

ML Lab Research Proposal

Identity-Token Adaptation for Cross-Species Animal Pose Estimation

Team: Ethan Wang (solo)
Date: July 22, 2026
Project: ITA: Identity-Token Adaptation for Cross-Species Pose on AP-10K
Target Compute: 1x AMD Instinct MI300X GPU, approx. 2 weeks

1. PROBLEM STATEMENT

Animal pose estimation toolkits like DeepLabCut, SLEAP, and SuperAnimal can track keypoints across dozens of species, but adapting to each new species still requires fine-tuning millions of backbone parameters. On the AP-10K benchmark (Yu et al., NeurIPS 2021), which covers 54 mammal species with a shared 17-keypoint skeleton, moving from a trained species pool to a held-out species typically means retraining an entire decoder head (8.5M parameters) or applying parameter-efficient methods like LoRA (87K parameters) or BitFit (14K parameters). We ask whether a single learned 768-dimensional token, roughly 11,000x fewer parameters than a full decoder fine-tune, can close the accuracy gap.

2. PROPOSED APPROACH

Identity-Token Adaptation (ITA) freezes a DINOv2-base backbone (Oquab et al., TMLR 2024; 86M parameters, never updated) and attaches a cross-attention decoder conditioned on one 768-dimensional token per species. Adapting to a new species means optimizing only that token. We chose DINOv2 because its self-supervised distillation objective produces patch features with strong spatial consistency: 49.0 mIoU on ADE20k segmentation with a frozen backbone versus 33.3 for MAE and 39.3 for OpenCLIP (Oquab et al., 2024). Two design choices are critical: cross-attention conditioning (FiLM-style modulation is too coarse to steer individual keypoints) and a hinge-margin loss that prevents the decoder from ignoring the token.

Preliminary results on 8 held-out AP-10K species show ITA reduces normalized keypoint error on all 8 (95% paired-bootstrap CIs excluding zero). On fox, the 768-parameter token reaches a 1.03x error ratio against the 8.5M-parameter decoder fine-tune. Against BitFit (Ben Zaken et al., ACL 2022) and LoRA rank-4 (Hu et al., ICLR 2022), ITA matches accuracy at 19 to 114x fewer parameters. Gains follow an inverted-U with feature-space distance: mid-distance species benefit most, near species less, far species least.

We now need to validate four claims for a main-conference submission: (1) the approach generalizes beyond DINOv2, (2) the inverted-U holds with more species, (3) retrieval-based token initialization enables rapid few-shot adaptation, and (4) the identity token can capture body-plan differences across non-mammalian skeletons.

3. DATASET

Dataset	Source	Size	Use
AP-10K	Yu et al., NeurIPS 2021	10,015 images 54 mammal species	Primary benchmark. 17-keypoint shared skeleton. Training pool + held-out species evaluation.

Dataset	Source	Size	Use
APT-36K	Yang et al., NeurIPS 2022	36,000 frames 30 species	Non-mammalian species (birds, fish) for multi-skeleton transfer (Experiment 4).
CUB-200	Wah et al., 2011	11,788 images 200 bird species	Fallback bird skeleton data (15 keypoints) if APT-36K bird annotations are insufficient.

Preprocessing

- All images resized to 224x224. DINOv2 patch features (1+256 tokens, 768-dim) extracted once and cached as .npy shards.
- AP-10K keypoints normalized to [0,1] relative to bounding box. Standard left-right flip augmentation.
- On-disk footprint per backbone: roughly 12 GB of cached features. Total across 5 backbones: roughly 60 GB, well under a 1 TB scratch budget.

4. EVALUATION METRICS

Experiment	Metric	Target / Interpretation
Cross-backbone validation (5 backbones x 3 species)	RMSE_norm, PCK@0.05/0.10/0.20	ITA efficiency ratio (768 params vs. full fine-tune) holds across DINOv1, MAE, CLIP, EVA-02, not just DINOv2
Inverted-U expansion (8 species total)	Adaptation gain vs. cosine distance; quadratic R-squared vs. linear	Significantly negative quadratic coefficient ($p < 0.05$) confirms inverted-U is not an artifact of 3 datapoints
Few-shot token init (3 strategies x 4 shot counts)	Convergence speed (epochs to 90% of full-training performance)	Retrieval-based init reaches 90% accuracy 2-3x faster than random init
Multi-skeleton transfer (birds, fish)	PCK@0.10 on non-mammalian species	PCK@0.10 > 0.3 for birds demonstrates identity token captures body-plan differences, not just texture

5. COMPUTE REQUIREMENTS

All experiments use cached frozen-backbone features, so each training run is lightweight. Estimates are scaled from measured single-GPU runs on Apple M2 and extrapolated to MI300X throughput.

Phase	Hardware	Est. Wall-Clock	Notes
Feature extraction (5 backbones x AP-10K + APT-36K)	1x MI300X	roughly 4 hours	One-time forward pass per backbone. Cached to disk as .npy shards.
Exp 1: Cross-backbone (5 backbones x 3 species x 10 seeds)	1x MI300X	roughly 3 days	510 GPU-hours of A100-equivalent work. MI300X accelerates via larger batch.
Exp 2: Inverted-U expansion (5 new species x 3 methods x 10 seeds)	1x MI300X	roughly 8 hours	50 GPU-hours. Fastest experiment, highest priority.

Phase	Hardware	Est. Wall-Clock	Notes
Exp 3: Few-shot init (3 strategies x 7 species x 4 K-values x 5 seeds)	1x MI300X	roughly 6 hours	45 GPU-hours. Plus roughly 1 hour for learned projector training.
Exp 4: Multi-skeleton (decoder redesign + non-mammal eval)	1x MI300X	roughly 4 hours	20 GPU-hours. Smallest compute, most manual annotation effort.
Total	1x MI300X	roughly 5 days within 2-week window	roughly 625 GPU-hours A100-equivalent. Remainder for reruns and ablations.

Software Dependencies

- Python 3.10+, PyTorch 2.x with ROCm backend.
- HuggingFace Transformers for backbone loading (DINOv2, MAE, CLIP, EVA-02 checkpoints).
- No CUDA-only kernels. All operations are standard PyTorch (nn.MultiheadAttention, AdamW, CosineAnnealingLR).
- No Docker required. Pure Python venv with pip install.

6. EXPECTED DELIVERABLES

- Cross-backbone comparison table: ITA efficiency ratio across DINOv2, DINOv1, MAE, CLIP, EVA-02 on 3 species with 10-seed CIs.
- Inverted-U validation plot: adaptation gain vs. feature-space distance for 8 species, with quadratic and linear regression fits and bootstrap CIs on the quadratic coefficient.
- Few-shot initialization benchmark: convergence curves for 3 initialization strategies across 4 shot counts, with epochs-to-90% as the headline metric.
- Multi-skeleton transfer results: PCK@0.10 on bird and fish species with the skeleton-agnostic decoder, plus cross-attention visualizations showing what the identity token attends to.
- An AAAI-27 main-conference paper draft with all tables and figures generated end-to-end from JSON result files, plus the full reproducibility codebase.

7. RISK & MITIGATION

Risk	Likelihood	Mitigation
ITA works only on DINOv2 and fails on MAE/CLIP backbones	Medium	Each backbone runs with a full-decoder control (no ITA). If frozen MAE + full decoder also fails, the issue is feature quality, not ITA. Report the ratio (ITA / full-decoder) to isolate the contribution.
Inverted-U pattern does not hold with 8 species	Medium	Also fit piecewise and monotonic regressions. A monotonic decrease is still publishable and corrects the 3-point preliminary observation. Test with phylogenetic distance (TimeTree) as an alternative metric.
Token space is not smooth, so retrieval-based initialization provides no benefit	Medium	Compute pairwise cosine similarity of all learned training tokens vs. feature-space similarity. If correlation < 0.3, report as a negative result and focus the paper on Experiments 1 and 2.

Risk	Likelihood	Mitigation
Small validation sets for rare species (< 30 images) produce high-variance estimates	High	Use 10 seeds per condition and report 95% bootstrap CIs. Exclude species with fewer than 15 validation images from headline claims. Flag sample-size limitations explicitly.
Wall-clock overrun on MI300X	Low	Experiments are prioritized: Exp 2 first (8 hours), then Exp 1 (3 days). If time runs short, drop Exp 4 (stretch goal) and reduce Exp 1 to 3 backbones (DINOv2, MAE, CLIP).