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

Ethan Y. Wang [*1] Aayan Alwani [*1]

Abstract

Markerless animal pose estimation is a standard tool in behavioral neuroscience, ethology, and conservation biology. Toolkits such as DeepLabCut, SLEAP, and SuperAnimal are routinely deployed across dozens of species, but adapting them to a new species requires fine-tuning millions of backbone parameters—an expensive step that must be repeated for every target species. We ask how far a much cheaper alternative can go: freeze the backbone entirely and adapt with a single learned token.

We propose Identity-Token Adaptation (ITA). A frozen DINOv2-base backbone (86 M parameters, never updated) produces patch features; a per-species learned token (d=768) conditions a small cross-attention decoder over those features. Adapting to a new species means optimizing only that token and, optionally, a small projection layer.

We evaluate on AP-10K (54 mammal species, shared 17-keypoint skeleton). Across eight held-out species at 10 seeds, token-only adaptation significantly reduces normalized keypoint error on every species; all eight 95% paired-bootstrap confidence intervals exclude zero. Three species (rabbit, fox, panther) meet a pre-registered 15% improvement threshold. On fox, the 768-parameter token achieves a 1.03× error ratio against an 8.5 M-parameter decoder fine-tune—essentially a tie at 11,118× fewer trainable parameters. Against standard PEFT baselines (BitFit, LoRA) on the same grid, ITA achieves comparable accuracy at 19–114× fewer parameters.

Two design choices are necessary: cross-attention identity injection and a token-utility auxiliary loss. Removing either collapses the adaptation signal. A pre-registered hypothesis—that interpolating over training-species tokens would beat random initialization—is falsified across three architectures, and we report this negative result. The practical upshot: a lab can stand up a per-species pose model from 768 trainable floats (≈ 3 KB), on a laptop, without unfreezing a backbone.

1. Introduction

Cross-species animal pose estimation underpins quantitative ethology and behavioral neuroscience. The standard workflow—fine-tune a backbone on labeled frames from each new species (Mathis et al., 2018; Pereira et al., 2022; Ye et al., 2024)—requires 10^6–10^8 parameter updates per species, plus the overhead of training, versioning, and storing a separate model for each one.

We test a much cheaper alternative. We freeze a self-supervised backbone (DINOv2-base, 86 M parameters) and never update it. A small cross-attention decoder reads the frozen features and predicts keypoints, conditioned on a single learned token per species. Adapting to a new species means optimizing only that token—768 numbers, about 3 KB. The question is empirical: on which species does this suffice, and where does it fall short?

Contributions. (C1) A parameter-versus-accuracy frontier. On AP-10K (Yu et al., 2021) with eight held-out species at 10 seeds, token-only adaptation reduces normalized keypoint error on every species, with all eight 95% paired-bootstrap intervals excluding zero. Three species (rabbit, fox, panther) meet a

[*]Equal contribution [1]Great Neck South High School. Correspondence to: Ethan Y. Wang <ethanwang202511@gmail.com>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)

pre-registered 15% improvement threshold.

(C2) A near-tie with full fine-tuning at 11,118× fewer parameters. On fox, the 768-parameter token reaches a 1.03× error ratio against an 8.5 M-parameter decoder fine-tune. A hyperparameter sweep of the comparator (Appendix) rules out under-training: the best-tuned ratio is 1.07×, still within the 15% threshold. Far-from-training species recover only 34–48% of the fine-tune's accuracy, consistent with a predicted floor effect.

(C3) Two necessary design choices. Cross-attention identity injection and a token-utility auxiliary loss are both required. Replacing cross-attention with simpler feature-wise modulation gives no adaptation signal. Removing the auxiliary loss lets the decoder ignore the token entirely. Both must be present.

(C4) A negative result, preserved. We pre-registered a stronger hypothesis: that interpolating over training-species tokens would produce a good initialization, beating random initialization with zero or one example. This failed across three architectures—all collapsed to outputting the average token regardless of input. We report the failure rather than drop it.

A non-monotone surprise. The improvement is not monotone in distance to the training pool. Rabbit (mid-distance) gains the most (−25.4%); the farthest species gain the least (4–6%); the nearest species fall in between. Cosine distance does not predict where token-only adaptation pays off.

2. Related Work

Parameter-efficient fine-tuning for vision. Methods like BitFit (Ben Zaken et al., 2022), LoRA (Hu et al., 2022), and visual prompt tuning (Jia et al., 2022) have shown that small parameter budgets can approach full fine-tuning on classification tasks. Most of this evidence is for classification, where the output is a single label. Pose estimation is dense regression—17 keypoint coordinates through a shared decoder—and our only conditioning channel is one token. Whether the prompt-tuning regime survives this setting, under cross-species shift, is the gap we fill.

Animal pose toolkits. DeepLabCut (Mathis et al., 2018), SLEAP (Pereira et al., 2022), and SuperAnimal (Ye et al., 2024) all adapt to new species by fine-tuning backbone weights. We hold the backbone fixed and map out where a single token suffices instead, on a controlled benchmark (AP-10K, 17-keypoint shared skeleton).

3. Method

3.1. Identity-Token Adaptation

An image X is passed through a frozen DINOv2-base backbone (Oquab et al., 2024) (f_θ; 86 M parameters, never updated), which returns P=256 patch features of dimension d=768. Each species g has a learned token $z_g \in \mathbb{R}^d$. A small cross-attention decoder ϕ_ψ takes the patch features and the species token as input and outputs 17 keypoint coordinates. The decoder (∼7.9 M params) consists of an input projection, a 2-layer transformer encoder, an identity-projection layer that injects the species token, 17 learned keypoint queries, two cross-attention blocks, and a coordinate head. The backbone is never updated—not during training, not during adaptation.

Adaptation budgets. To adapt to a new species $g^\star$, we optimize a subset of the model on k labeled frames. We define four nested budgets: 768 (token only); 296 K (token + identity-projection layer); 1.78 M (token + first cross-attention block); and 8.5 M (full decoder fine-tune, used as an upper-bound comparator). The 768-parameter token-only point is the headline: a per-species model that fits in 3 KB. Figure 1 shows the architecture and these four budgets.

3.2. Keeping the Decoder Sensitive to the Token

Without intervention, the decoder learns to predict keypoints from the patch features alone and ignores the species token. When this happens, swapping in a random token barely changes the output—the per-coordinate standard deviation across random tokens collapses to $\sim 10^{-4}$ in normalized coordinates. We fix this with a hinge-margin auxiliary loss: during training, we require the per-coordinate output spread across random tokens to stay above a target threshold. This forces the decoder to actually use the token as a conditioning signal.

3.3. Evaluation Setup

We hold out eight species spanning a range of distance to the training pool ($\cos_{\max} \in [0.46, 0.76]$) and evaluate at 10 random seeds each.

Metrics. $\mathrm{RMSE}_{\mathrm{norm}}$: root-mean-squared keypoint error normalized by image diameter (lower is better). PCK@0.05: fraction of keypoints within 0.05 image-diameters of ground truth (higher is better). $\cos_{\max}$: maximum cosine similarity between a held-out species' mean feature vector and any training species' mean (higher = closer to training data).

Pre-registered thresholds. (H1) Token-only adapta-

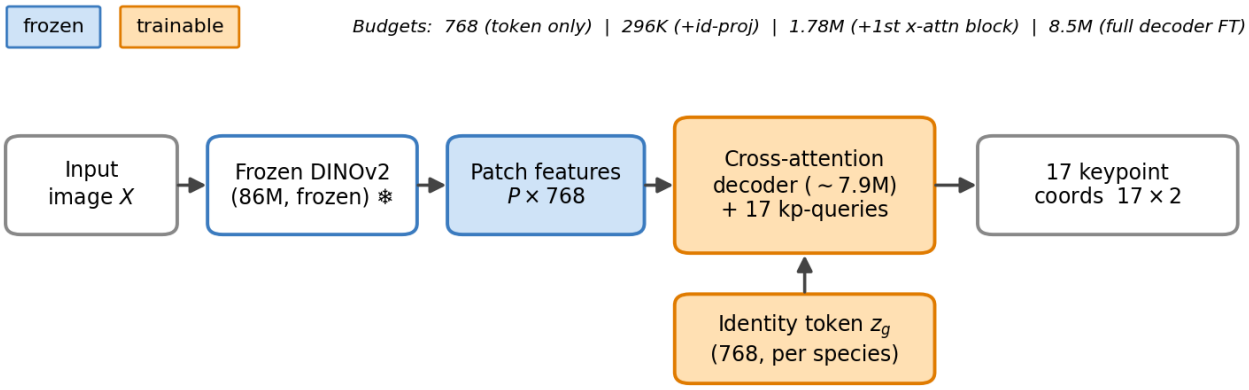

Figure 1. Identity-Token Adaptation. A frozen DINOv2-base backbone (blue, never updated) produces P patch features; a per-identity token z_g (orange) conditions a small cross-attention decoder that regresses 17 keypoints. At inference on a held-out species only the trainable (orange) parameters are updated, at one of four budgets: 768 (token only), 296K (+id-projection), 1.78M (+first cross-attention block), or 8.5M (full decoder fine-tune).

tion comes within 15% of full-decoder fine-tuning on near-and-mid species. (H5a) Interpolation initialization beats random initialization at $k \in \{0, 1\}$ by $\geq 15\%$ RMSE. The 15% margin is a practical-substitutability criterion: a user would accept the cheaper model if its error is no more than 15% worse. This threshold was fixed before any results were seen. Significance is assessed by paired bootstrap (1,000 replicates, paired by image).

4. Experiments

4.1. Pareto Frontier (C1)

Table 5 gives the full per-species results at each parameter budget. The 1.78 M point overfits the small held-out pools and never improves on the 296 K sweet spot.

4.2. Near-Tie with Full Fine-Tuning (C2)

On fox, the 768-parameter token reaches a 1.03× error ratio against an 8.5 M-parameter decoder fine-tune—essentially the same accuracy at 11,118× fewer parameters. At the 296 K budget, mid-and-near species (rabbit, fox, panther) recover 73–82% of the fine-tune's PCK@0.05 at 28× fewer parameters. Far-from-training primates recover only 34–48%, hitting a floor where the frozen features carry too little species-specific information.

4.3. PEFT baseline panel (C5)

We compare ITA against two standard parameter-efficient methods: BitFit (all decoder biases, 14,594 params) and LoRA (r=4 on every decoder linear, 87,560 params), run through the identical 5-shot, 50-step adaptation harness on all eight species at ten seeds. Table 2 reports the results.

Both PEFT methods edge ITA on raw RMSE by a small, statistically significant margin ($p < 10^{-4}$ for BitFit, $p = 0.004$ for LoRA, paired Wilcoxon), but at 19× and 114× more trainable parameters. PCK@0.05 is tied across all three methods. No single method dominates per-species: LoRA wins on far species, BitFit on mid-distance species, and ITA on fox and panther. The takeaway is that ITA is Pareto-favorable—comparable accuracy at one to two orders of magnitude fewer parameters—not that it beats PEFT on raw error.

4.4. Architectural Ablations (C3)

Quantitative ablation. We measure token sensitivity by computing prediction standard deviation across K=50 random tokens, with and without the auxiliary loss. With the loss, sensitivity is 3.9–6.7× higher (mean 5.2×) and never drops to the $\sim 10^{-4}$ floor seen without it. The auxiliary loss is a necessary enabler of the conditioning channel. However, the magnitude of token sensitivity does not predict the magnitude of improvement (Pearson r=0.08, p=0.85, n=8): it is a switch that must be on, not a dial. FiLM-only modulation (global scale-and-shift from the token) gives a null signal on every species—it is too coarse to steer attention to different patches per keypoint.

Non-monotone gain. The improvement peaks at mid-distance species (rabbit, −25.4%), not the nearest (panther, −16.3%) or farthest (4–6%). See Figure 2b

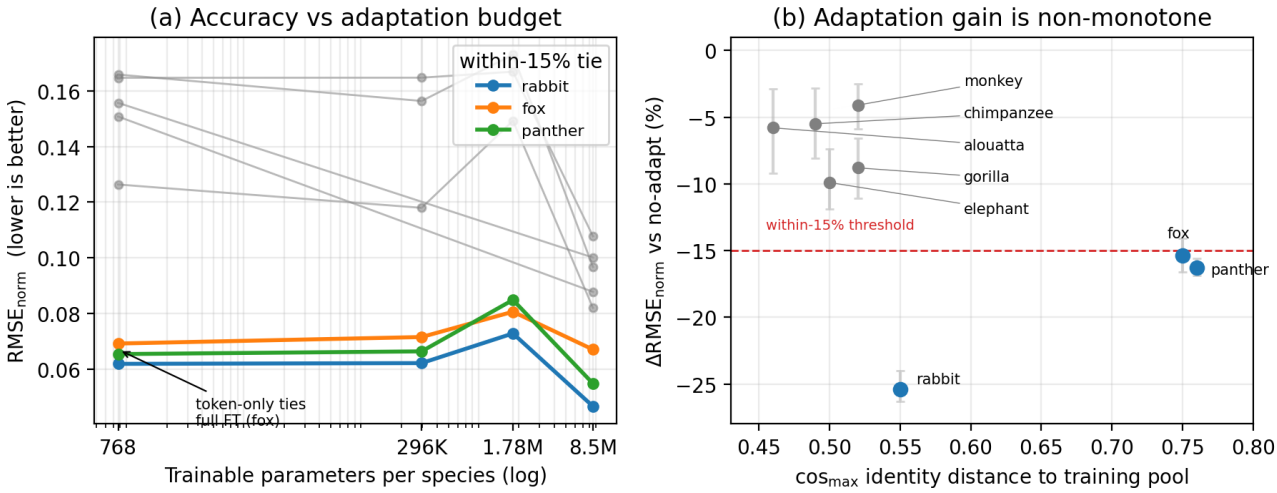

Figure 2. (a) $\mathrm{RMSE}_{\mathrm{norm}}$ versus trainable-parameter budget per species (log x; lower is better). For the within-15% species (rabbit, fox, panther) the 768-parameter token-only point already approaches the 8.5M full-decoder fine-tune; far-from-training primates (gray) keep improving with budget. (b) The gain over the no-adapt baseline is non-monotone in $\cos_{\max}$: rabbit (mid-distance, 0.55) gains most; far and very-near species gain less. Error bars are 95% across-seed bootstrap CIs.

Table 1. Token-only ITA (768-parameter budget) versus the no-adapt mean-token baseline across eight held-out species ($n_{\mathrm{seeds}} = 10$). $\Delta\mathrm{RMSE}_{\mathrm{norm}}$ is the mean % change (more negative is better); the 95% CI is the across-seed bootstrap interval and excludes zero on every species. Stars (★) mark the three species meeting the pre-registered within-15% threshold. The gain is non-monotone in $\cos_{\max}$ (Figure 2b).

Species ($\cos_{\max}$)	$\Delta\mathrm{RMSE}_{\mathrm{norm}}$	95% CI	≤15%
Rabbit (0.55)	−25.4%	[−26.3, −24.0]	★
Fox (0.75)	−15.4%	[−16.6, −14.1]	★
Panther (0.76)	−16.3%	[−16.9, −15.6]	★
Elephant (0.50)	−9.9%	[−11.9, −7.4]	
Gorilla (0.52)	−8.8%	[−11.1, −6.6]	
Alouatta (0.46)	−5.8%	[−9.2, −2.9]	
Chimpanzee (0.49)	−5.5%	[−8.1, −2.8]	
Monkey (0.52)	−4.1%	[−5.9, −2.5]	

Method	Params	$\mathrm{RMSE}_{\mathrm{norm}}$ (↓)	PCK@0.05 (↑)
ITA (token-only)	768	0.120 ± 0.005	0.295
BitFit	14,594	$\mathbf{0.113 \pm 0.004}$	0.297
LoRA (r=4)	87,560	0.114 ± 0.004	0.290

Table 2. PEFT baseline panel: mean ± SEM over 8 held-out species × 10 seeds, same adaptation harness. BitFit/LoRA edge ITA on RMSE (small, significant) at 19–114× more parameters; PCK@0.05 is tied. ITA is Pareto-favorable on parameter efficiency and best on the near-distance species (fox, panther). VPT omitted (would require patching the shared decoder forward).

Table 3. Architectural ablation: adaptation signal under three decoder variants.

Variant	Signal
X-attn id-inject + aux. loss	significant
FiLM-only modulation	null
X-attn, no aux. loss	null (collapsed)

and Appendix A.6 for the full analysis.

4.5. The H5a Binding Test, Falsified (C4)

All three converge to the same behavior: output something close to the average of training-species tokens, regardless of which demonstrations are provided. The pose-regression loss never penalizes this collapse, so the heads take the path of least resistance. We preserve this negative result, rename the method from ITI to ITA (dropping "Interpolation"), and note that a successful interpolation head would likely require an episodic loss that explicitly penalizes collapse to the mean.

Table 4. Direct test of learned interpolation-based initialization at $k \in \{0, 1\}$, three head architectures. None meets the pre-registered $\geq 15\%$ RMSE threshold.

Head architecture	Beats random $\geq 15\%$?
Cosine-softmax (mean-pooled)	no
Per-demo cross-attention	no
Explicit-supervision rescue	no

5. Discussion

What this buys a lab. A per-species pose model from 768 trainable floats ($\approx$ 3 KB), on a laptop, without unfreezing a backbone. On three of eight species the pre-registered 15% threshold is met; on fox the token ties full fine-tuning at 11,118× fewer parameters.

Token geometry. Learned tokens from morphologically similar species (e.g., the five primates) do cluster in $\mathbb{R}^{768}$, but the clustering tracks backbone-feature similarity ($\cos_{\max}$) rather than revealing independent structure. The tokens are useful as conditioning signals, not as a learned taxonomy.

Limitations. ITA is Pareto-favorable against BitFit and LoRA (comparable accuracy at 19–114× fewer parameters), but both edge it on raw RMSE. AdaptFormer, IA^3, and VPT (Jia et al., 2022) remain to be tested. The benchmark is mammal-only (AP-10K); birds and other clades are untested. Our decoder fine-tune comparator is an in-house proxy; a head-to-head comparison with the canonical SuperAnimal (Ye et al., 2024) or DeepLabCut (Mathis et al., 2018) fine-tune would strengthen the calibration.

Takeaway. Token-only adaptation can match full fine-tuning on near-and-mid-distance species at a few kilobytes per identity, but hits a floor on far-from-training species. The right reading is the per-species frontier, not a single headline number.

References

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Short Papers), 2022. arXiv:2106.10199.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), 2022. arXiv:2106.09685.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. Visual prompt tuning. In Proceedings of the European Conference on Computer Vision (ECCV), 2022. arXiv:2203.12119.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience, 21:1281–1289, 2018. doi: 10.1038/s41593-018-0209-y.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024. arXiv:2304.07193.

Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., Papadoyannis, E. S., Normand, E., Deutsch, D. S., Wang, Z. Y., McKenzie-Smith, G. C., Mitelut, C. C., Diez Castro, M., D'Uva, J., Kislin, M., Sanes, D. H., Kocher, S. D., Wang, S. S.-H., Falkner, A. L., Shaevitz, J. W., and Murthy, M. SLEAP: A deep learning system for multi-animal pose tracking. Nature Methods, 19: 486–495, 2022. doi: 10.1038/s41592-022-01426-1.

Ye, S., Filippova, A., Lauer, J., Schneider, S., Vidal, M., Qiu, T., Mathis, A., and Mathis, M. W. SuperAnimal pretrained pose estimation models for behavioral analysis. Nature Communications, 15, 2024. doi: 10.1038/s41467-024-48792-2.

Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., and Tao, D. AP-10K: A benchmark for animal pose estimation in the wild. In Advances in Neural Infor-

mation Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2021. arXiv:2108.12617.

A. Supplementary Material

A.1. Detailed Methodology

Identity-Token Adaptation. A frozen DINOv2-base backbone (Oquab et al., 2024) (f_θ; 86 M parameters, never updated) maps an image X to P patch features of dimension d=768. Each species g gets a learned token $z_g \in \mathbb{R}^d$. A cross-attention decoder takes both as input and outputs 17 keypoint coordinates. To adapt to a new species $g^\star$, we optimize only $z_{g^\star}$ (and optionally a small subset of the decoder) on a few labeled frames.

Token-utility auxiliary loss. Without it, the decoder ignores the token: prediction spread across random tokens collapses to $\sim 10^{-4}$. A hinge-margin term forces this spread above a target threshold, restoring the token as a usable conditioning signal. We tried cosine-contrastive and classification-head alternatives; the hinge margin was the most stable.

Evaluation. Eight held-out species, ten seeds, spanning $\cos_{\max} \in [0.46, 0.76]$. Pre-registered thresholds: (H1) within 15% of full fine-tune; (H5a) interpolation init beats random init at $k \in \{0, 1\}$ by $\geq 15\%$. Statistics: paired bootstrap by image (1,000 replicates). Primary metric: $\text{RMSE}_{\text{norm}}$; secondary: PCK@0.05.

A.2. Pareto Frontier and Parameter Budgets

We evaluate at four parameter budgets: 768 (token only), 296 K (token + projection layer), 1.78 M (token + cross-attention block), and 8.5 M (full decoder fine-tune). Table 5 gives the full matrix. The 296 K sweet spot recovers 73–82% of fine-tune accuracy on mid-and-near species and 34–48% on far species, at 28× fewer parameters. The 1.78 M budget overfits and never beats 296 K. The token-only point ties the fine-tune on fox at 11,118× fewer parameters.

Species ($\cos_{\max}$)	768	296K	1.78M	DFT 8.5M	ITA/DFT
alouatta (0.46)	0.151	–	–	0.088	1.72
chimpanzee (0.49)	0.166	0.156	0.173	0.097	1.72
elephant (0.50)	0.126	0.118	0.149	0.082	1.54
monkey (0.52)	0.156	–	–	0.100	1.56
gorilla (0.52)	0.165	0.165	0.167	0.108	1.53
rabbit (0.55)	0.062	0.062	0.073	0.047	1.33
fox (0.75)	0.069	0.072	0.081	0.067	1.03
panther (0.76)	0.065	0.066	0.085	0.055	1.19

Table 5. Per-species × parameter-budget $\text{RMSE}_{\text{norm}}$ (mean over 5–10 seeds). 768 = token-only ITA; 296K = +id-proj; 1.78M = +cross-attention block; DFT 8.5M = full decoder fine-tune; ITA/DFT = 768-param point vs. comparator. "–" = budget not run.

A.3. Full-Decoder-FT Comparator and Strength Sensitivity

Our comparator is a decoder fine-tune (8.5 M parameters, SGD), a proxy for the canonical SuperAnimal fine-tune at our scale. To rule out under-training, we swept learning rate and step budget on fox. The default setting (300 steps, lr 10^{-4}) gives RMSE = 0.0671; the best-tuned setting (800 steps, lr 3×10^{-4}) gives 0.0646. Against the best-tuned comparator, the 768-parameter token (0.0692) sits at a 1.07× ratio—still within the 15% threshold. The near-tie holds.

A.4. Auxiliary Loss, Token Sensitivity, and FiLM-Only Modulation

We re-measured token sensitivity (K=50 random tokens, $\mathcal{N}(0, 0.02^2)$) with and without the auxiliary loss on all eight species. With the loss, per-coordinate prediction spread is 3.9–6.7× higher (mean 5.2×) and stays well above the $\sim 10^{-4}$ floor seen without it. The loss is necessary—it must be on for the token to matter—but the magnitude of sensitivity does not predict the magnitude of improvement (Pearson r=0.08, p=0.85, n=8). It is a switch, not a dial.

FiLM-only modulation (global per-channel scale and shift from the token) gives a null adaptation signal on every species. Global affine modulation cannot route the token to specific patches; cross-attention can, because each

keypoint query attends selectively to the relevant patch features.

A.5. The H5a Binding-Test Falsification

For each held-out species we provide $k \in \{0, 1\}$ example frames, run an interpolation head to produce a candidate token, and compare its error against random initialization. The pre-registered bar is a $\geq 15\%$ improvement.

None of three architectures (cosine-softmax on mean-pooled demos, per-demo cross-attention, explicit-supervision rescue) meets this bar on any species. All three collapse to outputting something near the average of training tokens, regardless of the demos provided. The cause is the training signal: the pose loss never penalizes producing the average, so the heads have no reason to be more specific. A working interpolation head would likely need either more demonstrations or an episodic loss that explicitly penalizes collapse to the mean.

A.6. Identity Distance, Sample Complexity, and Cross-Domain Scouts

Non-monotone gain. Rabbit (mid-distance, cos =0.55) gains the most (-25.4%), chimpanzee (far, 0.49) gains the least ($\sim -5.5\%$), panther (near, 0.76) falls in between (-16.3%). Mid-distance species have the most headroom; far species lack information in the frozen features; near species are already well-served. Cosine distance alone does not predict where the token helps most.

Sample complexity. Improvement scales with sample budget k but plateaus around $k \approx 20$. Going from 20 to 50 frames gives little additional benefit. A useful model can be trained from $\sim$20 labeled frames.

Cross-domain scout. On a held-out SuperAnimal (Ye et al., 2024) species set absent from AP-10K, token-only adaptation significantly improves three of five species; the remaining two sit at the floor (very far from training data). The pattern matches the within-AP-10K results.

Substrate-invariance conjecture (future work). We hypothesize that a Wasserstein-2 threshold on backbone features may predict where token-only adaptation works across domains (pose, handwriting, speech). A feasibility check shows $\geq 2\sigma$ in/out-of-pool separation on AP-10K, but cross-domain testing is needed.

A.7. Deployment and Reproducibility

Because the backbone is frozen, its output can be cached: run each image through DINOv2 once, save the features, and reuse them for every adaptation run. This makes the eight-species, multi-seed sweep run in under twenty hours on a Mac laptop at zero cloud cost. To deploy on a new species: cache the backbone features, train the token on 20–50 labeled frames, and use it. No specialized hardware needed. Table 6 summarizes per-species recommendations.

Table 6. Per-species deployment recommendations.

Species	Recommended config	Param. budget
Rabbit, Fox, Panther (mid/near)	296 K Pareto sweet point	296,000
Fox (ultra-cheap)	Token-only ITA	768
Primates (far)	Full decoder fine-tune	8.5 M

A.8. Limitations and Broader Impact

Scope. We test on AP-10K (mammals only). Birds and other clades are a natural extension. Generalization to non-pose domains (handwriting, speech) is conjectured but not tested here.

PEFT baselines. BitFit and LoRA are included (Table 2); AdaptFormer, IA^3, and VPT (Jia et al., 2022) are not yet tested.

Broader impact. Token-only adaptation extends per-species pose estimation to field labs without specialized hardware. We see no direct misuse risk.

Reproducibility. The entire evaluation ran on a Mac laptop at zero cloud cost. Code, configs, and weights are released. All figures use fixed random seeds.