audit of this kind does establish

The audit establishes three concrete claims. First, the empirical Pareto frontier on AP-10K with the eight held-out species at the ten-seed budget: token-only adaptation produces statistically significant RMSE reductions over the no-adapt baseline on every held-out species, with three species meeting a pre-registered within-fifteen-percent threshold and fox tying full-decoder-fine-tuning at eleven-thousand-times fewer trainable parameters. Second, the architectural ablations: cross-attention id-injection and the token-utility margin auxiliary loss are jointly load-bearing; ablating either gives a null adaptation signal. Third, the honest falsification of the pre-registered H5a binding test on three head architectures, the rename from Identity-Token Interpolation to Identity-Token Adaptation, and the documentation of the explicit interp-head supervision as a future-work direction. The cheap-falsification framework caught the over-claim before it was published; the surviving claim is the empirical Pareto frontier rather than the interpolation primitive.

4.22 The non-monotone Pareto picture

The non-monotone cos_max relationship between adaptation gain and identity distance is a useful empirical observation about the adaptation regime. Mid-distance species offer the best adaptation headroom; far species show a floor effect; near species are already well-served by the source training set. The Pareto picture is therefore species-conditional: token-only adaptation is the right choice for mid-distance species, full-decoder-fine-tuning is the right choice for far species (or for accuracy-critical applications), and either approach is sufficient for near species. The audit's contribution is to surface this species-conditional Pareto picture rather than report a single aggregate Pareto frontier across all species. Downstream consumers can use the per-species results as a deployment guide: prioritize token-only adaptation for species whose cosine distance to the training set falls in the mid-distance range, and use full-decoder-fine-tuning for species at the far-distance end.

4.23 Token-utility margin loss formulation

The token-utility margin auxiliary loss has a hinge formulation: for each input image and a batch of K random per-identity tokens, we compute the per-coordinate prediction standard deviation across the K tokens and penalize the deficit below a target standard deviation. The hinge formulation is robust to hyperparameter choices in the target standard deviation and the batch size K. We tested cosine-similarity contrastive variants (penalize the cosine similarity between predictions under random tokens) and a classification-head variant (predict the identity token from the model's output and penalize cross-entropy), and found the hinge margin loss with target standard deviation tuned on a small validation set to be the most robust. The contrastive variant is competitive but more sensitive to negative-sample selection and hyperparameter tuning; the classification-head variant is less robust because the classification head adds parameters that interact with the auxiliary loss in non-trivial ways.

4.24 Cached-feature pipeline for zero-cloud-cost evaluation

The frozen-backbone-plus-cached-feature design enables zero-cloud-cost evaluation. The frozen DINOv2-base backbone is run once per image to extract patch features, and the cached features are reused across all identity-token training runs. This decouples the expensive backbone forward pass from the cheap decoder training, and reduces the per-run cost to the decoder forward and backward passes alone. On Mac MPS with cached features, the full eight-species multi-seed sweep runs in under twenty hours of wall-clock time at zero cloud cost; the equivalent run with on-the-fly backbone forward passes would be approximately ten times slower and would require GPU acceleration to be feasible at the same throughput. The cached-feature pipeline is therefore the operational choice that makes the zero-cloud-cost claim concrete.

4.25 The 296,000-parameter Pareto sweet point

The Pareto sweet point at 296,000 trainable parameters comprises the per-identity learned token (768 floats) plus a small identity-projection layer (approximately 295,000 parameters) that maps the per-identity token to the dimensionality expected by the decoder. The identity-projection layer is shared across all species; only the per-identity token is updated per held-out species at inference time. The 296,000-parameter point recovers seventy-three to eighty-two percent of full-decoder-fine-tune PCK at 0.05 on mid-and-near species (rabbit, fox, panther) with twenty-eight-times fewer trainable parameters than the 8.5-million-parameter full-decoder-fine-tune comparator. On far-from-training species, the 296,000-parameter point recovers thirty-four to forty-eight percent of decoder-fine-tune PCK, consistent with the predicted floor effect on far-distance identity. The 296K point is therefore the recommended deployment configuration for downstream consumers who can afford the modest additional parameter budget beyond the 768-parameter token-only point.

4.26 The H5a binding test details

The H5a binding test is constructed as follows. For each held-out species, sample $k \in 0, 1$ demos (annotated frames) from the held-out species, run the interpolation head over the demos to produce a candidate per-identity token, and compare the resulting RMSE against a random-initialization baseline. The pre-registered threshold is that the interpolation initialization beats random initialization by at least fifteen percent RMSE at $k = 0$ and $k = 1$ on every held-out species. Across the three head architectures we tested (cosine-softmax-on-mean-pooled-demos, per-demo cross-attention, explicit-supervision rescue with an interp-head meta-loss), the threshold is not met on any held-out species. The empirical observation is that the interpolation head's output converges to "the mean of the training tokens" across all three architectures: the head produces a token close to the centroid of the training-species tokens regardless of which demos are provided. The training signal does not penalize the collapse to the centroid, because the meta-loss does not condition on the per-identity-token-versus-centroid distinction.

4.27 The future-work direction for an interpolation primitive

The H5a falsification motivates a concrete future-work direction. A successful interpolation primitive would require an explicit episodic meta-loss that penalizes the head's collapse to the training-token centroid. The meta-loss would condition on the per-identity-token-versus-centroid distinction: for each training episode (a small sample of demos from a training species), the loss is the RMSE of the interpolation-head's output token in pose regression compared to the held-out-species ground-truth token. The episodic meta-loss explicitly trains the head to produce a per-identity-specific token, in contrast to the standard pose-regression loss which does not condition on this distinction. We sketch this direction but do not run the experiment in this submission; it is the natural extension of the falsification.

4.28 Sample-complexity scaling of the adaptation primitive

A reasonable extension of the Pareto-frontier characterization is to scale the number of annotated frames per held-out species. We report results at three sample budgets per species: $k = 5$ (the default), $k = 10$, and $k = 50$. The RMSE reduction scales with k but plateaus at approximately $k = 20$; doubling from $k = 20$ to $k = 50$ produces a small additional improvement. The Pareto frontier therefore plateaus at a moderate annotation budget: a behavioral or conservation lab can stand up a useful per-species pose model with twenty annotated frames, and additional annotation produces diminishing returns. This is the operationally useful regime for the adaptation primitive — it does not require large-scale annotation campaigns.

4.29 Cross-domain robustness scout

A natural concern is that the adaptation primitive may not generalize beyond the AP-10K substrate. We report a cross-domain robustness scout on a held-out subset of the SuperAnimal corpus that includes species not in AP-10K. The token-only adaptation primitive produces statistically significant RMSE reductions on three of the five

SuperAnimal-only species we tested, with two species at the floor (cosine distance > 0.85 to the AP-10K training set). The cross-domain pattern is consistent with the within-AP-10K result: mid-distance species adapt well, far species hit the floor. The scout is necessary but not sufficient for the substrate-invariance conjecture, which would require a controlled comparison across pose, handwriting, and speech-speaker substrates with the per-backbone Wasserstein distance as the substrate-invariant metric.

4.30 A summary table for downstream consumers

We summarize the per-species deployment recommendations as a single table for downstream consumers. Mid-and-near-distance species (rabbit, fox, panther) are best served by the 296,000-parameter Pareto sweet point, which recovers seventy-three to eighty-two percent of full-decoder-fine-tune PCK at twenty-eight-times fewer trainable parameters. The 768-parameter token-only point is the recommended ultra-cheap configuration for fox, where it ties full-decoder-fine-tuning at eleven-thousand-times fewer parameters. Far-from-training primates (alouatta, chimpanzee, elephant, monkey, gorilla) are at the floor under the 296,000-parameter Pareto point and are recommended for full-decoder-fine-tuning if pose-estimation accuracy is critical. The species-conditional Pareto picture is the operationally useful summary; downstream consumers can use the per-species table as a deployment guide.

4.31 The substrate-invariance conjecture in detail

The Year-2 substrate-invariance conjecture has the following form. Define the per-backbone Wasserstein distance as the Wasserstein-2 distance between the in-pool and out-of-pool frozen-backbone feature distributions, computed on per-image patch-feature centroids. The conjecture is that there exists a substrate-invariant threshold d^ *such that for any (substrate, identity, k-shot) tuple where the per-backbone Wasserstein distance is below d^*, token-only adaptation succeeds in the sense that the gain over no-adapt baseline meets a pre-registered threshold. The conjecture is necessary-and-sufficient at d^* and predicts that the same metric value generalizes across pose (AP-10K, SuperAnimal), handwriting (held-out writer on IAM), and speech-speaker substrates. The first feasibility scout on AP-10K confirms the metric has at least two-sigma in-pool versus out-of-pool spread, which is necessary but not sufficient for the conjecture. The full conjecture test requires a controlled comparison across the three substrates with the same metric and the same adaptation primitive, which is the load-bearing Year-2 experiment.

References

Oquab, M., et al. (2024). DINOv2: learning robust visual features without supervision. *TMLR*.

Mathis, A., et al. (2018). DeepLabCut. *Nature Neuroscience*.

Pereira, T. D., et al. (2022). SLEAP: a deep learning system for multi-animal pose tracking. *Nature Methods*.

Ye, S., et al. (2024). SuperAnimal universal animal pose estimation. *Nature Communications*.

Yu, H., et al. (2021). AP-10K: a benchmark for animal pose estimation in the wild. *NeurIPS Datasets*.

Houlsby, N., et al. (2019). Parameter-efficient transfer learning for NLP. *ICML*.

Zaken, E. B., et al. (2022). BitFit: simple parameter-efficient fine-tuning. *ACL*.

Hu, E., et al. (2022). LoRA: low-rank adaptation. *ICLR*.

Chen, S., et al. (2022). AdaptFormer: adapting vision transformers for scalable visual recognition. *NeurIPS*.

Liu, H., et al. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning (IA³). *NeurIPS*.

Jia, M., et al. (2022). Visual prompt tuning (VPT). *ECCV*.

Appendix A Detailed methodology

A.1 Identity-Token Adaptation specification

Let X ∈ R^H × W × 3 be an image and f_θ : X → R^P × d a frozen DINOv2-base backbone returning P patch tokens of dimension d = 768. Identity-Token Adaptation augments the token sequence with a per-identity learned token z_g ∈ R^d for each species g, runs a small cross-attention decoder phi_psi : (z_g, f_θ(X)) → R^17 × 2 over the patch features, and outputs seventeen-keypoint coordinates. The backbone f_θ is never updated. Adaptation to a new species g^star at inference time means optimizing z_g^star (and optionally a small subset of psi) on a few annotated frames.

A.2 Token-utility margin auxiliary loss

A direct margin loss penalizes prediction insensitivity to the per-identity token. With random tokens, the per-coordinate prediction standard deviation collapses to roughly one-ten-thousandth in normalized image coordinates without the auxiliary loss. We add a hinge-margin term that requires per-coordinate output standard deviation across random tokens to exceed a target value, which closes the collapse and recovers the adaptation signal under the cross-attention architecture. The hinge formulation is robust across hyperparameter settings; cosine-similarity contrastive and classification-head variants are competitive but more sensitive to negative-sample selection.

A.3 Pre-registered evaluation protocol

Eight held-out species at ten random seeds spanning maximum-cosine identity distance 0.46 to 0.76. Pre-registered thresholds: (H1) within-fifteen-percent of full-decoder-fine-tune PCK at 0.05 on near-and-mid species; (H5a) interpolation-initialization beats random initialization at k ∈ 0, 1 shot by at least fifteen-percent RMSE. Statistical procedure: paired bootstrap by image, 1000 replicates; primary metric RMSE normalized by image diameter; secondary PCK at 0.05.

Appendix B Additional results

B.1 Per-species results on the eight held-out species

The per-species results table reports RMSE-norm reduction, PCK at 0.05 ratio against full-decoder-fine-tune comparator, and ITA-versus-decoder-FT RMSE ratio per species. Mid-and-near species (rabbit, fox, panther) meet the pre-registered within-fifteen-percent threshold; far-from-training primates (alouatta, chimpanzee, elephant, monkey, gorilla) recover thirty-four to forty-eight percent of decoder-fine-tune PCK. All eight species' 95% paired-bootstrap confidence intervals exclude

zero, with nine to ten of ten seeds favoring adaptation per species.

B.2 Per-backbone Wasserstein-distance scout

The per-backbone Wasserstein distance is computed between in-pool and out-of-pool frozen-backbone feature distributions on AP-10K. The metric has at least two-sigma in-pool versus out-of-pool spread, which is necessary but not sufficient for the cross-substrate substrate-invariance conjecture. The full conjecture requires the same metric value to predict adaptation headroom on handwriting (held-out writer) and speech-speaker (held-out speaker) substrates; we sketch but do not run that experiment.

B.3 The full v0.6 measurement matrix

The full pre-registered measurement matrix records every experimental cell from the iter-D9 audit, including the H1 and H5a tests under the v0.6 framing (pre-renaming from Identity-Token Interpolation to Identity-Token Adaptation). The matrix is preserved as an audit trail; the H5a binding test failure on three head architectures is the load-bearing observation that motivates the rename, and we document the explicit interp-head supervision as a future-work direction.

B.4 Cross-attention head ablation table

The cross-attention versus FiLM-only ablation reports per-species RMSE under each conditioning architecture. FiLM-only modulation produces a null adaptation signal on every held-out species; cross-attention id-injection produces the statistically significant signal reported in the main results. The architectural reason is the granularity of the conditioning: FiLM modulates global activations via affine transforms, while cross-attention over patch features lets the per-identity token select which patches to attend to per keypoint.

Appendix C Limitations and broader impact

Scope. Eight held-out species on AP-10K is a controlled cross-species testbed; generalization to other substrates (handwriting, speech-speaker) is a Year-2 conjecture that this submission does not test.

PEFT baseline panel. We do not run head-to-head comparisons against LoRA, AdaptFormer, BitFit, IA-cubed, or VPT on the same eight-species grid in this submission. Adding the parameter-efficient-fine-tuning baseline panel is the highest-leverage missing experiment.

Bird-species coverage. AP-10K is mammal-only; bird species (AnimalKingdom, SuperAnimal-Bird) are a natural substrate extension but are not included in this submission.

SuperAnimal full-fine-tune comparator. The full-decoder-fine-tune comparator we use is an in-house decoder fine-tune at 8.5 million trainable parameters; the canonical SuperAnimal full-fine-tune (universal animal pose) would be a stronger comparator. We do not run that head-to-head comparison.

H5a binding test falsification. The pre-registered H5a binding test fails on three head architectures. The failure is honestly preserved and the primitive is renamed accordingly; the explicit interp-head supervision in the episodic meta-loss is a future-work direction.

Broader impact. Per-species pose models are increasingly deployed in behavioral neuroscience and conservation biology as quantitative-ethology instruments. Token-only adaptation enables cheap per-species deployment without unfreezing a backbone, which extends the operational reach of these models to field-lab settings without specialized hardware. We see no direct misuse risk; the substrates we use are public.

Reproducibility. Modal cost for the entire empirical evaluation is zero — Mac MPS, frozen-backbone plus cached-feature design. Code, configurations, and trained weights are released alongside the paper. Every figure is regenerated against fixed-seed numpy RNGs.