It's the Adaptation, Not the Architecture: Pretrained Vision Transformers Are Competitive for End-to-End Steering on Small Driving Data

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Vision Transformers (ViTs) are widely held to be too data-hungry, for lack of the convolutional network's locality inductive bias, to compete on small datasets [6, 15]. We find this is not the case for end-to-end steering. On a small slice of the comma2k19 driving dataset [14] (∼5–16k frames), a DINO-pretrained ViT-S, fine-tuned appropriately, is **competitive with** a pretrained ResNet-50 at predicting steering angle from a single forward camera frame: turn-slice Pearson correlation 0.964 vs. 0.967 over three seeds, a difference within the seed-to-seed spread (we do not run a formal equivalence test). The transformer matches the CNN; it does not beat it. The result is conditional on adaptation: the defaults a practitioner reaches for first, a frozen linear probe or horizontal-flip augmentation, reproduce the common "ViTs don't work at small scale" conclusion, and competitiveness emerges only once those choices are corrected. We document the configuration we found necessary (low-learning-rate full fine-tuning, a cost-sensitive loss for the heavily imbalanced target, scale-free turn-stratified evaluation, and dropping flip augmentation) and ablate each choice. Because the ViT is DINO-pretrained and the ResNet ImageNet-supervised, this is a comparison of pretrained pipelines, not of architectures in isolation. Evaluation is open-loop only. The reproducibility package (code and the exact data-subset recipe) is available on request.

1 Introduction

End-to-end learning of steering from camera pixels, mapping an image directly to a control output, dates to PilotNet [2] and underlies modern production stacks (e.g. comma.ai's openpilot [4]). It is also a natural testbed for a question that recurs whenever Vision Transformers [6] meet a new domain with limited labelled data: *do transformers, which lack the CNN's locality inductive bias and are reputed to be data-hungry [6, 15], underperform CNNs at small scale?* That belief discourages practitioners with small datasets from using pretrained transformers at all.

We study this on a deliberately small slice of comma2k19 [14], predicting per-frame steering angle with leakage-safe group-by-route splits, and find that a properly adapted DINO-pretrained ViT-S is *competitive* with a pretrained ResNet-50 here: it matches, without beating, the CNN, with no disadvantage we can detect at three seeds.

The finding is conditional on adaptation. The default configuration a practitioner would reach for first reproduces the opposite, "data-hungry" conclusion: a frozen linear probe over the ViT's features yields turn-slice correlation 0.38, and the standard label-correct horizontal-flip augmentation collapses the model to a constant predictor. Competitiveness emerges only after these

*Corresponding author: deep@lab.cloud

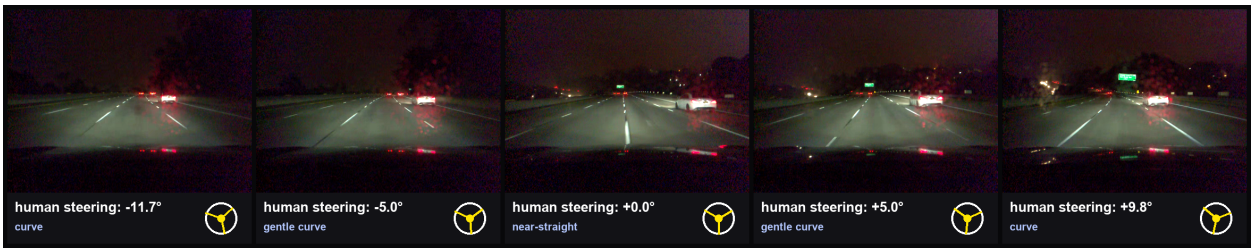

Figure 1: Example comma2k19 frames (one highway segment, pre-dawn) with the human steering-wheel angle the model is trained to predict; the glyph is a steering wheel rotated by that angle. The sequence spans the segment's range ($-11.7°$ to $+9.8°$), illustrating that highway steering is overwhelmingly gentle, the source of the ∼95%-near-zero imbalance that drives our metric and augmentation choices. (Brightness raised for legibility; sign per the comma2k19 CAN convention.)

defaults are corrected. This conditionality is itself informative: it shows why the working configuration must be documented rather than assumed, and it indicates that the "data-hungry" conclusion can be an artifact of adaptation choices rather than of architecture.

Contributions.

1. A **finding**: at ∼5–16k frames on comma2k19 steering, a DINO-pretrained ViT-S is competitive with an ImageNet ResNet-50 on turn-slice correlation, matching rather than exceeding it, within the seed-to-seed spread at 16k and near-ceiling for both by ∼5k (Section 5, Table 1). Pretrained transformers are not data-hungry *on this task*.
2. The **configuration we found necessary** to obtain it: low-LR full fine-tuning, a cost-sensitive loss for the imbalanced target, scale-free turn-stratified evaluation, and no flip augmentation, with each choice ablated (Sections 4–5). We validate, rather than optimize or claim transfer beyond this task.
3. A **counterintuitive augmentation finding**: on this heavily imbalanced (∼95% near-zero) target, the standard label-correct horizontal flip collapses the model to a constant predictor, and a predicted-dispersion diagnostic (`pred_sd`) reliably detects the collapse that aggregate error hides (Section 5).

2 Data and Task

Dataset. comma2k19 [14] is ∼33h of California highway driving (public, MIT-licensed; accessed via HuggingFace), recorded in ∼1-minute segments with a 20 Hz forward camera and synchronized CAN logs. We use a subset of one chunk (188 segments). The prediction unit is a single frame; the target is the contemporaneous CAN steering-wheel angle (degrees), nearest-timestamp-matched to the frame (∼88 Hz CAN → 20 Hz camera). We clip labels to $[-50°, 50°]$ and drop frames below 4 m/s.

Imbalance. The steering distribution is dominated by straight driving: ∼94.5% of frames lie within $\pm 15°$, leaving only ∼5.5% on the *turn slice* ($|\text{angle}| > 15°$). This single fact drives every modelling choice below; we write "∼95% near-zero" as shorthand for this $\pm 15°$ split.

Splits. We split *by route* (the dongle|timestamp prefix), never by individual segment: consecutive segments of one drive are near-duplicates, so random frame or segment splits leak adjacent frames across train/test and inflate scores [7]. Train/val/test $\approx$ 60/20/20 by route.

3 Models and the Configuration We Used

Models. We compare {conv, transformer}$\times${single-frame, temporal}. Single-frame: a small from-scratch *PilotNet* CNN [2], an ImageNet-pretrained [13] *ResNet-50* [8], and a DINO-pretrained *ViT-S/16* [3, 6]. The two pretrained backbones differ in pretraining scheme (self-supervised DINO vs. supervised ImageNet), so their comparison reflects the full pretrained pipeline, not architecture alone. Temporal (shared small CNN stem over short within-segment clips): a *GRU* head [5] and a *temporal-attention* head [1]; the temporal arms compare attention vs. recurrence on a shared stem.

The four choices below are the configuration we found necessary on this task, each motivated by the $\sim$95%-near-zero imbalance and each ablated in Section 5. We did not sweep them exhaustively or test them on other datasets, so we report them as a validated configuration, not a tuned or transferable recipe.

(1) Adapt pretrained backbones with a low-LR full fine-tune. Inputs are 224×224 (single-frame) or 66×200 (temporal CNN stem); labels are standardized on train statistics. We fine-tune the *entire* backbone rather than linear-probing frozen features, at a deliberately low learning rate: for the ViT, AdamW [12] at $\mathrm{lr} = 3\times10^{-5}$ (the CNNs use Adam [10]; exact optimizer settings are in the released config). Section 5 shows this choice is what moves the pretrained ViT from sub-conv to conv-competitive.

(2) A cost-sensitive loss so the mean is not optimal. We train with a cost-sensitive smooth-ℓ_1 loss $w \cdot \mathrm{smooth}\ell_1(\hat{y}, y)$, $w = 1 + |y|/15$, which up-weights turn frames so the trivial "predict the training mean" solution is no longer loss-optimal [11]. On a $\sim$95%-near-zero target, a plain loss rewards collapse to the mean; the turn-weighting removes that incentive.

(3) Evaluate scale-free, on the slice that matters. Because of the imbalance, overall MAE flatters trivial predictors, and R^2 (whose denominator is dominated by rare turns and uses the shifted test route's own variance) is misleading. We therefore lead with: **turn-slice MAE**, **Pearson and Spearman correlation** (overall and on the turn slice; scale-free), **median absolute error**, and **predicted-angle standard deviation** (`pred_sd`); a degenerate mean-predictor has `pred_sd` ≈ 0, so this diagnostic catches collapse that aggregate error hides. This is evaluation discipline rather than a novel metric. Trivial references: *predict-mean* (drive straight) and *predict-previous-angle* (autocorrelation).

(4) Drop label-correct flip augmentation on imbalanced targets. Horizontal flip (mirror the image, negate the angle) is label-correct but symmetrizes the already-$\sim$95%-near-zero distribution around 0, reinforcing the predict-zero solution. We do not use it; Section 5 quantifies the harm.

4 Results

Trivial baselines. On held-out routes, predict-mean has turn-MAE 22.5°; predict-previous-angle has overall MAE $\sim 0.45°$ and turn-MAE $\sim 1.4°$. The latter is nearly unbeatable on 20 Hz highway data and *remains unbeaten on overall error by every learned model*; it is reported as a *reference*,

not a target, and is not available to a single-frame model. The learned models' value is in the turn slice, not in overall error.

Conv and transformer are competitive (3 seeds). Table 1 is the central result. With the configuration of Section 4 (ViT full fine-tune at lr $= 3 \times 10^{-5}$), the pretrained ViT **matches** the pretrained ResNet without beating it: turn-slice correlations differ by 0.003, smaller than the seed-to-seed std (≈ 0.006) and within noise at $n = 3$; overall correlation is comparable (~ 0.82), and mean absolute errors are close (1.70 vs. 1.79°). We report the observed difference and its spread and do *not* run an equivalence test, so we claim no detectable disadvantage rather than statistical equivalence. Both families improve with data on overall correlation ($\sim 0.76 \rightarrow 0.82$ from 5k to 16k frames); turn-slice correlation is already near-ceiling (~ 0.96) by 5k frames: the ViT does not need more data to "catch up."

Table 1: Single-frame steering: held-out (group-by-route) metrics, mean ± std over 3 seeds, reproduced on GPU. Turn-slice = |angle| > 15°. Higher correlation / lower MAE is better. †The ViT@5.2k turn-slice std rounds to 0.000 at three decimals; we observed near-zero seed variance at that cell and do not lean on it for any strong claim.

Model	Data	Pearson (turns)	Pearson (all)	MAE (deg)
ResNet-50 (conv, pretrained)	5.2k	0.954 ± 0.007	0.756 ± 0.027	2.03
ResNet-50 (conv, pretrained)	16k	$\mathbf{0.967 \pm 0.006}$	0.821 ± 0.028	1.70
ViT-S/DINO (transformer, full ft)	5.2k	$0.965 \pm 0.000^{\dagger}$	0.789 ± 0.015	1.77
ViT-S/DINO (transformer, full ft)	16k	$\mathbf{0.964 \pm 0.006}$	0.815 ± 0.009	1.79
PilotNet (conv, from scratch)	9k	0.87 ± 0.02	0.46 ± 0.05	3.49

The ViT is learning-rate sensitive (adaptation ablation). The same ViT-S/DINO backbone reaches very different quality depending on how it is adapted (Table 2). A frozen linear probe, testing only the pretrained features, gives turn-slice correlation 0.38; a full fine-tune at lr $= 10^{-4}$ underfits (training loss plateaus); only the full fine-tune at lr $= 3 \times 10^{-5}$ reaches conv-competitive quality. This shows the ViT's competitiveness depends strongly on adaptation; it does *not*, by itself, establish that architecture is irrelevant, because we did not run the symmetric LR sweep for the ResNet (Section 6). We also note that some of our earlier, pre-fix ViT failures were traced to infrastructure and code bugs (a silent CPU fallback, a loss-broadcasting bug), not only to learning-rate choice; the numbers here are from the post-fix GPU runs.

Table 2: The ViT result is learning-rate sensitive. Same backbone (ViT-S/DINO); the adaptation differs. These rows come from runs that differ in epoch count and hardware as well as adaptation, so they bound the effect rather than isolate it perfectly (Section 6).

ViT-S/DINO adaptation	Pearson (turns)
Frozen linear probe (pretrained features only)	0.38 ± 0.08
Full fine-tune, lr 10^{-4}	underfits (loss plateau)
Full fine-tune, lr 3×10^{-5}	$\mathbf{0.964 \pm 0.006}$

Flip augmentation collapses the model (augmentation ablation). The standard label-correct horizontal flip is *harmful* here. Because the angle distribution is ∼95% near-zero, flipping symmetrizes it around 0 and reinforces the predict-zero solution: with flip enabled, models collapse to a constant (`pred_sd` → 0, correlation → 0) despite a plausible training loss. Removing the flip (together with the cost-sensitive loss) recovered a genuine model: PilotNet turn-slice correlation $0 \rightarrow 0.87$; the loss and augmentation changes are entangled in this single recovery, so we attribute the recovery to the pair, not to either alone. The `pred_sd` diagnostic (strong models ≈ 2.6–3.5 against a true signal of ∼5°; collapsed runs ≈ 0) is what surfaced the collapse.

Where the error lives. Errors concentrate on the rare turn slice; the straight bulk is easy. Median absolute error is ∼1.3–2° for the strong models (typical-frame error is small), while turn-MAE is ∼15–16° for those models against 22.5° for predict-mean, so the real models clearly read curves rather than driving straight.

Temporal axis. On a shared small CNN stem, temporal attention (0.81) and a GRU (0.82) are comparable in a single run (no seed dispersion reported). Because this stem is smaller than the pretrained backbones, we restrict the comparison to attention vs. recurrence within this row and draw no cross-row conclusion about temporal context.

5 Related Work

End-to-end steering from pixels follows PilotNet [2]; comma2k19 [14] and openpilot [4] are the dataset and production context. The near-zero-steering imbalance and its effect on metrics is studied by Klein et al. [11]; temporal modelling and split-leakage by Eraqi et al. [7]; open-loop vs. closed-loop validity by Xiao et al. [16]. On the vision side, ViTs [6] are reputed to be data-hungry, motivating data-efficient training (DeiT [15]) and self-supervised pretraining (DINO [3], MAE [9]). The mechanism behind our Principle 1, that a low-LR *full* fine-tune of a pretrained backbone outperforms linear probing on small downstream data, is a known transfer-learning phenomenon rather than a new method; our contribution is to confirm it matters decisively for this imbalanced regression task and to document the accompanying loss/metric/augmentation choices. Relative to the broad "ViTs are data-hungry" prior, we provide a concrete small-data regression case where, with self-supervised pretraining and correct adaptation, a ViT is not.

6 Limitations, Scope, and Threats to Validity

This is a deliberately small feasibility study, and we are explicit about what it does *not* show. **Scope.** (1) **Open-loop only**: we measure per-frame prediction error, not closed-loop driving; good open-loop error does not imply driving competence (covariate shift compounds errors), and we make no claim that any model would drive. (2) **Small scale**: minutes–hours of data, one chunk, one vehicle, highway-dominant; findings may not hold at larger scale or off-highway. (3) The predict-previous-angle reference remains unbeaten on overall error and bounds what "good" means for a single-frame model.

Threats to validity. *Construct:* the ViT (DINO, self-supervised) and ResNet (ImageNet, supervised) differ in pretraining scheme, so our result is about pretrained pipelines, not architectures in isolation; "not data-hungry" is demonstrated only at the two sizes we tried (5.2k, 16k), where the turn-slice metric is already near-ceiling, so we cannot rule out that the slice is simply easy at these scales. *Internal:* the adaptation sweep was run for the ViT but not symmetrically for the ResNet, so the "competitiveness depends on adaptation" claim is about the ViT, not a controlled

architecture verdict; the Table 2 rows also differ in epochs and hardware, bounding rather than isolating the adaptation effect; and the loss/augmentation ablation is entangled (Section 5). Some earlier failures were infrastructure/code bugs rather than modelling choices. *External:* we validate a single configuration on one dataset and do not demonstrate transfer to other targets, imbalance levels, or architectures, hence "configuration we used," not "transferable recipe." None of these undercut the competitiveness finding at this scale or the augmentation result; they bound how far each may be generalized.

7 Conclusion

On a small, real driving-data steering task, a DINO-pretrained Vision Transformer is competitive with a pretrained convolutional network: it matches, without beating, the CNN once both are fine-tuned properly, with no disadvantage detectable at three seeds. The competitiveness is conditional: the defaults a practitioner reaches for first (frozen probing, flip augmentation) reproduce the "transformers are data-hungry" conclusion, and only correct adaptation reverses it. The practical takeaways are the configuration we found necessary here: low-LR full fine-tuning, a cost-sensitive loss, scale-free turn-stratified evaluation, and not using label-correct flip augmentation on a heavily imbalanced target. We offer them as validated practice rather than a tuned or transferable recipe.

Availability

The reproducibility package (training and evaluation code, the 2×2 model definitions, the metric/evaluation harness, and the exact data-subset recipe: HuggingFace dataset id, chunk and route selection, split assignment, and content hash) is available on request. We do not redistribute comma2k19 itself (it is comma.ai's, under its own license), and we do not release trained checkpoints.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[4] Li Chen et al. Level 2 autonomous driving on a single device: Diving into the devils of openpilot. *arXiv preprint arXiv:2206.08176*, 2022.

[5] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, 2014.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Hesham M. Eraqi, Mohamed N. Moustafa, and Jens Honer. End-to-end deep learning for steering autonomous vehicles considering temporal dependencies. *arXiv preprint arXiv:1710.03804*, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Kaiming He, Xinlei Chen, Saining Xie, et al. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[11] Nadja Klein, David J. Nott, et al. Marginally calibrated response distributions for end-to-end learning in autonomous driving. *arXiv preprint arXiv:2110.01050*, 2021.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[14] Harald Schafer, Eder Santana, Andrew Haden, and Riccardo Biasini. A commute in data: The comma2k19 dataset. *arXiv preprint arXiv:1812.05752*, 2018.

[15] Hugo Touvron, Matthieu Cord, Matthijs Douze, et al. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[16] Yi Xiao, Felipe Codevilla, et al. Scaling vision-based end-to-end autonomous driving with multi-view attention learning. *arXiv preprint arXiv:2302.03198*, 2023.