WHEN SHOULD THE DECISION CHANGE? COUNTERFACTUAL PROBING OF VISION-LANGUAGE MODELS FOR AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

In autonomous driving, a small change in another road user's behavior can make a previously appropriate ego decision unsafe. How reliably do vision-language models adapt their driving decisions to such changes? We introduce DRIVEWHEN, a counterfactual probing framework that examines model decisions across controlled alternatives to the same traffic scene. The framework combines scene interventions, context-dependent action requirements, and paired evaluation to assess both responsiveness when a different action is required and stability when the original action remains valid. Grounded in real nuScenes scenes, we construct paired videos through four atomic interventions on existing scene entities: leftward and rightward lateral intrusions designed to introduce hazards, forward departures designed to alleviate hazards, and appearance replacements serving as controls. Human review independently assigns the required ego action to each original and counterfactual video, allowing the same motion intervention to warrant either retaining or revising the original action across different traffic contexts. Evaluation of five vision-language models reveals frequent failures to revise their actions when the altered traffic situation requires a different response. These findings expose a gap between producing a decision for an individual scene and responding to changes in the interactions that inform that decision. DRIVEWHEN tests whether models can distinguish visually similar traffic situations and revise their decisions only when the altered context warrants it, providing a targeted diagnostic of response reliability in safety-critical driving.

1 INTRODUCTION

Vision-language models (VLMs) increasingly connect visual understanding to reasoning and action in autonomous driving (Xu et al., 2024; Shao et al., 2024; Tian et al., 2025; Hwang et al., 2025). Because driving is safety-critical, their decisions must remain appropriate as traffic evolves. A local change in another road user's behavior can require a new ego action even when the road layout remains similar. Evaluating whether VLMs revise or retain decisions accordingly is therefore essential.

Yet reliability under decision-relevant changes in complex traffic remains unclear. Conventional benchmarks score videos independently, allowing high accuracy from scene priors or frequent labels even when a model ignores changed interactions. Accuracy and Macro-F1 indicate whether predictions match labels, not whether the model responds to the new evidence. DriveBench identifies reliability issues in visual inputs, data, and metrics (Xie et al., 2025), while Driving on Memory shows that strong scores can arise from previously recorded observations without viewing current traffic (Löwens et al., 2026). Existing benchmarks therefore provide limited evidence of correct responses to interaction changes, motivating a controlled evaluation framework.

Counterfactual video generation offers a promising route to this evaluation, but turning an edited clip into a valid decision test poses two connected challenges. First, a complete evaluation framework must connect a controlled intervention, verification of the resulting video, reviewed actions for both versions, and paired criteria for determining whether the model should retain or revise its decision. The intervention itself cannot serve as the expected label, because the same motion change may or may not alter the appropriate ego action. Second, benchmark construction requires a recorded scene

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **When should a counterfactual edit change the driving decision?** **(a)** Controlled interventions create alternative situations, but the edit alone does not determine the correct response; colored alternatives are decision questions rather than model predictions. **(b)** Reviewed examples show that added and removed risk can each require retention or revision. A red light preserves `BRAKE_OR_STOP`, while a near cut-in changes `SLOW` to `BRAKE_OR_STOP`; an already moving ego vehicle preserves `KEEP`, while a departing lead vehicle changes `SLOW` to `KEEP`. Arrows report the reviewed action pairs.

with a suitable existing object, enough space for a controlled trajectory, a visually coherent edited video, and a clear ego-action judgment in both versions. Together, these constraints make benchmark construction demanding: every retained pair must preserve the surrounding traffic context, realize the intended intervention clearly and coherently, and support a reliable ego-action judgment for both the original and counterfactual videos.

The correct response therefore depends on the complete traffic context. As Figure 1 shows, added or removed risk can each require either adjustment or retention. Consequently, unchanged predictions divide into *correct retention* and *missed adjustment*, while changed predictions divide into *correct adjustment* and *unnecessary change*.

We address these challenges with DRIVEWHEN, which turns controlled edits to recorded driving scenes into context-dependent decision tests. Using CompoSIA (Zhan et al., 2026), we intervene on objects already present in a scene and retain edited videos that pass visual and semantic review. Human reviewers assign the appropriate ego action independently to the original and counterfactual videos; the relation between these actions then determines whether a model should retain its decision, become more conservative, or become less conservative. Our contributions are:

- **A unified framework for context-dependent decision evaluation.** We connect controlled object intervention, video verification, reviewed ego actions for both videos, and paired behavioral criteria within a single evaluation framework. This formulation distinguishes necessary decision adjustment from correct retention instead of treating every intervention as a request to change the action.
- **A reviewed counterfactual benchmark grounded in real driving scenes.** We screen recorded scenes for suitable existing objects, apply four predefined trajectory or appearance interventions, and retain 150 video pairs after reviewing edit quality and action clarity. The benchmark includes motion edits that require different responses in different traffic contexts, preventing the intervention type from acting as the answer label.
- **Evidence that improved action scores need not imply better counterfactual responses.** Across five VLMs, directional response on pairs requiring change reaches at most 13.16%.

Task-specific action supervision improves paired accuracy by 8.15 percentage points while reducing directional response by 3.52 points, exposing a gap that aggregate action scores alone can obscure.

2 Related Work

2.1 Vision-language driving models

End-to-end driving learns planning from sensor inputs, with query-based and vectorized architectures integrating perception and planning Hu et al. (2023b); Jiang et al. (2023). VLMs extend these systems toward open-ended understanding and interactive question answering Hwang et al. (2025); Sima et al. (2024), as well as language-guided control Xu et al. (2024); Shao et al. (2024); Tian et al. (2025). Driving benchmarks assess visual question answering and structured reasoning Marcu et al. (2024); Nie et al. (2024); Wang et al. (2025); Xie et al. (2025). In parallel, generative world models predict scene evolution Hu et al. (2023a); Yang et al. (2024); Gao et al. (2024), while recent approaches model dependencies among scene elements and their interactions Yan et al. (2026); Moon et al. (2026). These directions supply models for understanding, planning, and generation. Our focus is the appropriateness of action recommendations under controlled changes, which requires comparable alternative scenes and explicit expected behavior.

2.2 Driving behavior evaluation

Behavioral testing examines model capabilities through systematic perturbations with known expected outcomes. Contrast sets probe local decision boundaries with semantic edits (Gardner et al., 2020), and CheckList organizes tests around functionality, invariance, and directional expectations (Ribeiro et al., 2020). For driving, DriveBench motivates input controls and fine-grained diagnostics (Xie et al., 2025); Driving on Memory uses stored scene observations as diagnostic controls (Löwens et al., 2026); and CCTVBench pairs real accident videos with generated alternatives to evaluate contrastive consistency over mutually exclusive hypotheses (Zhou et al., 2026). These studies demonstrate the value of controlled comparisons beyond independent-sample scores. We apply this principle to ego-action recommendations: reference actions are established for both contexts, and their relation determines whether a model should adjust or retain its decision.

2.3 Counterfactual construction

Counterfactual explanations typically search for small input modifications that change a model prediction (Wachter et al., 2018; Mothilal et al., 2020). In vision, STEEX constrains edits semantically, OCTET optimizes object-centric position and appearance representations, and SAFE uses saliency to guide explanations for driving systems (Jacob et al., 2022; Zemni et al., 2023; Samadi et al., 2023). These approaches primarily operate on static images and connect construction to a target model's prediction change. Video generation adds temporal and geometric control: Ctrl-V uses bounding-box sequences, DriveEditor supports 3D-guided object editing, and CompoSIA disentangles scene structure, object identity, and ego behavior (Luo et al., 2024; Liang et al., 2025; Zhan et al., 2026). DriveWHEN builds on these video-editing capabilities to construct context-dependent driving-decision tests. Candidate retention depends on edit validity and whether the traffic context supports a clear reference action; a prediction change by a particular model is not an acceptance criterion.

3 The DriveWHEN Framework

Figure 2 presents DriveWHEN in three stages. First, we separate the traffic factors that remain fixed from the state of the object being edited. Second, we modify that object's trajectory or appearance and synthesize a counterfactual video while preserving the surrounding scene. Third, we assign an ego action to each video and evaluate both action correctness and whether the model retains or revises its decision as required.

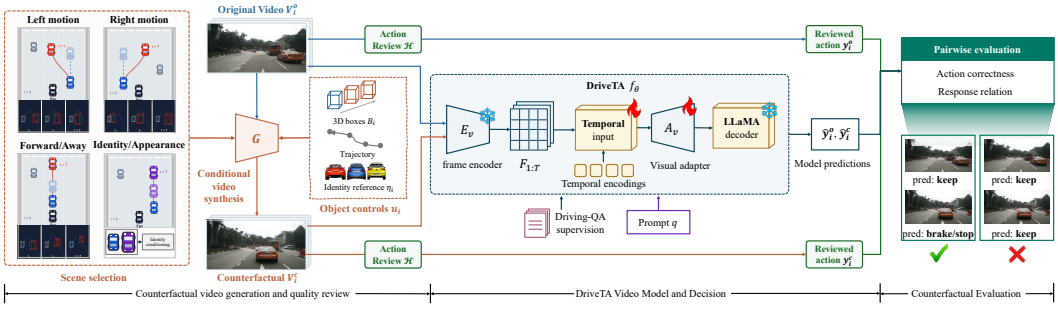

Figure 2: **The DriveWHEN framework.** (a) Four interventions modify selected scene objects through trajectory or appearance controls. Projected geometry, scene conditions, and appearance references guide video synthesis. Generated clips undergo quality review and reference-action annotation. (b) The same model processes each video separately; evaluation checks action correctness and the required response relation. DriveTA extends an image-based driving model to video input through frame encoding, temporal embeddings, and driving-QA fine-tuning. (c) Reviewed actions and model predictions are compared to determine whether the model correctly retains or adjusts its decision.

3.1 Driving Decision Models

Represent a short traffic sequence as $S = (C, Z)$, where C contains the road layout, signals, observed ego states, and non-target objects, and Z contains the selected object's state sequence, including its 3D trajectory and appearance. The observation function $\mathcal{O}$ maps the scene state to sensor input X, and a driving model predicts an action or trajectory $\hat{d}$ under task or navigation conditions q:

$$S = (C, Z), \qquad X = \mathcal{O}(S), \qquad \hat{d} = \pi_\theta(X, q). \tag{1}$$

We instantiate this interface with front-view video $V = \mathcal{O}_{\text{front}}(S)$ and a short-horizon longitudinal action. The visual encoder E_θ maps the video to feature representation z, and the language decoder D_θ combines z with prompt q to produce the decision. We use f_θ to denote the complete video-to-action model:

$$z = E_\theta(V), \qquad \hat{y} = D_\theta(z, q) = f_\theta(V, q), \qquad \hat{y} \in \mathcal{A}. \tag{2}$$

The action space is $\mathcal{A} = \{\texttt{KEEP}, \texttt{SLOW}, \texttt{BRAKE_OR_STOP}\}$. `KEEP` denotes maintaining speed or requiring no additional deceleration; `SLOW` denotes slowing down, maintaining cautious low speed, or starting slowly when appropriate; and `BRAKE_OR_STOP` denotes braking or remaining stopped. We ask how changing Z—the selected object's trajectory or appearance—alters $\hat{y}$ while C is held fixed. The relevant traffic context is the resulting scene (C, Z'), rather than the edited object alone.

3.2 Object-Level Counterfactual Interventions

We create alternative interactions by editing objects and trajectories already present in real traffic scenes. DriveWHEN applies predefined 3D motion or appearance changes to the selected object and renders the resulting continuous video. The reviewed action is then determined from the complete edited traffic context.

Counterfactual evaluation problem. Let $\mathcal{T}_{u_i}$ denote an intervention operator parameterized by control u_i. It transforms the selected object state Z_i while leaving the shared scene factors C_i unchanged. The original and counterfactual scenes and observations are

$$\begin{aligned} S_i^o &= (C_i, Z_i), & V_i^o &= \mathcal{O}_{\text{front}}(S_i^o), \\ S_i^c &= (C_i, \mathcal{T}_{u_i}(Z_i)), & V_i^c &= G(C_i, \mathcal{T}_{u_i}(Z_i)), \end{aligned} \tag{3}$$

where $\mathcal{O}_{\text{front}}$ extracts the recorded front-view observation and G renders the edited video. The common C_i specifies the intended shared conditions; their preservation in the rendered clip is checked during quality review. Let Q verify that the edit is visible, the video is usable, and the traffic

context supports a clear action judgment. Let $\mathcal{H}$ assign the reviewed action without using the tested model. The resulting evaluation set is

$$\mathcal{D}_{\text{cf}} = \{(V_i^o, V_i^c, y_i^o, y_i^c) \mid Q(V_i^o, V_i^c) = 1, \quad y_i^v = \mathcal{H}(V_i^v),\ v \in \{o, c\}\}. \tag{4}$$

Define the action rank by $r(\texttt{KEEP}) = 0$, $r(\texttt{SLOW}) = 1$, and $r(\texttt{BRAKE_OR_STOP}) = 2$. The required response relation is

$$g_i = \operatorname{sgn}[r(y_i^c) - r(y_i^o)] \in \{-1, 0, +1\}, \tag{5}$$

where $g_i = +1$ means a more conservative action, $g_i = -1$ a less conservative action, and $g_i = 0$ retention. Thus, retention is a required response for some counterfactual pairs rather than an invalid or ineffective edit.

Trajectory-grounded control. We next instantiate the abstract control u_i as a trajectory or appearance edit applied to an existing scene object. Write $Z_i = (B_i, \eta_i)$, where $B_i = \{b_{i,t}\}_{t=0}^{T-1}$ is the observed 3D box sequence and η_i is the appearance condition. Each box contains its center, dimensions, and heading, $b_{i,t} = (x, y, z, \ell, w, h, \psi)_{i,t}$. We specify

$$u_i = \big(\{\Delta b_{i,t}\}_{t=0}^{T-1}, \eta_i^{\text{ref}}\big), \qquad \mathcal{T}_{u_i}(Z_i) = \big(\{b_{i,t} + \Delta b_{i,t}\}_{t=0}^{T-1}, \eta_i^{\text{ref}}\big). \tag{6}$$

Motion edits add predefined center displacements to the observed trajectory while retaining the original appearance condition. Appearance edits replace the appearance reference while preserving the trajectory.

There are four atomic edits: move left, move right, move forward/away, and replace appearance. In the first-frame ego coordinate system, x points right and y points forward. For $T = 33$ frames and $t = 0, \ldots, T-1$, the center displacements, in meters, are

$$(\Delta x_t, \Delta y_t) = \begin{cases} (-3.5 \sin \phi_t, -0.2t), & \text{move left}, \\ (+3.5 \sin \phi_t, -0.6t), & \text{move right}, \\ (0, +0.25t), & \text{forward-away}, \\ (0, 0), & \text{appearance replacement}, \end{cases} \qquad \phi_t = \frac{\pi t}{2T}. \tag{7}$$

Here t is a frame index. The lateral edits move a vehicle from an adjacent lane toward the ego lane, while the forward-away edit increases the distance between a lead vehicle and the ego vehicle. Appendix A gives the target-position and free-space requirements. Edited boxes are projected into the front camera and passed to CompoSIA with appearance and scene conditions (Zhan et al., 2026). The recorded ego trajectory and non-target conditions are retained. Review removes ineffective edits and visible artifacts; the original and counterfactual videos are then each assigned a reviewed action.

Appendix Figure 5 shows the four intervention mechanisms and demonstrates that edit type does not determine the action label: the same motion edit can preserve or change the reviewed ego action depending on the scene.

3.3 Constructing the DriveWHEN Benchmark

We use the 850 scenes of nuScenes v1.0-trainval (700 training and 150 validation scenes) (Caesar et al., 2020). From each scene's first sample, we extract 33 consecutive front-view frames from the 12 Hz stream with interpolated 3D annotations. Not every scene can support every edit. A lateral edit requires a vehicle in an adjacent region and enough free space in the ego lane for it to move without overlapping other vehicles. A forward-away edit requires a lead vehicle and enough open space ahead. Scenes without a suitable target or enough space are skipped. In total, 725 scenes support at least one edit, yielding 1,121 generation tasks, where each task is one scene–intervention combination. Human review checks target presence, edit realization and visibility, hallucinations, deformation, interpenetration, and disruptive non-target changes. Retained clips must also support a clear traffic-action judgment. Reviewing each complete original and counterfactual video produces 150 usable pairs from 138 source scenes.

We call an action requirement context-dependent when the intervention category alone is insufficient to determine whether the ego action should change. The relation is derived from the reviewed actions for the complete original and counterfactual traffic contexts.

Table 1: Composition of the DRIVEWHEN benchmark by intervention and context-dependent action requirement.

Edit	Pairs	Unchanged	More conservative	Less conservative
Appearance	62	62	0	0
Forward-away	14	3	0	11
Move left	55	7	48	0
Move right	19	2	17	0
Total	150	74	65	11

Table 1 shows that 76 of the 88 motion-edited pairs require an action change, while 12 retain the original action. All 62 appearance-edited pairs require retention. The retained motion cases are important: a lateral intrusion may leave `BRAKE_OR_STOP` unchanged when a red light already requires stopping, and a departing lead vehicle may leave `KEEP` unchanged when the ego vehicle is already moving appropriately. Conversely, nearby cut-ins and the removal of a blocking lead vehicle often change the reviewed action. Motion editing is therefore not itself a label for changing the decision.

The target objects in DRIVEWHEN are not limited to passenger cars. Across the benchmark, edited objects include cars and trucks as well as pedestrians, bicycles, motorcycles, and traffic cones. The data also vary in object position, motion path, interaction distance, and when the edited behavior becomes visible, as illustrated in Appendix Figure 7. This variation prevents the task from reducing to a single cut-in template or a change at a fixed video position.

3.4 DRIVETA: ADAPTING AN IMAGE-BASED DRIVING MODEL TO VIDEO

To include an image-based driving baseline, we develop DRIVETA, a temporal extension of the DriveLM challenge LLaMA-Adapter V2 baseline (Sima et al., 2024; Gao et al., 2023). CLIP (Radford et al., 2021) encodes each front-view frame; temporal embeddings and visual adaptation modules expose the sequence to LLaMA. Driving-QA supervision uses DriveLM front-view question–answer data to connect multi-frame observations with driving language.

DriveTA is adapted on 696 DriveLM scenes using eight frames per video while CLIP and LLaMA remain frozen. For the common 33-frame evaluation, we interpolate the learned temporal positions to the full sequence. This stage uses no DriveWHEN labels; Appendix D gives the complete training configuration, and Section 4.4 studies DriveWHEN-specific supervision separately.

3.5 CONTEXT-DEPENDENT PAIRED BEHAVIORAL EVALUATION

For each pair, the same model weights and action prompt are used in separate inference calls, yielding $\hat{y}_i^o = f_\theta(V_i^o, q)$ and $\hat{y}_i^c = f_\theta(V_i^c, q)$. After decoding semantic actions, we compute

$$c_i = \mathbf{1}[\hat{y}_i^o = y_i^o \ \wedge \ \hat{y}_i^c = y_i^c], \qquad \hat{g}_i = \mathrm{sgn}[r(\hat{y}_i^c) - r(\hat{y}_i^o)]\,. \tag{8}$$

The indicator c_i checks both action predictions, while comparing $\hat{g}_i$ to g_i checks the response relation. Correct retention requires $g_i = 0$ and $c_i = 1$; correct adjustment requires $g_i \neq 0$ and $c_i = 1$. A changed prediction with $g_i = 0$ is an unnecessary change, whereas an unchanged prediction with $g_i \neq 0$ is a missed adjustment. Relation agreement alone is insufficient: a reference pair `KEEP`→`SLOW` and prediction pair `KEEP`→`BRAKE_OR_STOP` have the same direction but different counterfactual actions. We aggregate these complementary criteria in Section 4.1; Appendix B.2 illustrates the scoring rules.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and inputs. We evaluate DRIVETA, Qwen2.5-VL-7B, LLaVA-Video-7B, Qwen3-VL-8B, and LLaVA-OneVision-7B (Bai et al., 2025b; Zhang et al., 2025; Bai et al., 2025a; Li et al., 2025).

Table 2: Comparison on the full DRIVEWHEN benchmark of 150 video pairs (%). Every learned model receives the same 33 front-view frames. Results are averaged after decoding three semantically equivalent action-option orders. Bold marks the best learned model; Always KEEP is a diagnostic policy rather than a learned baseline.

Model	PairAcc	Macro-F1	Bal. Acc.	DIR-strict	App. PairAcc
DRIVETA	**18.22**	28.60	**41.61**	7.46	**42.47**
Qwen2.5-VL-7B	13.78	21.54	33.60	8.77	29.57
LLaVA-Video-7B	15.33	**30.12**	39.75	**13.16**	32.26
Qwen3-VL-8B	12.44	22.12	36.26	7.02	28.49
LLaVA-OneVision-7B	17.56	29.40	39.20	8.77	37.63
Always KEEP	20.67	19.39	33.33	0.00	45.16

The latter four use publicly available pretrained weights. The full evaluation uses all 150 pairs, with each model receiving the same 33 front-view frames per video and predicting one of the three semantic actions. Appendix B.2 documents the prompt format, output parsing, and aggregation protocol.

Metrics. We report five primary metrics. Macro-F1 and balanced accuracy are the class-averaged F1 and recall. Three paired metrics assess the task-specific requirements:

$$\text{PairAcc} = \frac{1}{N}\sum_{i=1}^{N} c_i, \qquad N = 150, \tag{9}$$

$$\text{DIR}_{\text{strict}} = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} \mathbf{1}[\hat{g}_i = g_i], \qquad \mathcal{C} = \{i : g_i \neq 0\}, \quad |\mathcal{C}| = 76, \tag{10}$$

$$\text{Appearance-PairAcc} = \frac{1}{|\mathcal{A}_{\text{app}}|}\sum_{i\in\mathcal{A}_{\text{app}}} c_i, \qquad |\mathcal{A}_{\text{app}}| = 62. \tag{11}$$

PairAcc is the fraction of pairs for which both action predictions are correct. DIR-strict measures directional agreement on the 76 action-changing pairs and assigns zero credit to an unchanged prediction. Appearance PairAcc measures exact paired correctness on the 62 appearance-edited pairs, whose reviewed actions require retention. Invalid outputs count as errors without reducing the denominator. All evaluated models produce valid semantic outputs under the primary protocol. We additionally report relation-specific recall to distinguish retention from the two adjustment directions.

4.2 COMPARISON RESULTS

Table 2 compares five learned models under the same 33-frame protocol, jointly measuring per-video action recognition, exact paired correctness, directional adjustment, and appearance retention.

The learned models reach only 12.44–18.22% PairAcc and 7.02–13.16% DIR-strict. DRIVETA leads the learned models in PairAcc, balanced accuracy, and appearance retention, whereas LLaVA-Video leads in Macro-F1 and directional response. No learned model leads on both paired correctness and directional response.

The Always KEEP policy sharpens this distinction. It obtains 20.67% PairAcc—higher than every learned model—yet scores zero on DIR-strict. Repeated labels allow a constant prediction to solve some pairs without responding to the intervention. PairAcc is therefore informative only when read together with the direction-sensitive metric. DRIVEWHEN requires a model to determine when an altered interaction matters, rather than rewarding a decision change after every edit.

4.3 COUNTERFACTUAL RESPONSE AND VISUAL–TEMPORAL EVIDENCE

Figure 3a separates unchanged, more-conservative, and less-conservative pairs. The learned models recall 79.73–88.29% of unchanged relations, but only 5.64–13.33% of more-conservative and 9.09–15.15% of less-conservative responses. The retention–adjustment gap appears across all five

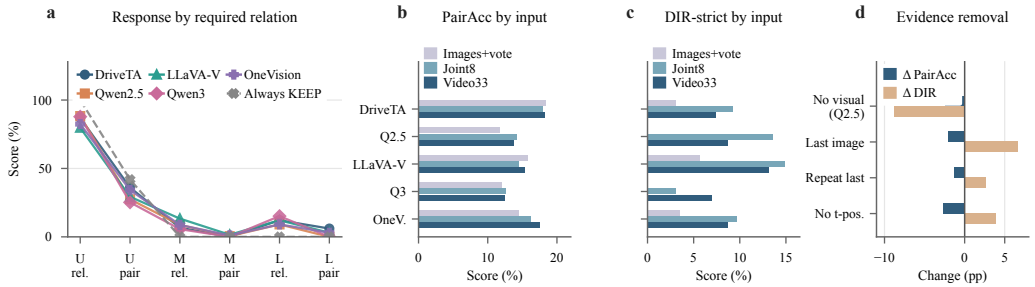

Figure 3: **Counterfactual response and visual–temporal evidence.** (a) Compact model profiles for unchanged (U), more-conservative (M), and less-conservative (L) pairs. "rel." is relation recall; "pair" is exact PairAcc. (b,c) Ablation on temporal information: PairAcc and DIR-strict when eight sampled frames are judged independently and voted (Images+vote), presented jointly (Joint8), or replaced by the complete 33-frame video (Video33). (d) Evidence controls compare the full video with no visual input, the final image alone, the final image repeated across 33 positions, and video input without temporal-position embeddings; positive values favor the control over the full-video result.

Table 3: Task-specific adaptation on 75-pair test set (%). DIR-strict uses the 41 pairs that require an action change in the recorded adaptation evaluation.

Method	PairAcc	Macro-F1	Bal. Acc.	DIR-strict
DriveTA	16.44	27.87	39.68	**8.13**
DriveTA + action SFT	**24.59**	**36.05**	**40.59**	4.61
DriveTA + pairwise loss	23.70	35.78	39.85	5.42

architectures. Exact PairAcc is lower still, showing that a prediction may move in the correct direction without reaching the correct action.

Panels (b–d) test visual–temporal evidence. Jointly presenting eight frames improves DIR-strict over independent frame voting for every model, whereas increasing from eight to 33 frames yields no uniform gain. We retain all 33 frames because sparse sampling may miss brief peripheral events. Removing visual input drives Qwen2.5-VL's DIR-strict to zero, while final-frame and temporal-position controls affect DRIVETA's action and directional scores differently. Preserving temporal evidence and using it for counterfactual response are therefore distinct capabilities; complete results appear in Appendix E.

4.4 TASK-SPECIFIC ADAPTATION

The base DRIVETA accepts video but is not trained for the three-action DriveWHEN task. We split the 150 pairs by source scene into 60 training, 15 validation, and 75 test pairs. Action SFT supervises each video's action; the pairwise loss additionally encourages equal distributions for matching labels and the annotated direction otherwise.

Action SFT raises PairAcc by 8.15 points and Macro-F1 by 8.18, with BRAKE/STOP recall increasing from 0 to 50.88% (Table 3). DIR-strict nevertheless falls from 8.13% to 4.61%, and the pairwise loss adds no further PairAcc gain. Task supervision improves action recognition without producing better counterfactual response; Appendix F shows the class-recall redistribution.

4.5 QUALITATIVE ANALYSIS

Figure 4 makes the paired evaluation concrete through four representative outcomes: correct adjustment, missed adjustment, correct retention, and unnecessary change. Each case shows the original and counterfactual videos at two aligned, manually selected timestamps. The model nevertheless

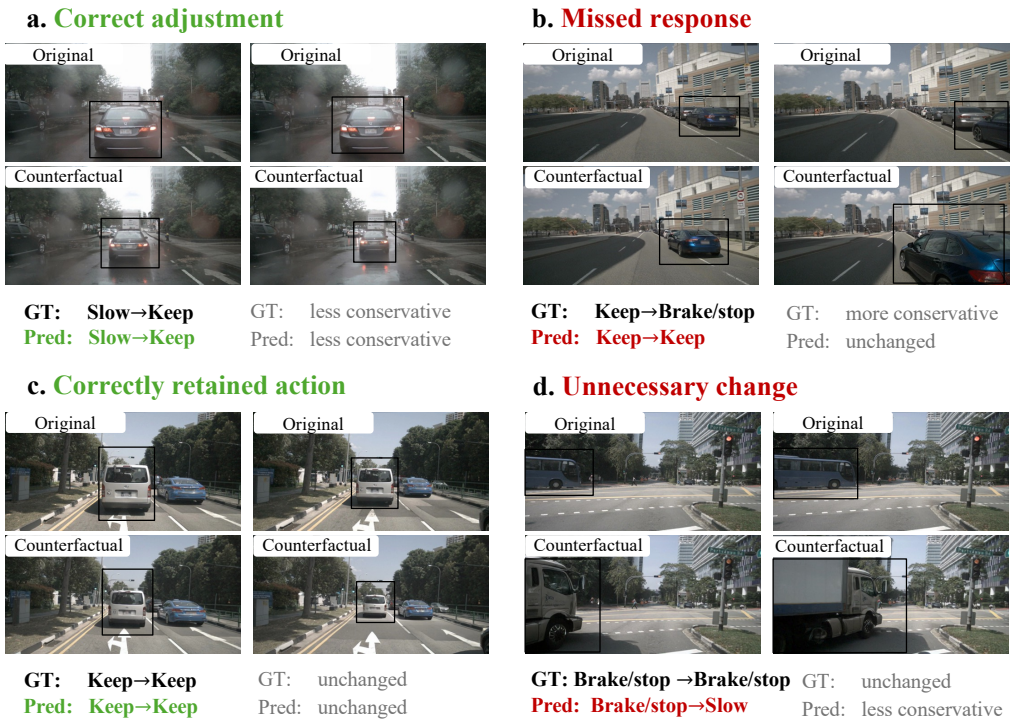

Figure 4: **Four counterfactual response patterns.** Each case contains an Original row and a Counterfactual row at two aligned timestamps.

processes each complete video independently; the displayed frames serve only to visualize the relevant interaction.

Figure 4a and Figure 4b examine pairs for which the appropriate action should change. In Figure 4a, the model correctly relaxes its action after the leading vehicle moves away and releases the original constraint. In Figure 4b, it fails to respond when a lateral intrusion introduces a new hazard. Figure 4c and Figure 4d examine the complementary requirement of retaining a valid action. A departing vehicle need not alter an existing KEEP decision when the ego vehicle can already proceed, while a lateral intervention need not change BRAKE/STOP when a red light remains the binding constraint. These cases demonstrate that the behavioral consequence of an intervention is determined by the complete traffic context: a model should change its action when the edited interaction becomes decision-relevant and preserve it when the existing constraint still governs the scene. Additional cases in Appendix C distinguish stable but incorrect predictions from correct invariance and show how a model may respond in the required direction without reaching the correct final action.

5 CONCLUSION

We introduced DRIVEWHEN, a counterfactual probing framework that connects controlled interventions on existing road users with reviewed ego-action labels and paired behavioral evaluation. By assigning an action to both the original and counterfactual videos, DRIVEWHEN evaluates whether a model should revise its decision or retain it under the complete traffic context. Across five VLMs, the benchmark exposes a consistent gap between action recognition and counterfactual response: models preserve decisions far more readily than they execute more- or less-conservative adjustments. Counterfactual pairs make this capability directly measurable by separating correct retention from repeated error and necessary adjustment from arbitrary change. DRIVEWHEN therefore provides a targeted diagnostic of whether vision–language driving models respond to the interaction that actually governs the ego decision, strengthening evaluation for safety-critical driving.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and revision, English-language editing, literature organization, figure preparation, and development of parts of the data-generation and experimental workflow. CompoSIA was used to synthesize the counterfactual videos described in Section 3.2. Reference actions were assigned through human review independently of the evaluated VLMs. Reported experimental values are taken from the experimental records; illustrative responses in Figure 1 are explicitly distinguished from measured predictions. The authors are responsible for the final manuscript, experimental claims, and accompanying artifacts.

REPRODUCIBILITY STATEMENT

The anonymized supplement includes the frozen 150-pair manifest, reviewed action annotations, inference prompts, evaluation code, and the temporal implementation of DRIVETA. The provided scripts derive action relations from the paired labels and compute all reported metrics. Section 3 and Appendices A–D document the framework, benchmark construction, scoring, and video adaptation. The generated videos will be released in accordance with the applicable nuScenes data-use terms.

REPRODUCIBILITY STATEMENT

We provide an anonymized supplementary package containing the frozen manifest and action annotations for all 150 counterfactual pairs, the three semantically equivalent option mappings and prompts, the evaluation code, and the project-specific temporal implementation of DRIVETA. The evaluation code recomputes the gold relation directly from the paired semantic actions and implements the metrics reported in the paper. Section 3 formalizes the interventions and evaluation criteria, while Appendices A, B.2, and D document benchmark construction, scoring, and driving-video adaptation. The generated videos are derived from nuScenes and are excluded from the anonymous review package; their public release will follow the applicable data-use terms.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL Technical Report. arXiv preprint, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. arXiv preprint, 2025b.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11621–11631, 2020.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, et al. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. arXiv preprint, 2023.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 91560–91596, 2024.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, 2020. Association for Computational Linguistics.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A Generative World Model for Autonomous Driving. arXiv preprint, 2023a.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-Oriented Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862, 2023b.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: End-to-End Multimodal Model for Autonomous Driving. *Transactions on Machine Learning Research*, 2025.

Paul Jacob, Éloi Zablocki, Hédi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. STEEX: Steering Counterfactual Explanations with Semantics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 387–403, 2022.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8350, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*, 2025.

Yiyuan Liang, Zhiying Yan, Liqun Chen, Jiahuan Zhou, Luxin Yan, Sheng Zhong, and Xu Zou. DriveEditor: A Unified 3D Information-Guided Framework for Controllable Object Editing in Driving Scenes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5164–5172, 2025.

Christian Löwens, Thorben Funke, and Alexandru Paul Condurache. Driving on Memory. arXiv preprint, 2026.

Ge Ya Luo, Zhi Hao Luo, Anthony Gosselin, Alexia Jolicoeur-Martineau, and Christopher Pal. Ctrl-V: Higher Fidelity Video Generation with Bounding-Box Controlled Object Motion. arXiv preprint, 2024.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. LingoQA: Visual Question Answering for Autonomous Driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 252–269, 2024.

Seokha Moon, Minseung Lee, Joon Seo, Jinkyu Kim, and Jungbeom Lee. Causality-Aware End-to-End Autonomous Driving via Ego-Centric Joint Scene Modeling. arXiv preprint, 2026.

Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617. Association for Computing Machinery, 2020.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2Drive: Towards Interpretable and Chain-based Reasoning for Autonomous Driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 292–308, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, 2020. Association for Computational Linguistics.

Amir Samadi, Amir Shirian, Konstantinos Koufos, Kurt Debattista, and Mehrdad Dianati. SAFE: Saliency-Aware Counterfactual Explanations for DNN-based Automated Driving Systems. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 5655–5662, 2023.

Hao Shao, Yuxuan Hu, Letian Wang, Steven L. Waslander, Yu Liu, and Hongsheng Li. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15120–15130, 2024.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with Graph Visual Question Answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 256–274, 2024.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4698–4726. PMLR, 06–09 Nov 2025.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, 31 (2):841–887, 2018.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22442–22452, 2025.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6585–6597, 2025.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

Tianyi Yan, Huan Zheng, Dubing Chen, et al. CausalDrive: Real-time Causal World Models for Autonomous Driving. arXiv preprint, 2026.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized Predictive Model for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14662–14672, 2024.

Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. OCTET: Object-aware Counterfactual Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15062–15071, 2023.

Yifan Zhan, Zhengqing Chen, Qingjie Wang, Zhuo He, Muyao Niu, Xiaoyang Guo, Wei Yin, Weiqiang Ren, Qian Zhang, and Yinqiang Zheng. Composing driving worlds through disentangled control for adversarial scenario generation. In *European Conference on Computer Vision*, pp. 467–487. Springer, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data. *Transactions on Machine Learning Research*, 2025.

Xingcheng Zhou, Hao Guo, Rui Song, Walter Zimmer, Mingyu Liu, André Schamschurko, Hu Cao, and Alois Knoll. CCTVBench: Contrastive Consistency Traffic VideoQA Benchmark for Multimodal LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21665–21684, San Diego, California, United States, 2026. Association for Computational Linguistics.

APPENDIX

A DriveWHEN Benchmark Construction **13**

A.1 From nuScenes Clips to Reviewed Pairs 13

A.2 Intervention Types 13

A.3 Action Space and Benchmark Profile 14

B Evaluation Protocol **15**

B.1 Models and Inference 15

B.2 Metrics 16

C Qualitative Analysis of Paired Decisions **16**

D DRIVETA Architecture and Driving-Video Adaptation **18**

E Supplementary Benchmark Analysis **18**

E.1 Response Across Required Relations and Interventions 18

E.2 Input Organization and Evidence Controls 19

F DriveWHEN Task-Specific Adaptation **19**

A DRIVEWHEN BENCHMARK CONSTRUCTION

This section explains how DriveWHEN converts a real driving clip into a controlled pair. The construction retains the scene, ego trajectory, and surrounding traffic while modifying one existing road user or roadside object. The paired videos therefore expose whether a model responds to a local change without replacing the broader scene context.

A.1 FROM NUSCENES CLIPS TO REVIEWED PAIRS

We scan the 850 annotated scenes in nuScenes v1.0 trainval (700 training and 150 validation scenes). From the beginning of each scene, we extract 33 consecutive front-camera frames at 12 Hz, corresponding to approximately 2.67 seconds between the first and last frame. The frames, 3D boxes, and ego poses provide the source clip and the geometry needed to select an editable actor.

For each intervention, we identify an existing actor that satisfies the intervention-specific spatial and safety conditions in Table 4. We then apply a predefined trajectory or appearance transformation while retaining the original ego motion. CompoSIA renders the edited sequence with a Wan2.1-T2V-1.3B backbone. DRIVEWHEN uses CompoSIA as its video renderer and contributes the counterfactual task design, reviewed benchmark, and paired evaluation protocol.

The scan yields 1,121 generation tasks across 725 scenes that support at least one edit. One task denotes one *(scene, intervention)* combination, so a scene can support more than one intervention. The pool contains 28 forward-away, 256 move-left, 568 move-right, and 269 appearance tasks. After reviewing edit visibility, temporal coherence, and decision relevance, we freeze 150 pairs from 138 source scenes for evaluation.

A.2 INTERVENTION TYPES

Let (x, y) denote the target center in the first-frame ego coordinate system, with x pointing right and y forward. For frame t, define $\phi_t = (\pi/2)t/T$ with $T = 33$. The target actor is selected before rendering; scenes without a valid actor are skipped.

Table 4: Counterfactual interventions. Every edit acts on one existing scene actor while the ego trajectory remains unchanged.

Intervention	Initial target region	Controlled edit	Context condition
Forward-away	$-1 < x < 1,\ 0 < y < 10$	$\Delta x_t = 0,\ \Delta y_t = 0.25t$	The acceleration corridor is free within 10 m.
Move left	$1 < x < 5,\ 0 < y < 20$	$\Delta x_t = -3.5 \sin \phi_t,\ \Delta y_t = -0.2t$	The ego lane ahead is clear for at least 15 m.
Move right	$x < -1,\ 0 < y < 20$	$\Delta x_t = +3.5 \sin \phi_t,\ \Delta y_t = -0.6t$	The ego lane ahead is clear for at least 15 m.
Appearance	$-1 < x < 1,\ y > 0$	Preserve the trajectory and replace the actor with a fixed appearance template	A leading object is present in the target region.

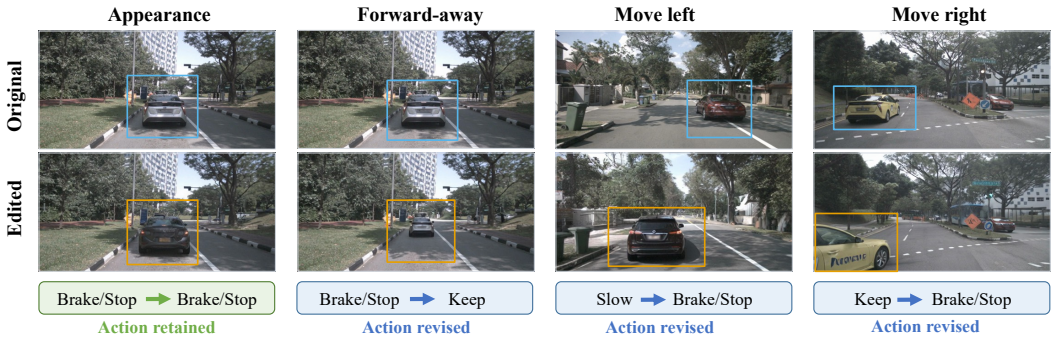

Figure 5: **Four counterfactual intervention types in DRIVEWHEN.** Each column presents corresponding frames from an original video and its edited counterpart. Appearance replacement modifies the target's visual identity; forward-away increases its longitudinal separation; move-left and move-right introduce lateral motion relative to the ego vehicle. The action pair beneath each example gives the reviewed video-level ego actions before and after editing. The appearance example retains the original action, whereas the three illustrated motion edits require an action revision. These examples demonstrate the intervention mechanisms rather than defining a fixed action outcome for each edit type. Colored boxes identify the edited actor for visualization and do not represent model attention.

Forward-away removes a potential longitudinal constraint. Move-left and move-right introduce lateral motion toward the ego path. Appearance changes the rendered identity while retaining motion. These meanings describe the intervention, not the action label: the surrounding context still determines whether the ego action should change.

The four intervention channels are illustrated in Figure 5. Each intervention modifies an existing scene actor while preserving the surrounding traffic context and recorded ego motion. The examples also highlight the distinction between a visual intervention and its behavioral consequence: the appropriate ego action may either remain unchanged or be revised, depending on how the edited interaction affects the complete scene.

A.3 ACTION SPACE AND BENCHMARK PROFILE

We manually review both videos in every pair and assign KEEP, SLOW, or BRAKE/STOP. SLOW includes cautious deceleration, low-speed continuation, and slow starting; BRAKE/STOP includes active braking and continuing to wait. The labels are frozen before the final comparison.

We order the actions by conservativeness,

$$r(\text{KEEP}) = 0, \qquad r(\text{SLOW}) = 1, \qquad r(\text{BRAKE/STOP}) = 2, \tag{12}$$

and derive the required relation from the two reviewed actions. This yields 74 unchanged, 65 more-conservative, and 11 less-conservative pairs.

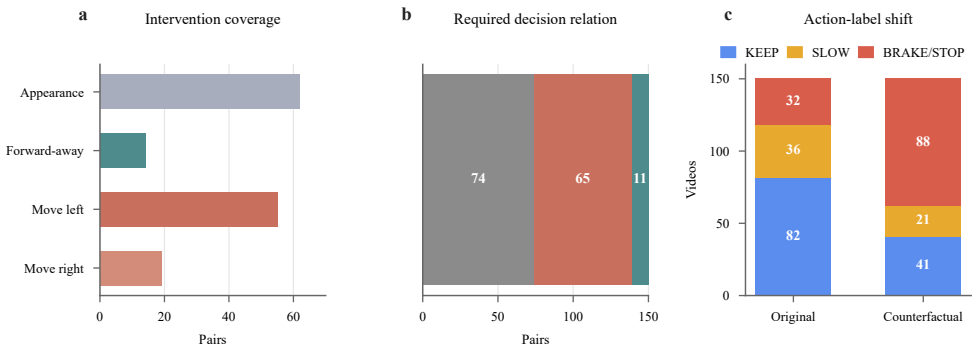

Figure 6: **DriveWHEN profile.** (a) Coverage of the four interventions in the reviewed 150-pair benchmark. (b) Distribution of the required action relation derived from the two frozen action labels. (c) Action-label composition before and after intervention. The shift toward BRAKE/STOP reflects the prevalence of safety-critical lateral interactions, while the unchanged and less-conservative subsets prevent the task from reducing to "always become more conservative."

Figure 6 summarizes the DriveWHEN more directly than a second count table. Appearance edits contribute 62 pairs, forward-away 14, move-left 55, and move-right 19. Original labels contain 82 KEEP, 36 SLOW, and 32 BRAKE/STOP videos; counterfactual labels contain 41, 21, and 88, respectively. DriveWHEN therefore combines action preservation, risk introduction, and constraint release.

Interaction diversity. Aggregate intervention counts do not fully characterize the difficulty of DRIVEWHEN. As shown in Figure 7, the reviewed videos contain diverse traffic participants, interaction geometries, and temporal locations of decision-relevant changes. Even within a nominal intervention such as move-left, the affected actor may only drift slightly toward the ego lane, enter rapidly from close range, merge gradually in the distance, cross at a sharp angle, or originate from an opposing lane. Consequently, recognizing the edit is not sufficient: a model must determine when and how the local change becomes relevant to the ego decision.

B EVALUATION PROTOCOL

Each model views the original and counterfactual videos in separate forward passes and returns one semantic action for each. It is not told which actor was edited, what intervention was used, or what the paired video contains.

B.1 MODELS AND INFERENCE

We evaluate DRIVETA, Qwen2.5-VL-7B, LLaVA-Video-7B, Qwen3-VL-8B, and LLaVA-OneVision-7B. The architecture and preceding driving-video adaptation of DRIVETA are described in Appendix D. The other four models use public pretrained weights and their native video or multi-image processors.

Every model receives 33 frames at 520×260 resolution under deterministic decoding with a maximum of 160 generated tokens. The original and counterfactual videos are processed in separate forward passes: the model receives neither the intervention type nor the paired alternative. All main benchmark results are computed over the complete set of 150 reviewed pairs.

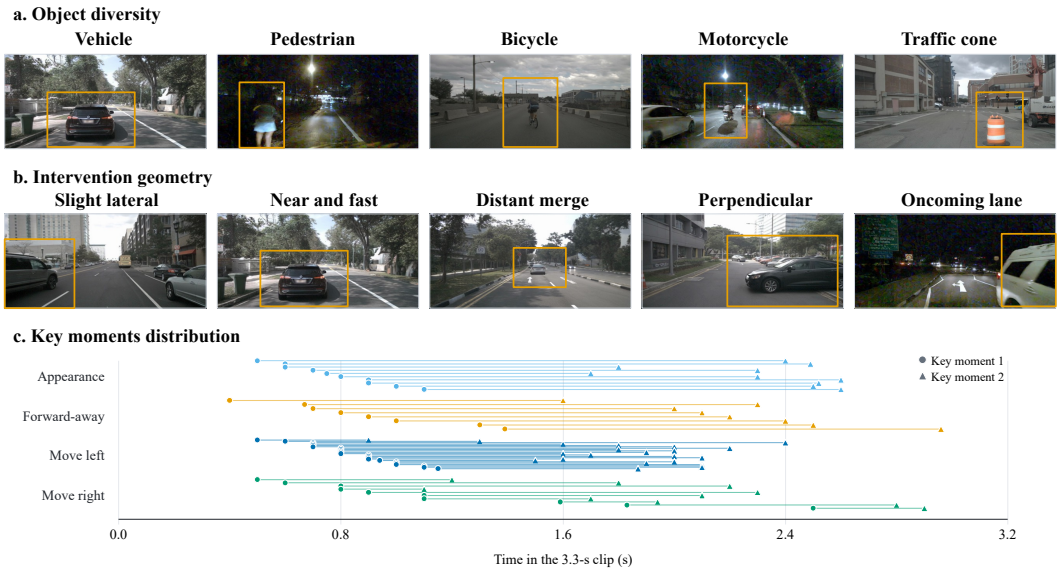

Figure 7: **Actor, geometric, and temporal diversity in DRIVEWHEN. (a)** Reviewed cases involve visually diverse traffic participants and roadside objects, including vehicles, pedestrians, bicycles, motorcycles, and traffic cones. **(b)** The interventions produce varied interaction geometries, from slight lateral motion and distant merging to close-range intrusion, perpendicular crossing, and motion from an opposing lane. **(c)** Two representative moments manually selected from each displayed case occur at different positions throughout the 33-frame video, showing that the informative change is not confined to a fixed temporal window. Circles and triangles denote the first and second selected moments, respectively; they are visualization points rather than annotations of event onset and termination. Boxes in panels (a) and (b) identify the relevant actor for visualization and do not represent model attention.

B.2 METRICS

For pair i, let y_i^o, y_i^c be the reviewed original and counterfactual actions and $\hat{y}_i^o, \hat{y}_i^c$ the predictions. Exact paired correctness is

$$c_i = \mathbf{1}[\hat{y}_i^o = y_i^o]\mathbf{1}[\hat{y}_i^c = y_i^c], \qquad \text{PairAcc} = \frac{1}{N}\sum_{i=1}^{N} c_i. \tag{13}$$

Define the reviewed and predicted relations as

$$g_i = \text{sgn}(r(y_i^c) - r(y_i^o)), \qquad \hat{g}_i = \text{sgn}(r(\hat{y}_i^c) - r(\hat{y}_i^o)). \tag{14}$$

For $\mathcal{C} = \{i : g_i \neq 0\}$,

$$\text{DIR-strict} = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}} \mathbf{1}[\hat{g}_i = g_i]. \tag{15}$$

Relation recall applies the same equality test separately to $g_i = 0, +1, -1$. Macro-F1 averages class F1 over all 300 videos, balanced accuracy averages the three class recalls, and Appearance PairAcc restricts paired correctness to the 62 appearance pairs. Prediction agreement on an appearance pair is reported separately as stability because stable error is not correct invariance.

C QUALITATIVE ANALYSIS OF PAIRED DECISIONS

Aggregate scores can conceal qualitatively different decision behaviors. Figure 8 complements the main-paper examples by isolating two distinctions that are central to the DRIVEWHEN evaluation: prediction agreement is not necessarily correct invariance, and a response in the correct direction does not necessarily reach the correct action.



Figure 8: **Additional paired decision patterns.** Each case shows aligned frames from the original and counterfactual videos together with the reviewed and predicted action pairs. **(a) Stable but wrong:** DRIVETA predicts `SLOW` for both appearance-paired videos, while both reviewed actions are `BRAKE_OR_STOP`. **(b) Correct direction, wrong action:** for a move-left intervention, Qwen3-VL changes from `KEEP` to `SLOW`, recognizing the need for a more conservative response but not reaching the reviewed `BRAKE_OR_STOP` action. **(c) Constraint release, partial response:** for a forward-away intervention, Qwen3-VL relaxes from `BRAKE_OR_STOP` to `SLOW`, whereas the released constraint supports `KEEP`. Models process the two complete videos independently; the displayed frames are selected only for visualization. Boxes mark the relevant actor and do not represent model attention.

Figure 8a shows why unchanged predictions cannot by themselves be interpreted as successful invariance. The model is stable across the pair, yet both action predictions are incorrect. This distinction motivates reporting Appearance PairAcc alongside prediction stability: the former requires the retained action to be correct, whereas the latter records agreement even when the shared prediction is wrong.

Figure 8b and Figure 8c expose a different ambiguity. Both predictions move in the reviewed direction, but they stop at an intermediate action. In the move-left case, the model detects increased risk but responds with `SLOW` where the edited interaction requires `BRAKE_OR_STOP`. In the forward-away case, it detects that the constraint has weakened but relaxes only to `SLOW` rather than `KEEP`. A

Table 5: DRIVETA driving-video adaptation and evaluation configuration.

Component	Configuration
Base architecture	DriveLM challenge LLaMA-Adapter V2 baseline
Driving adaptation data	696 DriveLM scenes
Training input	Eight front-view frames per video
Training stages	Initial stage: 3 epochs; subsequent front-view stages: 1 and 3 epochs
Frozen components	CLIP encoder and LLaMA language backbone
Updated components	Temporal embeddings and visual–language adaptation modules
Trainable parameters	Approximately 61.7M
Hardware	Two RTX 3090 GPUs initially; one RTX 3090 for subsequent stages (24 GB per GPU)
DRIVEWHEN evaluation	33 frames; eight learned temporal embeddings linearly interpolated to the evaluation length

relation-only score can recognize these partial responses, while PairAcc requires the correct action on both videos. Together, the cases demonstrate why DRIVEWHEN evaluates both the direction of a response and its final action endpoint.

D DRIVETA ARCHITECTURE AND DRIVING-VIDEO ADAPTATION

DRIVETA provides a video-capable driving baseline for the DRIVEWHEN evaluation. It extends the DriveLM challenge LLaMA-Adapter V2 baseline with learnable temporal embeddings over independently encoded front-view frames. CLIP extracts frame-level visual features, while the visual–language adaptation modules expose the resulting sequence to the LLaMA language decoder. The CLIP encoder and LLaMA backbone remain frozen throughout training. Only the temporal embeddings and adaptation modules are updated, together comprising approximately 61.7M trainable parameters.

The driving-video adaptation uses 696 DriveLM scenes and eight frames per training video. Training comprises an initial three-epoch stage followed by two front-view adaptation stages lasting one and three epochs, respectively. The initial stage was run on two NVIDIA RTX 3090 GPUs, while each subsequent stage used one RTX 3090 GPU. At evaluation time, the eight learned temporal embeddings are linearly interpolated to the 33-frame input length used by DRIVEWHEN. No DRIVEWHEN action labels or counterfactual pairs are used during this adaptation.

Table 5 summarizes the architecture and driving-video adaptation configuration.

E SUPPLEMENTARY BENCHMARK ANALYSIS

The main figure exposes the dominant trends; the following tables provide the complete values.

E.1 RESPONSE ACROSS REQUIRED RELATIONS AND INTERVENTIONS

Table 6 disaggregates model behavior according to whether the reviewed action should remain unchanged, become more conservative, or become less conservative. Relation recall is high on unchanged pairs but substantially lower on both adjustment directions. Exact PairAcc is lower still, showing that moving in the correct direction does not guarantee reaching the correct final action. This breakdown provides the complete numerical counterpart to Figure 3(a).

Table 7 reports exact paired correctness for each intervention. Performance is concentrated in appearance controls, whereas both lateral interventions remain particularly challenging. Move-left PairAcc reaches at most 3.03%, and move-right PairAcc reaches at most 3.51% across the five learned models. The result shows that responding to a newly introduced interaction is considerably more demanding than preserving a decision under an appearance change.

Table 6: Relation recall and exact PairAcc within each required relation (%). Values average the three semantic option orders.

	Unchanged		More cons.		Less cons.	
Model	Recall	PairAcc	Recall	PairAcc	Recall	PairAcc
DRIVETA	86.94	**36.04**	6.67	0.00	12.12	**6.06**
Qwen2.5-VL-7B	**88.29**	27.48	8.72	0.51	9.09	0.00
LLaVA-Video-7B	79.73	29.28	**13.33**	**1.54**	12.12	3.03
Qwen3-VL-8B	87.84	25.23	5.64	0.00	**15.15**	0.00
LLaVA-OneVision-7B	82.43	34.23	8.72	1.03	9.09	3.03
Always KEEP	100.00	41.89	0.00	0.00	0.00	0.00

Table 7: PairAcc by intervention (%), averaged over the three semantic option orders.

Model	Appearance	Forward-away	Move left	Move right
DriveTA	**42.47**	7.14	0.00	0.00
Qwen2.5	29.57	4.76	1.82	3.51
LLaVA-Video	32.26	7.14	**3.03**	1.75
Qwen3	28.49	7.14	0.00	0.00
OneVision	37.63	**9.52**	1.82	**3.51**

E.2 INPUT ORGANIZATION AND EVIDENCE CONTROLS

Table 8 compares three ways of presenting visual evidence: independently classifying eight sampled images and voting, jointly processing the same eight frames, and processing the complete 33-frame video. Joint presentation consistently improves DIR-strict over independent image voting, while increasing the sampled sequence from eight to 33 frames does not produce a uniform gain across models.

Table 9 further tests whether model outputs depend on visual and temporal evidence. Removing visual input eliminates the directional response of Qwen2.5-VL, while final-frame and temporal-position controls alter PairAcc and DIR-strict differently for DRIVETA. Together, these controls distinguish access to temporal evidence from the ability to translate that evidence into the required action response.

F DRIVEWHEN TASK-SPECIFIC ADAPTATION

The adaptation study tests whether the video adapter can learn the DRIVEWHEN action space from limited supervision and whether improved action recognition transfers to correct action adjustment across a pair. Table 10 reports the source-scene-disjoint training, validation, and test split.

Table 8: Input-organization ablation (%). Each cell reports PairAcc / DIR-strict. Images+vote and Joint8 use the same sampled timestamps.

Model	Images+vote	Joint8	Video33
DRIVETA	18.44 / 3.07	18.00 / 9.21	18.22 / 7.46
Qwen2.5-VL-7B	11.78 / 0.00	14.22 / 13.60	13.78 / 8.77
LLaVA-Video-7B	15.78 / 5.70	14.44 / 14.91	15.33 / 13.16
Qwen3-VL-8B	12.00 / 0.00	12.67 / 3.07	12.44 / 7.02
LLaVA-OneVision-7B	14.44 / 3.51	16.22 / 9.65	17.56 / 8.77

Table 9: Visual and temporal evidence controls (%). "App. stability" is prediction agreement on appearance pairs and can include stable error.

Model	Scope	Input	PairAcc	DIR	App. stab.	App. pair
Qwen2.5	three orders	Full video	13.78	8.77	88.17	29.57
Qwen2.5	three orders	No visual	11.33	0.00	100.00	27.42
Temporal	fixed prompt	Full video	21.33	5.26	87.10	46.77
Temporal	fixed prompt	Final image	19.33	11.84	80.65	41.94
Temporal	fixed prompt	Repeat final $\times 33$	20.00	7.89	82.26	45.16
Temporal	fixed prompt	Remove temp. pos.	18.67	9.21	88.71	41.94
Qwen2.5	fixed prompt	Full video	17.33	11.84	88.71	37.10
Qwen2.5	fixed prompt	Final image	12.67	3.95	85.48	29.03
Qwen2.5	fixed prompt	Repeat final $\times 33$	20.67	5.26	93.55	43.55

Table 10: Source-scene-disjoint adaptation split.

Split	Pairs	Scenes	Appearance	Move left	Move right	Forward-away
Train	60	55	24	22	8	6
Validation	15	14	6	5	3	1
Test	75	69	32	28	8	7

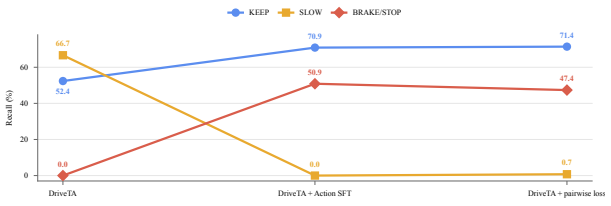

Figure 9: **Task adaptation redistributes action recall on the 75-pair test set.** Action SFT raises KEEP recall from 52.38% to 70.90% and BRAKE/STOP recall from 0% to 50.88%, while SLOW recall falls from 66.67% to 0%. Adding the pairwise loss leaves the same pattern largely unchanged. The adaptation gain is therefore a class redistribution rather than a uniform improvement across actions.

Let p^o and p^c denote the semantic distributions. Action SFT minimizes

$$\mathcal{L}_{\text{action}} = -\frac{1}{2}\left[\log p^o(y^o) + \log p^c(y^c)\right]. \tag{16}$$

For equal actions, the pairwise variant uses Jensen–Shannon divergence. With $m = (p^o + p^c)/2$,

$$\mathcal{L}_{\text{inv}} = \frac{1}{2}D_{\text{KL}}(p^o \| m) + \frac{1}{2}D_{\text{KL}}(p^c \| m). \tag{17}$$

For unequal actions, define $e(p) = p(\text{SLOW}) + 2p(\text{BRAKE/STOP})$. With gold direction $g \in \{-1, +1\}$,

$$\mathcal{L}_{\text{dir}} = \max\{0,\, 0.5 - g[e(p^c) - e(p^o)]\}. \tag{18}$$

The complete objective is

$$\mathcal{L}_{\text{pair}} = \mathcal{L}_{\text{action}} + 0.5\mathbf{1}[g = 0]\mathcal{L}_{\text{inv}} + 0.5\mathbf{1}[g \neq 0]\mathcal{L}_{\text{dir}}. \tag{19}$$

The frozen annotation selects the active relation term. Motion edits do not automatically trigger the directional loss.

Both variants update approximately 61.7M adaptation parameters for at most five epochs using AdamW with learning rate 10^{-5}, weight decay 0.01, gradient clipping at 1.0, mixed precision, and an effective batch of four pairs. Checkpoints are selected by validation action loss. The three runs vary data order and are not independent repetitions over all sources of training randomness.

Figure 9 explains why PairAcc can improve while DIR-strict declines. The original model uses SLOW frequently but never predicts BRAKE/STOP; SFT reverses this imbalance and becomes more likely to repeat a high-conservativeness action across a pair. The pairwise relation loss does not restore a balanced three-action response.