2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Robustness Is a Function, Not a Number: A Factorized Comprehensive Study of OOD Robustness in Vision-Based Driving

Amir Mallak[1] Alaa Maalouf[1]
[1]University of Haifa
Correspondance: mallak002@gmail.com

***Abstract*— Out-of-distribution (OOD) robustness in vision-based autonomous driving is often reduced to a single number, hiding what breaks a policy and by how much. We adopt a factorized view, decomposing environments along five axes: scene (rural/urban), season, weather, time (day/night), and agent mix; and measure performance under controlled k-factor perturbations ($k \in \{0, 1, 2, 3\}$). Using closed-loop control in VISTA, we benchmark FC, CNN, and ViT policies, train compact ViT heads on frozen foundation-model (FM) features, and vary in-distribution (ID) support in scale, diversity, and temporal context. (1) ViT policies are markedly more OOD-robust than comparably sized CNN/FC, and adding FM features yields state-of-the-art success at a latency cost. (2) Naive temporal inputs (multi-frame) do not beat the best single-frame baseline. (3) The largest single-factor drops are rural $\rightarrow$ urban and day $\rightarrow$ night ($\sim 31\%$ each); actor swaps $\sim 10\%$ and moderate rain $\sim 7\%$; several season shifts are drastic, and combining a time flip with other changes further degrades performance. (4) FM-feature policies stay above 85% under three simultaneous changes; non-FM single-frame policies take a large first-shift hit, and all no-FM models fall below 50% by three changes. (5) Interactions are non-additive: some pairings (e.g., urban-night) partially offset, whereas season–time combinations are especially harmful. (6) Training on winter/snow is most robust to single-factor shifts, while a rural+summer baseline gives the best overall OOD performance. (7) Scaling traces/views of the same configuration improves robustness (about $+11.8$ points from 5 to 14 traces), yet targeted exposure to hard conditions can substitute for scale. (8) Using multiple ID environments broadens coverage and strengthens weak cases (urban OOD $60.6\% \rightarrow 70.1\%$) with a small ID drop; single-ID preserves peak performance but in a narrow domain. These results yield actionable design rules for OOD-robust driving policies.**

I. Introduction

Autonomous driving systems must operate far beyond the narrow slice of conditions seen during training. Real roads combine many shifting *factors*: scene layout (rural vs. urban), time of day, season, weather, and the mix of nearby agents (vehicles, pedestrians, animals). Small changes along any one axis can subtly alter visual appearance, dynamics, and affordances; combined changes amplify these effects. Despite rapid progress in perception and end-to-end control, a core open question remains: *which factors matter most for out-of-distribution (OOD) robustness, and how should we design the in-distribution (ID) training pipeline support to prepare for them?*

To answer this, we advocate a *factorized* view of distribution shift. Rather than treating "OOD" as a monolith, we explicitly decompose the environment into semantically meaningful axes and evaluate policies under controlled k-factor perturbations—i.e., test conditions that differ from the in-distribution (ID) support on exactly k factors. This yields interpretable robustness profiles: performance as a function of *how many* factors change and *which* factors change. Our analysis shows that such decomposition exposes sensitivities that are obscured by aggregate OOD scores.

Why this matters? Safety-critical deployment hinges on reliable behavior under inevitable distribution shift, e.g., night drives after a model trained at noon, first snow of the season after a summer-only dataset, or an unexpected animal entering the roadway. A factorized evaluation makes robustness *diagnosable*: practitioners can identify, for example, that weather+night degrades steering more than scene+agents, or that balancing time-of-day in the ID set yields larger gains than balancing season, given a fixed data budget. Such insights directly inform data collection, simulation curriculum design, and model selection.

A. Our contributions

Motivated by the discussion above, we present the first *systematic, factorized* experimental study of generalization in vision-based autonomous driving. We quantify how (i) the *training-data factors* included in the ID set, (ii) the *type and number* of test-time distribution shifts, (iii) the *policy architecture* (MLP, CNN, ViT) and the use of *foundation-model features*, and (iv) key *design choices* (data budget, ID diversity vs. scale, single-frame vs. sequence) impact OOD robustness. Specifically, we contribute:

- **A factorized OOD framework.** We formalize the environment as a Cartesian product of factor sets and define k-factor OOD shells via a Hamming distance over factors. This provides a precise and reproducible way to construct ID/OOD splits and to attribute errors to specific axes of variation.
- **Systematic architectural comparison.** Under matched training budgets and protocols, we benchmark FC, CNN, and ViT policies on closed-loop metrics and regression error, reporting robustness as a function of the number and identity of shifted factors.
- **What to include in the ID set.** We vary the ID support along selected factors *at constant data budget, and raising budget in terms of diversity and quantity of the same ID* to quantify which axes are most valuable for

OOD generalization and when broad coverage trades off with ID specialization.

- **Foundation-model features for control.** Using frozen DINO/BLIP-2 patch descriptors with a compact ViT policy head, we isolate the contribution of generic visual features to OOD robustness and analyze how these benefits interact with the choice of ID support.
- **Temporal context.** We compare single-frame policies to sequence-based models to assess whether short histories mitigate specific factor shifts (e.g., adverse weather) and how temporal aggregation complements foundation-model features.

II. Related Work

From modular stacks to end-to-end policies Classical systems used a modular stack (perception→prediction→planning→control) that was reliable yet prone to compounding errors. End-to-end control dates to ALVINN [1] and has since advanced to pixels to steering and learned affordances [2]–[6]. Conditional imitation learning adds high-level commands [7], while later analyses expose limits of pure behavior cloning [8]. We retain the end-to-end setting and ask which architectural biases (MLP/CNN/ViT) and which training distributions best support robustness under controlled shifts.

Generalization and robustness under distribution shift. OOD sensitivity—across towns, weather, lighting, and traffic—has been documented repeatedly; for example, performance drops starkly in new towns or adverse weather even when ID results look strong [8]. Common remedies include domain randomization and augmentation [9], and domain adaptation; yet open-loop gains often fail to translate to closed-loop safety. We complement these lines by *factorizing* shift along semantically meaningful axes (scene, time, season, weather, agents) and measuring robustness as a function of *how many* and *which* factors change.

Foundation models for vision and their use in driving Large-scale pretraining yields image encoders whose features transfer broadly: CLIP aligns images with language for robust zero-shot recognition [10], while DINO learns strong self-supervised ViT representations with emergent semantics [11]; BLIP-2 efficiently couples frozen vision encoders to LLMs [12]. Driving-specific pretraining has leveraged diverse web or fleet data for policy representations [13]–[18]. We operationalize this idea in control by feeding *frozen, patch-wise* features (DINO/CLIP/BLIP-2) to a compact policy head and quantifying when such features improve OOD robustness—and along which factors.

Structured, factorized evaluation Simulation enables controlled manipulations of environment factors. CARLA [19] popularized *New Town* and *New Weather* splits; NoCrash [8] contrasted traffic density and weather to expose failure modes. Data-driven simulators like VISTA [20] reproject real logs to photorealistic, closed-loop scenes, supporting reproducible stress tests. We formalize factorization by defining k-factor OOD *shells* via a Hamming distance over factors, enabling matched-budget, per-axis attribution rather than a single aggregate OOD.

Temporal modeling for control. Temporal context improves driving decisions over single-frame policies. Early FCN–LSTM models fused video history for egomotion prediction [21], and recent end-to-end approaches use spatial–temporal Transformers for perception, prediction, planning [22] or explicit temporal/global reasoning [23]. We directly compare single-frame and sequence-based policies (temporal ViT and RNN heads) under our factorized OOD protocol to reveal which shifts benefit most from temporal aggregation.

Benchmarks and surveys. A growing literature benchmarks end-to-end stacks and catalogs open challenges in robustness, causality, and evaluation [24], [25]. Simulators like CARLA and VISTA remain central for closed-loop, controllable, and reproducible experiments [19], [20]. Our work contributes a *methodological* addition—factorized OOD shells with matched-budget comparisons across architectures, training supports, and temporal context—intended to complement existing benchmarks and inform data curation for real-world deployment.

III. Experimental Setup

We first present the exact questions we aim to study.

A. Key questions we address

This work answers the following practical questions:

- **Q1 — Architecture vs. robustness.** Under matched training protocols, which backbone (FC, CNN, ViT) is intrinsically more resilient to specific factor shifts?
- **Q2 — Role of foundation-model features.** Do frozen patch-wise features from DINO/BLIP-2 confer uniform robustness, or do they target specific axes (e.g., lighting) while being neutral elsewhere?
- **Q3 — Temporal context.** Do short frame histories improve robustness (and for which factors), compared to single-frame policies?
- **Q4 — Which factors matter most?** Among scene, time, season, weather, and agent mix, which single-factor shifts degrade performance the most?
- **Q5 — How many changes can a policy tolerate?** How does performance decay as the number of shifted factors k increases (1, 2, 3), and is the decay monotonic?
- **Q6 — Are factor interactions additive?** Do combinations such as night+snow hurt more than the each of their parts (super-additivity), or less?
- **Q7 — Training data choices.** Under which settings is it better to train a model so that it can generalize to unseen configurations? For example, is it more effective to train on night or day data, in summer or winter conditions, or in rural versus urban environments?
- **Q8 — Data diversity.** Does increasing ID *diversity* (more factor coverage) help OOD generalization?
- **Q9 — Data Scale.** Does increasing *quantity* of a narrow ID help OOD generation?
- **Q10 — Specialization vs. Generalization.** How does increasing the diversity of the ID data trade off special-

ization (performance on a targeted ID subset) against generalization (robustness to OOD shifts)?

B. Task Formulation and Platform

We study closed-loop control for autonomous driving in the *VISTA* simulator [26], a photorealistic, data-driven environment designed for learning-based autonomy. At simulation step t, the agent receives either a single RGB frame $I_t \in \mathbb{R}^{H\times W\times 3}$ or a short sequence of τ frames $I_{t-\tau:t}$, and must predict continuous controls $\hat{\theta}_t$ (steering angle) and $\hat{g}_t$ (throttle/gas). The policy hence realizes a mapping $\pi : I_{t-\tau:t} \mapsto \big(\hat{\theta}_t, \hat{g}_t\big)$. Unless otherwise stated, we use single-frame input (τ=0); the multi-frame setting is introduced in Sec. III-L.

C. Environment Factorization and Distribution Shifts

To make distribution shifts precise, we factorize the environment along five semantically meaningful axes that strongly affect driving:

- **Scene type:** $\mathcal{S} = \{\text{rural}, \text{urban}\}$
- **Season:** $\mathcal{S}\rceil = \{\text{summer}, \text{winter}, \text{spring}, \text{fall}\}$
- **Weather:** $\mathcal{W} = \{\text{dry}, \text{rain}, \text{snow}\}$
- **Time of day:** $\mathcal{T} = \{\text{day}, \text{night}\}$
- **Agents/characters:** traffic actors and non-vehicle entities (e.g.,$\mathcal{A} = \{\text{cars}, \text{animals (etc.)}\}$).

The full environment configuration space is the Cartesian product of the above factor levels

$$\boxed{\mathcal{E} \;=\; \mathcal{S}\times\mathcal{T}\times\mathcal{S}\rceil\times\mathcal{W}\times\mathcal{A}},$$

and we denote a specific environment configuration by a tuple $e = (s, t, \sigma, w, a) \in \mathcal{E}$.Thus, the *in-distribution* (ID) training set is supported on a designated subset $\mathcal{E}_{\text{ID}} \subseteq \mathcal{E}$.

D. Levels of OOD shifts

A *k-factor OOD* test condition, written $e' \in \mathcal{E}^{(k)}_{\text{OOD}}$, differs from the ID support on exactly k factors (with all other factors matched). We evaluate across $k \in \{0, 1, 2, 3\}$ and report results per-factor (which factor changed) and per-k (how many factors changed). For each study, we: (i) specify $\mathcal{E}_{\text{ID}}$; (ii) construct test suites for all admissible k-factor changes that are feasible in VISTA; and (iii) ensure no scene instances from $\mathcal{E}_{\text{ID}}$ appear in OOD tests. When $\mathcal{E}_{\text{ID}}$ is a mixture (Sec. III-K), we stratify by factor to avoid inadvertent leakage.

E. Basic Policy Models and Training

We compare three vision policy backbones trained end-to-end from images to controls: (1) **FC (Fully-Connected):** a shallow MLP operating on spatially downsampled/flattened pixels. (2) **CNN:** a standard convolutional network with global pooling and a control head. (3) **ViT:** a Vision Transformer with patch embedding and a control head. All models output $(\hat{\theta}, \hat{g})$ each step. We optimize the weighted regression Mean Squared Error (MSE) loss

$$\mathcal{L} = \lambda_\theta \,\text{MSE}(\theta, \hat{\theta}) + \lambda_g \,\text{MSE}(g, \hat{g}),$$

with $\lambda_\theta, \lambda_g$ fixed across studies. Unless otherwise noted, image encoders are frozen, and the policy model is trained from scratch on the specified ID distribution.

F. Foundation-model Feature Policies (Sec. III-J).

For the ViT policy head, we also consider policies that consume *frozen* patch-wise features from large-scale pre-trained encoders (DINO and BLIP-2). Given a frame I_t, we extract per-patch descriptors $\{z_{t,p}\}$ (as explained in [27], [28] for DINO, and in [15], [17] for BLIP2) and feed them as tokens into the policy backbone (a compact ViT), which is trained to predict controls while the feature extractor is frozen. This isolates the effect of generic, Internet-scale features on OOD robustness.

G. Evaluation Metrics and Protocol

We report closed-loop performance using **Route completion** (%) and **Infraction counts** (collisions, off road, lane departures). Each configuration uses 100 ID and 100 OOD episodes; we report mean $\pm$ std. Statistical comparisons use paired tests on matched episodes with Holm correction. Splits are fixed per study; hyperparameters are shared unless an ablation changes them.

H. Study S1: Architecture Robustness and OOD Factorized Shifts

We train FC, CNN, and ViT on a fixed ID distribution $\mathcal{E}_{\text{ID}}$ and evaluate on OOD sets with $k \in \{1, 2, 3\}$ factor changes. We analyze: (1) sensitivity curves as a function of k (*how many* factors change), and how it affects each policy model, (2) per-factor robustness: *which* factor changes affect each policy the most (the least), (3) interactions between factors (e.g., night+snow vs. night+rain) and their effect, and inherently, and (4) architecture robustness to all of these factors and level of changes. This isolates architectural inductive biases with identical data budgets and training schedules.

I. Study S2: Effect of the ID Training Distribution

We repeat S1 while redefining $\mathcal{E}_{\text{ID}}$ to examine how the choice of training factors influences OOD generalization. Specifically, we consider two alternative ID configurations from $\mathcal{E}$. In addition, to study the effect of training data size under a fixed configuration, we train the same ViT model on that configuration using 1, 5, and 14 traces.

We quantify how the choice of ID factors impacts robustness profiles from S1 for each architecture, levels and factors of change, by repostring the metric defined in Section III-G.

J. Study S3: Foundation-Model Features with the Best Backbone

From S1 we select the strongest backbone (empirically ViT) and repeat S1 using frozen patch-wise features from DINO, Clip, and BLIP-2. We train compact policy heads (ViT-based) on top of features and evaluate k-factor OOD shifts. We ask: (1) Do generic features improve OOD robustness relative to training from scratch? (2) Are gains uniform across factors (e.g., weather vs. time-of-day)? (3) How do FM-feature policies interact with the ID choice from S2? (4) Which foundation model helps generalize better? on what factors? and what level of change?

K. Study S4: Data Scale and Diversity vs. Specialization

We vary both *quantity* and *diversity* of ID data and compare:

1) **Single-ID→Same-ID:** train and test on the same single configuration (upper bound on specialization).
2) **Single-ID→Other-ID:** train on one configuration, test on a different single configuration (pure shift).
3) **Multi-ID→Single-ID:** train on a mixture of configurations, test on a single target configuration (does diversity harm specialization?).

Each condition is evaluated for $k \in \{0, 1, 2, 3\}$ factor changes, with and without FM features (from S3). This study disentangles the effect of more examples and *varied* examples in the ID set to in and out of distributions.

L. Study S5: Temporal Context—Single Frame vs. Sequence

We compare *single-frame* policies (τ=0) against *multi-frame* policies using short histories (τ>0). For the temporal models we evaluate: (1) **ViT-Temporal:** a ViT backbone operating on per-frame tokens with a lightweight temporal aggregator (e.g., temporal pooling or attention across frames). (2) **RCNN-Temporal:** a CNN encoder with a recurrent head over frame-level embeddings. We re-run S1, S2, and S4 under both temporal settings to quantify how motion cues affect OOD robustness and whether temporal context complements FM features.

M. Implementation Details

We use AdamW (lr 10^{-4}) with cosine decay, standard normalization, and early stopping on validation MSE. Closed-loop evaluation uses a 200-step horizon with shared seeds and routes across models.

IV. Results

In this section, we present our results from a top-level perspective: we begin with a broad summary of the findings and then delve into the detailed analysis.

Terminology. In the following figures and discussion, we adopt the naming convention for each trained model

$$\langle \text{FeatureExtractor} \rangle + \langle \text{PolicyArchitecture} \rangle \ (n_{\text{train}}, n_{\text{val}}),$$

where n_{train} and n_{val} denote the number of traces used for training and validation, respectively. The term ⟨FeatureExtractor⟩ refers to the foundation model (FM) backbone used to extract features (DINO/CLIP/BLIP2). If no feature extractor is used, we write `Pass+<PolicyArchitecture>`. The term ⟨PolicyArchitecture⟩ refers to the policy model (e.g., CNN, fully connected (FC), or Vision Transformer (ViT)). Whenever multiple frames are used in training, we append `T=16` to specify training on 16 consecutive frames. **Environment naming convention.** To concisely encode environment conditions, we use 5\6-character tags (e.g., `RFDNC`, `RSuDDC`, `USuRDC`), where each position denotes a specific semantic factor: **Scene**, **Season**, **Weather**, **Time**, and **Actor**. The first character encodes **Scene** as `R` (rural) or `U` (urban); the second\third, **Season** as `Su` (Summer) or `Sp` (Spring) or `W` (winter) or `F` (Fall); the third\forth, **Weather** as `D` (dry), `R` (Rain), or `S` (Snow); the fourth\fifth, **Time** as `D` (day) or `N` (night); and the fifth\sixth, **Actor** as `C` (Car) or `A` (Animal). As an example, `RSuDDC` refers to a Rural-Summer-Dry-Day-Car; `UFRNA` denotes Urban-Fall-Rain-Night-Animal. This notation enables compact, interpretable reference to specific environmental configurations throughout the paper.

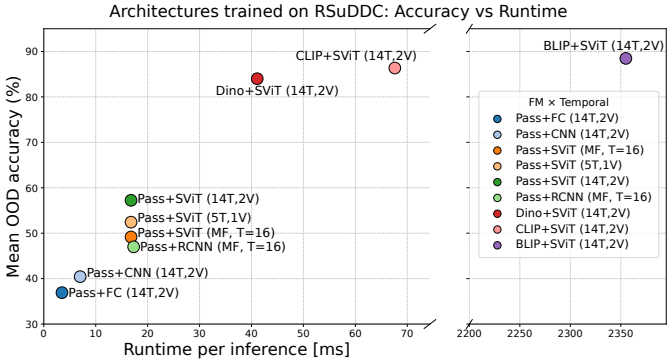


Fig. 1: Accuracy as a function of runtime across model variants.

A. Architectures and training choices

Architecture vs. robustness. For context, we trained CNN, FC, and a simple ViT on the same in-distribution setting (Rural–Summer–Dry–Day–Car) using 14 traces without any feature extractor. In addition, we trained another ViT on only 5 traces from the same configuration to evaluate performance under limited data availability. We then trained a ViT with each of the three feature extractors (DINO/CLIP/BLIP2), and finally, a set of models trained on the last 16 frames as input (see Section III-L for details).

Fig. 1 summarizes runtime per inference vs. mean closed-loop success across OOD scenarios. With a fixed (non-FM) extractor, upgrading the policy architecture markedly improves robustness: Pass+FC and Pass+CNN reach 36.9% and 40.4%, whereas Pass+ViT attains 57.2% at ~ 16.8 ms ($\approx +16.8$ points over CNN for $\approx 2.4\times$ latency). This indicates that:

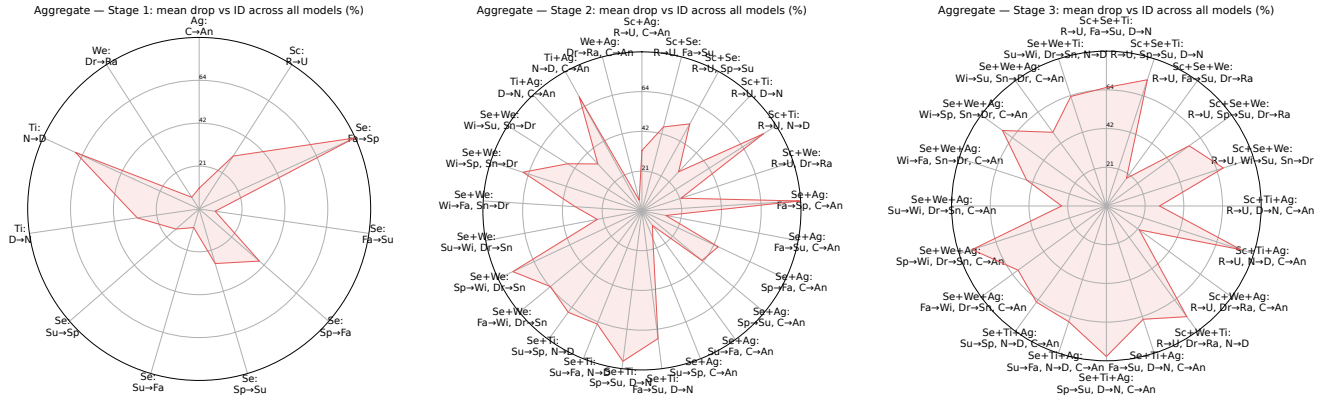
Takeaway 1

Even without external pretraining,**ViT is a stronger policy and substantially boosts out-of-distribution performance compared to a CNN of the same size.**

Role of foundation-model features. Replacing the no-FM extractor with FM features yields a large jump in accuracy: Dino+ViT and CLIP+ViT achieve 84.0% and 86.4%, respectively, while BLIP2+ViT reaches 88.5%. These gains come with higher cost: relative to Pass+ViT (~ 16.8 ms), Dino requires $\sim 2.4\times$ runtime, CLIP $\sim 4\times$, and BLIP2 is over two orders of magnitude slower (~ 2355 ms). Thus

Takeaway 2

Foundation models features offer state-of-the-art robustness but at significant latency, which may limit real-time deployment.



(a) Single factor (b) Double factor (c) Triple factor

Fig. 2: Effect of changes across one, two, and three simultaneous factors. **Key:** Sc, scene (R, rural; U, urban), Se, season (Wi, winter; Sp, spring; Su, summer; Fa, fall), We, weather (Dr, dry; Ra, rain; Sn, snow), Ti, time (D, day; N, night), Ag, agents (C, car; An, animal).

Temporal context. Pass+ViT (MF, T=16) attains 49.2%, which is −8.1 points below Pass+ViT (single-frame, 14T,2V) at 57.2%. Pass+RCNN (MF) reaches 47.0%, outperforming Pass+CNN (40.4%) but still below both single-frame ViT variants (57.2% at 14T,2V; 52.4% at 5T,1V). These results suggest:

Takeaway 3

Naïvely adding multi-frame context does not surpass the best single-frame baseline.

Thus, temporal aggregation must be designed carefully (e.g., alignment and motion-aware fusion) to translate into consistent OOD gains at comparable latency.

B. OOD Environmental Factor Shifts and Their Effect

Which factors matter most? We quantify robustness by the *drop* in closed-loop accuracy relative to each model's in-distribution (ID) baseline. Single-factor shifts (Fig. 2a) reveal a clear ordering of difficulty: switching **scene** from rural→urban and **time** from day→night are the two dominant sources of degradation, each causing about a 31% average drop (Scene: 31.15%, Time: 31.00%). In contrast, changing the **actor** from car→animal is comparatively mild (10.27%), and light **weather** variation (dry→rain) is smallest (6.86%). Seasonal shifts are typically modest (e.g., summer→fall: 9.62%, summer→spring: 15.23%), but certain seasonal discontinuities are severe (fall→spring: 84.63%). Large time reversals also hurt in the opposite direction (night→day: 67.4%). These patterns are consistent with intuitive factors: urban scenes add visual density, clutter, and occlusion; illumination inversions (day↔night) fundamentally alter photometrics and signal-to-noise; by comparison, swapping the actor class perturbs semantics more than geometry and thus affects ego-motion control less; moderate rain impacts appearance but preserves most structural cues.

Two- (Fig. 2b) and three-factor (Fig. 2c) shifts amplify these trends. Pairings that *include time flips* balloon the drop: season+time combinations (e.g., spring→summer with day→night) reach 81.0%, while fall→summer with day→night is 68.8%. Scene+time (rural→urban with day→night) remains challenging, but more moderate on average (28.6%), reflecting partial complementarity between increased texture/occlusion and low-light noise. Triples that include a time inversion similarly dominate (e.g., rural→urban + spring→summer + day→night: 72.9%; season+time+actor with spring→summer + day→night + car→animal: 82.6%). In contrast, combinations *without* time flips-such as scene+weather+actor (rural→urban + dry→rain + car→animal: 22.3%) or season+weather+actor with summer→winter and snow—remain substantially more manageable (≈24.5%).

Takeaway 4

The primary axes of brittleness are illumination (*time*) and distributional density/geometry (*scene*). Seasonal shifts matter when they induce large textural/illumination discontinuities (e.g., fall→spring). Actor identity is least impactful for low-level control, and moderate precipitation alone is relatively benign.

These results suggest: targeted data collection for nocturnal/low-light and dense urban conditions should yield the greatest robustness gains across architectures.

How many changes can a policy tolerate? We summarize tolerance by averaging accuracy at each change level (0, 1, 2, 3 changed factors) across environments. As shown in Fig. 3, models with foundation-model (FM) features remain remarkably stable: BLIP+ViT holds at 88.96/87.55/89.82% for 1/2/3 changes, CLIP+ViT at 86.89/84.77/89.18%, and DINO+ViT at 86.20/81.86/85.33%. In contrast, lightweight non-FM policies degrade steadily as changes accumulate: Pass+ViT (single-frame) drops from 100% → 63.28% → 56.28% → 49.43%, Pass+CNN from 100% → 40.88% → 40.66% → 39.20%, and the multi-frame variants converge near the high-30s by three changes (Pass+ViT MF: 61.0/60.51/46.32/36.91%; Pass+RCNN MF: 59.0/54.02/45.25/39.36%). Three patterns emerge:

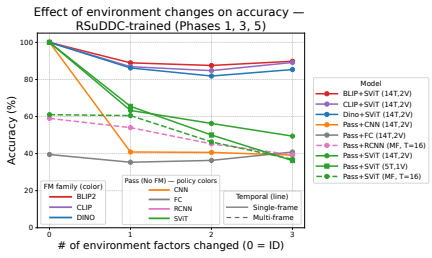


Fig. 3: Accuracy as a function of environment factors changes across model variants.

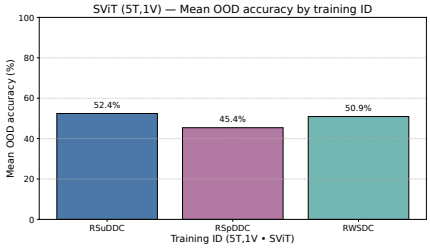


Fig. 4: Training in-Distribution (ID) vs accuracy.

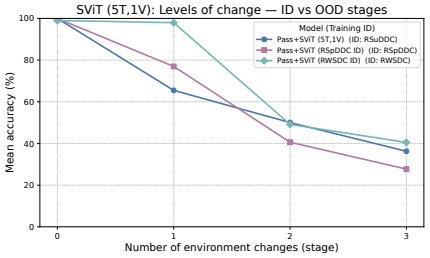


Fig. 5: Number of environment changes vs accuracy.

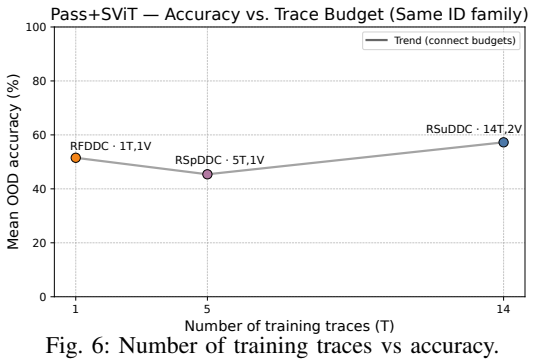


Fig. 6: Number of training traces vs accuracy.

Takeaway 5

(i) The *first* factor change produces the dominant drop for non-FM single-frame models (e.g., Pass+ViT loses 36.7 points at one change, with only modest additional declines thereafter); (ii) temporal (multi-frame) policies are more resilient to the first change, but their robustness erodes once a second change is introduced; and (iii) FM-based extractors maintain performance above $\sim$85% under two or three simultaneous shifts, whereas non-FM models fall below 50% by the third change.

Are factor interactions additive? No, interactions are decidedly non-additive. As can be seen in Fig. 3, on FM models, accuracy at three changes occasionally *rebounds* above the two-change level (e.g., BLIP+ViT: 87.55% $\rightarrow$ 89.82%, CLIP+ViT: 84.77% $\rightarrow$ 89.18%, DINO+ViT: 81.86% $\rightarrow$ 85.33%), indicating that certain factor combinations are not simply "harder" in aggregate. Aggregated per-factor analyses reinforce this: a scene+time shift (rural$\rightarrow$urban with day$\rightarrow$night) yields a mean drop of 28.63%, which is *less* than either scene alone (31.15%) or time alone (31.00%), suggesting partial compensation (e.g., urban lighting at night restoring salient edges). Conversely, pairings that involve *season* with a time flip can be strongly super-additive (e.g., spring$\rightarrow$summer with day$\rightarrow$night: 81.02% drop), and triples that stack season+time+{actor/scene} often remain severe (e.g., spring$\rightarrow$summer + day$\rightarrow$night + car$\rightarrow$animal: 82.62%). Taken together,

Takeaway 6

The first change imposes the largest penalty, and subsequent changes can either *dampen* or *amplify* difficulty depending on the axes involved—illumination flips (day$\leftrightarrow$night) are the key amplifier, while some scene+time pairings are partially mitigating.

C. Training Data choices

Under which settings is it better to train a model? Controlling for architecture and budget (ViT, 5T/1V), we compare three training IDs: *RSuDDC* (rural–summer–dry–day–car), *RSpDDC* (spring–dry–day), and *RWSDC* (winter–snow–day). Averaged over all OODs (Fig. 4), **RSuDDC** yields the best overall robustness (52.4%), followed closely by **RWSDC** (50.9%) and then **RSpDDC** (45.4%). Stage-wise trends (Fig. 5) reveal complementary strengths. With *one* factor changed, the model trained on **RWSDC** is strikingly resilient (97.95%), outperforming both **RSpDDC** (76.96%) and **RSuDDC** (65.48%). This indicates that exposure to adverse winter/snow conditions promotes invariances that translate exceptionally well to single-axis shifts. As changes *compound*, **RSuDDC** and **RWSDC** remain comparatively stable, while **RSpDDC** degrades fastest: at two changes the means are 50.04% (RSuDDC), 49.14% (RWSDC), vs. 40.60% (RSpDDC); at three changes 36.25% (RSuDDC), 40.49% (RWSDC), vs. 27.70% (RSpDDC).

Takeaway 7

Takeaway. If deployments are expected to see *isolated* shifts (one factor at a time), training on winter/snow (**RWSDC**) provides the strongest single-change robustness with only a small tradeoff in global average. For *compounded* shifts, both **RSuDDC** and **RWSDC** are safer choices than **RSpDDC**. For broad, mixed-condition operation where overall average matters most, **RSuDDC** remains the most reliable single-ID training choice.

Training data scale. Holding the architecture fixed (ViT) while varying the number of training traces (Fig. 6) shows a clear—though not strictly monotonic—benefit from scale. Moving from 5T/1V (**RSpDDC**) to 14T/2V (**RSuDDC**) increases the mean OOD accuracy from 45.4% to 57.2% ($\uparrow$ 11.8 points). Even a 1T/1V model (**RFDDC**) reaches 51.5%, outperforming the 5T/1V **RSpDDC** model despite far fewer traces. This indicates that *what* is seen can rival *how much* is seen: scale helps, but content alignment (e.g., exposure to harder seasonal/appearance factors) can offset smaller budgets. Overall, when deployment budgets allow,

TABLE I: Blip+Svit trained on 1 vs. 2 vs. 3 IDs: closed-loop success on ID and OODs (%).

Env	Stage	Blip+Svit (1 Id: Rsuddc)	Blip+Svit (2 Ids: Rsuddc+ +Rsudnc)	Blip+Svit (3 Ids: Rsuddc+ +Rsudnc+ +Usuddc)
RSuDDC (ID)	0	100.0	98.3	98.9
RFDDC	1	96.6	98.6	98.5
RSpDDC	1	98.2	97.1	96.2
RSuDDA	1	97.3	96.2	96.9
RSuDNA	1	—	80.8	85.7
RSuDNC	1	92.1	—	—
USuDDA	1	—	—	87.5
USuDDC	1	60.6	70.1	—
USuDNA	1	—	—	86.8
USuDNC	1	—	89.2	—
USuRDC	1	—	—	93.1
RFDDA	2	99.2	94.7	98.7
RSpDDA	2	99.3	96.5	96.7
RSuDNA	2	86.2	—	85.7
RWSDC	2	99.1	99.4	94.1
USuDDA	2	63.6	—	—
USuDNC	2	84.8	—	—
USuRDA	2	—	—	94.1
USuRDC	2	80.7	78.7	—
RWSDA	3	98.7	99.4	97.7
USuDNA	3	86.0	—	—
USuRDA	3	84.8	81.0	—

higher trace count and view diversity (14T/2V) yields the strongest average robustness.

Takeaway 8

More training traces and views improve robustness on average, but targeted exposure to challenging factors can partially substitute for scale.

Training data diversity. Diversity across IDs (BLIP+ViT; 1 vs. 2 vs. 3 IDs) - See Table I, trades a small loss on the nominal ID for broader OOD gains. On the ID itself (RSuDDC), accuracy is 100% (1 ID), 98.3% (2 IDs), and 98.9% (3 IDs). By contrast, several OODs improve with diversity: RFDDC rises from 96.6% (1 ID) to 98.6% (2 IDs), and USuDDC jumps from 60.6% (1 ID) to 70.1% (2 IDs). Diversity also unlocks strong performance on previously challenging domains—e.g., USuRDC appears at 93.1% for the 3-ID model in stage 1. For comparison, in stage 2 the same OOD appears for the 1-ID and 2-ID models at 80.7% and 78.7%, respectively, which is consistent with the trend that modest ID diversity can unlock stronger robustness on urban regional shifts. Some shifts see mild regressions (e.g., RSpDDA: $99.3\% \rightarrow 96.5\% \rightarrow 96.7\%$; RWSDC: $99.1\% \rightarrow 99.4\% \rightarrow 94.1\%$), consistent with finite-capacity trade-offs when fitting multiple distributions. Net effect: multi-ID training broadens coverage and raises performance on previously weak axes (urban/ USu), while incurring small drops on a few well-covered seasonal/time variants.

Takeaway 9

Multi-ID training broadens coverage and lifts weak axes (urban and USu), while only slightly reducing accuracy on the nominal ID.

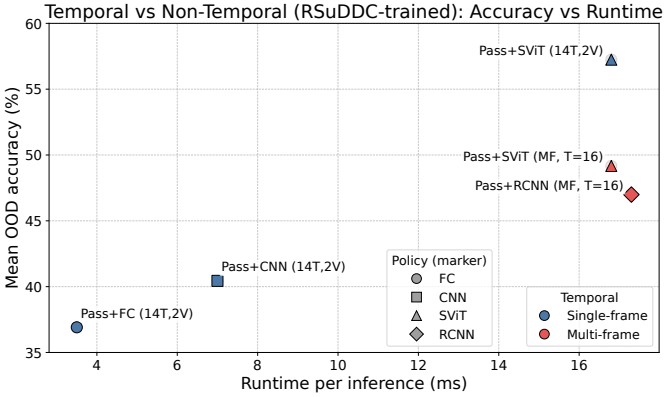


Fig. 7: Temporal vs Non-Temporal: Run-Time vs accuracy.

Specialization vs. generalization. As can be seen from Table I, single-ID training excels on closely related conditions (e.g., RSpDDA at 99.3% with 1 ID), reflecting sharp specialization around the nominal domain. Multi-ID training, however, delivers stronger *generalization* along axes underrepresented by the base ID—e.g., substantial gains on urban and USu variants (USuDDC: $60.6\% \rightarrow 70.1\%$ with 2 IDs; new USuRDC at 93.1% with 3 IDs). In practice, if deployment environments are concentrated near a single domain, specialization preserves peak ID performance. When coverage across diverse scene/region factors is paramount, modest multi-ID diversity yields broader robustness with only a slight reduction on the nominal ID.

Takeaway 10

If deployment is concentrated near one domain, specialization preserves peak ID performance. If coverage across diverse scene and region factors matters, modest multi-ID diversity yields broader robustness with minimal ID loss.

D. Temporal Information

Holding the Pass backbone fixed and comparing models at comparable runtime (See Fig. 7), the single-frame ViT attains the highest mean OOD accuracy, 57.24% at $16.8\,\mathrm{ms}$. The multi-frame ViT, at the same runtime, reaches 49.17% (-8.07 points relative to single-frame ViT), and the multi-frame RCNN at $17.3\,\mathrm{ms}$ reaches 46.99% (-10.25 points). Temporal models nevertheless improve substantially over the simpler single-frame baselines: MF-ViT exceeds FC (36.91%) and CNN (40.44%) by $+12.26$ and $+8.73$ points, respectively, while MF-RCNN exceeds them by $+10.08$ and $+6.55$ points. Overall, in this setting the strongest results come from a high-quality single-frame ViT, with temporal fusion yielding consistent gains over FC/CNN but not surpassing the single-frame ViT at matched compute.

Takeaway 11

At similar runtime, single-frame ViT delivers the best mean OOD accuracy (57.24%). Multi-frame models provide sizable gains over FC/CNN ($\approx +10$ points on average), but do not outperform the single-frame ViT in this training regime.

V. Conclusion

We evaluate out-of-distribution robustness in vision-based driving by factorizing scene, season, weather, time, and agent mix, and measuring performance under controlled k-factor shifts. In VISTA closed-loop tests, ViT heads on BLIP-2, CLIP, or DINO features achieved top OOD accuracy (88.5%, 86.4%, 84.0%), staying above 85% under three shifts, while non-FM baselines fell below 50%. Not all factors are equal: rural→urban and day→night each caused $\sim 31\%$ drops, actor swaps $\sim 10\%$, and light rain $\sim 7\%$, with seasonal flips sometimes severe. Interactions were non-additive: scene+time could be sub-additive, season+time often compounded. Temporal aggregation (Pass+ViT MF 49.2%) did not surpass the best single-frame baseline (57.2%), though it beat FC/CNN. Training scale and design also mattered: more traces improved robustness ($+11.8$ from 5 to 14), targeted hard cases substituted for scale, and multi-ID training broadened coverage (urban OOD $60.6\% \rightarrow 70.1\%$) at minor ID cost. Overall, these results yield practical rules for building reliable closed-loop driving policies under multi-factor shifts. Limitations include the use of simulation and a coarse discrete factorization, future work should validate the main takeaways in real world driving and refine the factors to finer, potentially continuous levels.

References

[1] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 1, 1989, pp. 305–313.

[2] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *arXiv:1604.07316*, 2016.

[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2722–2730.

[4] A. Amini, G. Rosman, S. Karaman, and D. Rus, "Variational end-to-end navigation and localization," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8958–8964.

[5] T.-H. Wang, W. Xiao, M. Chahine, A. Amini, R. Hasani, and D. Rus, "Learning stability attention in vision-based end-to-end driving policies," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 1099–1111.

[6] W. Xiao, T.-H. Wang, R. Hasani, M. Chahine, A. Amini, X. Li, and D. Rus, "Barriernet: Differentiable control barrier functions for learning of safe robot control," *IEEE Transactions on Robotics*, 2023.

[7] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4693–4700.

[8] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9329–9338.

[9] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9650–9660.

[12] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning (ICML)*, 2023.

[13] Q. Zhang, Z. Peng, and B. Zhou, "Learning to drive by watching YouTube videos: Action-conditioned contrastive policy pretraining," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 111–128.

[14] D. Chen and P. Krähenbühl, "Learning from all vehicles," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 222–17 231.

[15] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, and D. Rus, "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 6687–6694.

[16] A. Maalouf, N. Jadhav, K. M. Jatavallabhula, M. Chahine, D. M. Vogt, R. J. Wood, A. Torralba, and D. Rus, "Follow anything: Open-set detection, tracking, and following in real-time," *arXiv preprint arXiv:2308.05737*, 2023.

[17] M. Chahine, A. Quach, A. Maalouf, T.-H. Wang, and D. Rus, "Flex: End-to-end text-instructed visual navigation with foundation models," 2024. [Online]. Available: https://arxiv.org/abs/2410.13002

[18] A. Mallak, E. Aasi, S. Sreeram, T.-H. Wang, D. Rus, and A. Maalouf, "See less, drive better: Generalizable end-to-end autonomous driving via foundation models stochastic patch selection," *arXiv preprint arXiv:2601.10707*, 2026.

[19] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[20] A. Amini, T. Wang, I. Gilitschenski, W. Schwarting, Z. Liu, S. Han, S. Karaman, and D. Rus, "VISTA 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2419–2426.

[21] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2174–2182.

[22] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 533–549.

[23] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, and Y. Liu, "Reasonnet: End-to-end driving with temporal and global reasoning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 723–13 733.

[24] Z. Zhu and H. Zhao, "A survey of deep rl and il for autonomous driving policy learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 043–14 065, 2022.

[25] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *arXiv:2306.16927*, 2023.

[26] A. Amini, T.-H. Wang, I. Gilitschenski, W. Schwarting, Z. Liu, S. Han, S. Karaman, and D. Rus, "Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2419–2426.

[27] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep ViT features as dense visual descriptors," *arXiv preprint arXiv:2112.05814*, 2021.

[28] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari *et al.*, "Conceptfusion: Open-set multimodal 3d mapping," *arXiv preprint arXiv:2302.07241*, 2023.