DO DRIVING WORLD-ACTION MODELS IMAGINE WHAT THEY PLAN? IMAGE-ACTION CONSISTENCY FOR EVALUATION AND LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving world-action models (WAMs) couple action prediction with visual imagination: the generated future should depict the motion the model plans. This correspondence requires direct evaluation. We introduce Image-Action Consistency (IAC), a framework for measuring this relationship and learning from the motion expressed in generated futures. Visual motion is recovered without access to the evaluated action, enabling comparison in a shared yaw-progress space. The Alignment Score (AS) measures within-pair motion agreement, the Response Consistency Score (RCS) evaluates paired yaw responses under controlled condition changes, and the Grounding Score (GS) measures geometric agreement with independently logged futures. Scores are reported together with measurement coverage. On held-out logged videos, visual progress estimates achieve 90.1% accuracy under the coarse tolerance rule on supported intervals at 88.2% coverage. Conditional agreement can hide failures to respond. Under the tested DriveWAM protocol, yaw directions agree in 92.5% of pairs with detectable responses in both outputs, yet only 49.5% of 745 valid pairs produce material action responses. Paired visual-motion targets extracted from generated futures further provide auxiliary supervision for action learning, improving PDMS by 1.54 points on DriveWAM and 1.00 point on WorldDrive over matched-budget action-only fine-tuning.

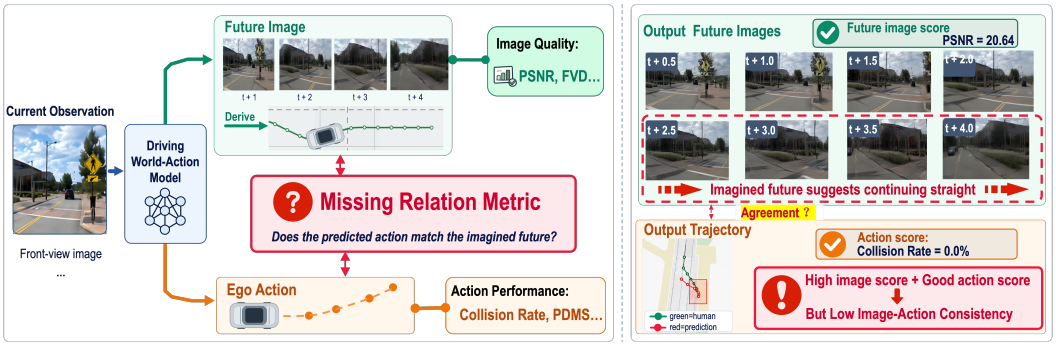
1 INTRODUCTION

Driving world-action models (WAMs) jointly generate a future video and a native action prediction within the same inference run (Wang et al., 2026; Gui et al., 2026; Shi et al., 2026; Zhang et al., 2025). This pairs a driving decision with a visual account of its predicted future. For that account to represent the model's plan, the motion visible in the video must agree with the native action over the same horizon. A video depicting a different maneuver cannot serve as a visual prediction of that plan. Video-quality and planning metrics assess appearance, distributional fidelity, action accuracy, and driving outcomes (Zhang et al., 2018; Unterthiner et al., 2018; Hu et al., 2023; Dauner et al., 2024). Establishing a coherent joint prediction requires evidence about the relationship between the outputs, as illustrated in Fig. 1(a).

Figure 1(b) illustrates a generated future whose apparent motion departs from the paired trajectory. Trajectory-based evaluations have begun to examine this relationship (Zhou et al., 2026; Liu et al., 2026). Static agreement leaves conditional behavior unresolved: two outputs can agree while both remain unchanged under a condition intervention. When that intervention is intended to alter motion, evaluation must check whether both outputs change in corresponding directions. Internal agreement also leaves open whether the generated future matches the recorded scene. These are distinct questions about agreement within a prediction, correspondence across predictions, and geometric agreement with an independently logged future.

Obtaining this evidence requires a common motion representation and a visual measurement that does not use the evaluated action to fit the agreement. Otherwise, the comparison partly reflects the

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.



(a) Separate image and action evaluation **(b) A case of Image-Action Inconsistency**

Figure 1: **Motivation for Image-Action consistency.** (a) Conventional visual-quality and planning metrics evaluate the two outputs separately, without directly checking their motion agreement. (b) A failure case from SUV (Yuan et al., 2026): the generated video initially suggests a slight right turn, but later frames (dashed box) drift toward an apparently straight-ahead continuation, inconsistent with the paired predicted trajectory (red). The generated scene also appears to depart from the observed road layout. The recorded human trajectory is shown in green. This case illustrates a motion disagreement that separate video and planning evaluations do not directly test.

evaluator's fitting procedure. Occlusion, scene dynamics, and generation artifacts also limit which motions can be recovered reliably. An unavailable visual measurement establishes neither agreement nor disagreement. Observation coverage is therefore part of the result.

We introduce Image-Action Consistency (IAC) to make these checks operational. A candidate-blind visual estimator and an action adapter map paired outputs into yaw and interval longitudinal progress. The Alignment Score (AS) measures within-pair motion agreement; the Response Consistency Score (RCS) measures paired yaw responses under controlled condition changes; and the Grounding Score (GS) compares generated and independently logged future geometry. Because recovered motion is expressed in action-compatible quantities, supported visual estimates can also serve as detached auxiliary targets for action learning. We retain the original action objective and freeze the video generator and visual estimator.

The evaluation reveals why this decomposition matters. Under the tested DriveWAM protocol, directions agree in 92.5% of pairs where both outputs have detectable yaw responses, but only 49.5% of the 745 valid pairs produce material action responses. High conditional direction agreement therefore coexists with an overall RCS of 38.1. Reporting both quantities exposes a response limitation that conditional agreement alone would conceal.

In summary, our contributions are:

- **Candidate-blind measurement of joint predictions.** We define a shared yaw-progress representation and a measurement protocol that reports agreement together with visual support.
- **A diagnostic decomposition of consistency.** AS, RCS, and GS distinguish within-pair agreement, paired yaw responses, and agreement with logged future geometry. The evaluation exposes response limitations hidden by conditional direction agreement.
- **Action learning from generated motion.** With video generation and visual estimation frozen, paired IAC targets improve PDMS by 1.54 points on DriveWAM and 1.00 point on WorldDrive over matched-budget action-only fine-tuning.

2 Related Work

2.1 World-Action Models for Autonomous Driving

Driving WAMs connect future scene prediction with action generation through different architectures (Wang et al., 2026). IAC evaluates models that export a future video and a native action from the same inference instance, including DriveWAM (Shi et al., 2026), Epona (Zhang et al., 2025), WorldDrive (Gui et al., 2026), and DriveVA (Liu et al., 2026). The evaluation operates on these paired outputs and applies across coupling architectures. Related approaches, including PWM (Zhao et al., 2026) and DriveLaW (Xia et al., 2026), couple future prediction with planning or condition video on action.

2.2 Video-Action Consistency

Visual scores such as FVD and planning scores such as PDMS evaluate the two outputs separately (Unterthiner et al., 2018; Dauner et al., 2024). Closer prior work already compares recovered visual motion with action information. ACT-Bench (Arai et al., 2024) evaluates action fidelity to supplied instructions and trajectory conditions. DrivingGen (Zhou et al., 2026) compares video-recovered trajectories with supplied ego trajectories using ADE and DTW. DriveVA (Liu et al., 2026) reconstructs video motion by visual odometry and compares it with the predicted trajectory after a two-dimensional similarity transform. IAC instead scores model-produced video-action pairs with candidate-blind visual calibration and explicit coverage, adds the Response Consistency Score (RCS) for paired responses to a controlled change in conditioning, and reports a Grounding Score (GS) of the generated future versus an independently logged future. The distinction is the measurement protocol and the paired response, not the use of reconstructed visual motion. Appendix C provides an extended discussion.

3 Methodology

IAC evaluates native video-action pairs through motion alignment (AS), paired responses to controlled condition changes (RCS), and geometric agreement with logged futures (GS). We call the model's native action the *candidate*. A visual estimator is *candidate-blind* when its motion estimates and calibration do not use that candidate or quantities computed from it. Fig. 2 summarizes the measurement and learning pipeline.

3.1 World-Action Model Interface

Given context x, control condition c, and sampling randomness ξ, a world-action model produces a future video and native action:

$$(V_c, a_c) = M_\theta(x, c, \xi), \qquad V_c = \{I_{c,t}\}_{t=1}^{T}. \tag{1}$$

The *native action* a_c is the action exported by the WAM in the same inference run as the generated video, before evaluator-side post-processing, reconstruction, or optimization. IAC also receives observed history H and evaluator metadata κ. We evaluate a four-second horizon using $K = 4$ one-second intervals. Reference futures are obtained independently from logged or realized states and are never derived from model predictions.

3.2 A Shared Motion Space for Video and Action

To compare heterogeneous video and action outputs, we map both modalities into a shared ego-motion space:

$$\mathbf{z}_c^s = (\psi_c^s, \mathbf{p}_c^s), \qquad \mathbf{p}_c^s = (p_{c,1}^s, \ldots, p_{c,K}^s), \qquad s \in \{F, A\}. \tag{2}$$

Here, F and A denote visual and action measurements. Both branches express interval displacement in the ego frame at the start of that interval and project it onto the local forward axis.

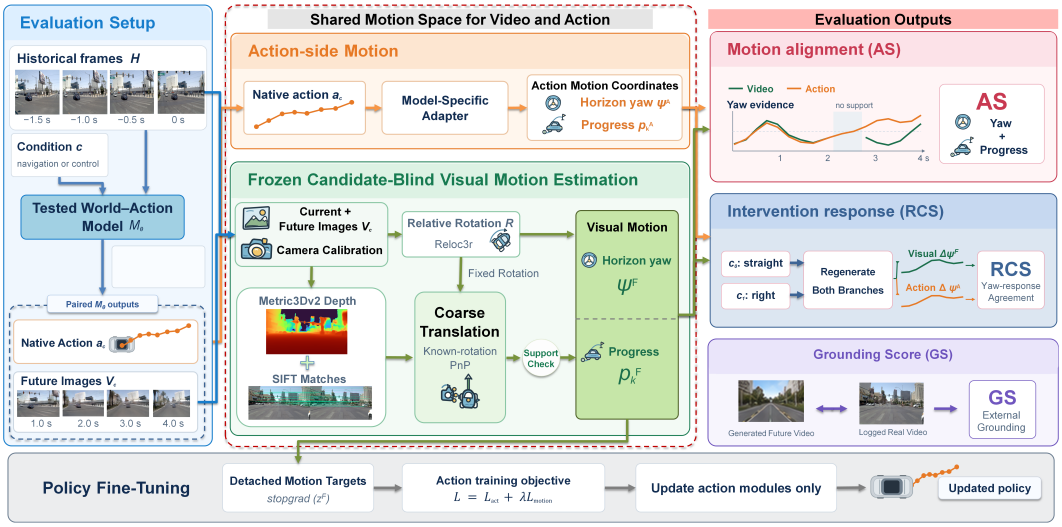


Figure 2: **Overview of IAC.** A tested world-action model (WAM) produces a generated future V_c and native action a_c from the same inference run. AS maps both outputs independently into a shared ego-motion space of yaw ψ and interval longitudinal progress p_k. The action branch uses a model-specific adapter, whereas the frozen candidate-blind visual branch estimates motion from the current/future images and evaluator metadata κ, without reading the candidate action. AS measures paired video-action agreement; RCS compares the same-source change in visual yaw against the change in native yaw under a controlled intervention; and GS compares frozen candidate-blind geometric descriptors between the generated and the logged future. Independent action execution is reported separately as external task validation. Supported visual motion estimates also supervise policy fine-tuning, with the video generator and visual estimator held fixed.

Table 1: **Notation.**

Symbol	Meaning
ψ^F, ψ^A	visual and action yaw
p_k^F, p_k^A	interval progress from visual and action branches
K	number of scored intervals
$\beta(\cdot)$	coarse progress binning function
$\Delta\psi^F, \Delta\psi^A$	visual and action responses
GS	grounding score

Visual motion estimation. The visual branch is candidate-blind: it uses only images, timestamps, and camera calibration, and its measurements are fixed before the native actions are accessed. For yaw, Reloc3r (Dong et al., 2025) estimates relative rotations between adjacent frames, which are transformed to the ego frame and accumulated over valid intervals. For progress, we combine Metric3Dv2 depth (Hu et al., 2024) with SIFT correspondences (Lowe, 2004), and recover camera translation with the estimated rotation held fixed. The recovered displacement is transformed to the ego frame at the interval start; its longitudinal component defines $p_{c,k}^F$.

Action motion. A model-specific adapter converts the native action into ego poses over the same K intervals. The action motion comprises final yaw and interval longitudinal progress:

$$\psi_c^A = \psi_{c,K}, \qquad p_{c,k}^A = e_x^\top \Delta r_{c,k}^A, \tag{3}$$

where $\Delta r_{c,k}^A$ is the native trajectory displacement over interval k, expressed in the ego frame at that interval's start, and e_x is the local forward unit vector.

Measurement support. A visual progress estimate is retained only when the geometric evidence is sufficient: at least 40 correspondences, an inlier fraction of at least 0.10, and a median reprojection error of at most 3 pixels. In the reported AS and RCS evaluations, visual yaw is accumulated over all four one-second intervals. Visual and action yaw use the same four-second horizon, including both conditions of each RCS pair. Unsupported measurements remain undefined rather than being assigned zero values. For coarse progress comparison, we quantize nonnegative displacement using thresholds of 0.5, 3, 6, and 9 m, corresponding to `stop`, `very short`, `short`, `medium`, and `long`. We bin $\max(p^F, 0)$ for visual progress and p^A for action progress, and regard two estimates as consistent when they fall in the same or adjacent bins. The bin boundaries and agreement tolerance are calibrated on logged futures and fixed before evaluation.

3.3 Motion Consistency Evaluation

Alignment Score (AS). AS measures coarse motion agreement between a generated future and its native action. Progress is compared over supported intervals, and yaw direction is compared on actions whose turn magnitude reaches the fixed deadband $\tau_\psi = 0.0128$ rad. We define

$$Q_{\tau_\psi}(y) = \begin{cases} 0, & |y| < \tau_\psi, \\ \operatorname{sgn}(y), & |y| \geq \tau_\psi. \end{cases} \tag{4}$$

Only examples satisfying $Q_{\tau_\psi}(\psi_i^A) \neq 0$ are included in yaw agreement evaluation:

$$\begin{aligned} A^p_{i,k} &= \mathbf{1}\left[|p^F_{i,k} - p^A_{i,k}| \leq \max(3\,\mathrm{m}, 0.5\,p^A_{i,k}) \ \vee \ B_\beta\big(\max(p^F_{i,k}, 0),\, p^A_{i,k}\big) = 1\right], \\ A^\psi_i &= \mathbf{1}\left[\operatorname{sgn}(\psi^F_i) = \operatorname{sgn}(\psi^A_i)\right], \quad \text{defined when } Q_{\tau_\psi}(\psi^A_i) \neq 0. \end{aligned} \tag{5}$$

Here, $B_\beta(u, v) = \mathbf{1}[|\beta(u) - \beta(v)| \leq 1]$, where $\beta(p)$ assigns $\max(p, 0)$ to the progress bins in Sec. 3.2. Clipping applies only to bin lookup; the metric-tolerance test retains the signed residual. Yaw agreement evaluates turn direction, not magnitude.

We report conditional agreement and measurement coverage separately. Let $\mathcal{I}_p$ contain supported progress intervals and $\mathcal{I}_\psi$ contain eligible turning examples with visual yaw support.

$$\begin{aligned} P_{\mathrm{cond}} &= \frac{\sum_{(i,k)\in\mathcal{I}_p} A^p_{i,k}}{|\mathcal{I}_p|}, & Y_{\mathrm{cond}} &= \frac{\sum_{i\in\mathcal{I}_\psi} A^\psi_i}{|\mathcal{I}_\psi|}, \\ AS_{\mathrm{cond}} &= 100\sqrt{P_{\mathrm{cond}} Y_{\mathrm{cond}}}, & AS &= AS_{\mathrm{cond}}\sqrt{C_p C_\psi}, \\ C_p &= \frac{|\mathcal{I}_p|}{KN}, & C_\psi &= \frac{|\mathcal{I}_\psi|}{N_\psi}. \end{aligned} \tag{6}$$

Here, N counts evaluated video-action branch rows, and N_ψ counts those with non-zero turn labels after applying Q_{τ_ψ}. Multiple branches from one source share a source identity. $\mathrm{AS}_{\mathrm{cond}}$ summarizes agreement on supported measurements; AS weights it by the geometric mean of progress and yaw coverage. Ratios with zero denominators, and scores that depend on them, are undefined. AS combines marginal progress and yaw agreement, not the fraction of pairs passing both channels. Its yaw component compares signs on action-defined turns without a visual magnitude threshold. Low AS can reflect disagreement or limited observability; low coverage alone does not establish a motion error.

Response Consistency Score (RCS). RCS measures whether video and action outputs respond consistently to the same intervention. For each condition pair (c_0, c_1), we keep the observed history and sampling randomness fixed and compare the resulting motion changes across the two outputs. The visual and action responses are

$$\Delta\psi^F_i = \psi^F_{i,c_1} - \psi^F_{i,c_0}, \qquad \Delta\psi^A_i = \psi^A_{i,c_1} - \psi^A_{i,c_0}. \tag{7}$$

We quantize each response as $D_\tau(v) = \operatorname{sgn}(v)\mathbf{1}[|v| \geq \tau]$. The action threshold is $\tau_A = 0.0128$ rad and the visual threshold is $\tau_F \approx 0.01332$ rad; both are fixed before evaluation. For N_{pair} contract-valid intervention pairs, excluding invalid inputs,

$$\mathrm{RCS} = \frac{100}{N_{\mathrm{pair}}} \sum_i \mathbf{1}\left[D_{\tau_A}(\Delta\psi^A_i) \neq 0 \ \wedge \ D_{\tau_F}(\Delta\psi^F_i) = D_{\tau_A}(\Delta\psi^A_i)\right]. \tag{8}$$

A pair receives credit only for measurable, material yaw responses in the same direction. Valid pairs with no material response or unavailable visual measurements remain in the denominator and contribute zero to the score. An unavailable response is not imputed as zero motion. RCS measures paired response correspondence under intervention, not causal influence between the output branches. It does not test whether either response follows the intended semantics of the condition.

3.4 Reference-Based Geometric Grounding

Grounding Score (GS). GS measures geometric agreement between a generated future and an independently logged future from the same source scene, providing a reference-based check separate from internal video-action agreement. A frozen candidate-blind reader combines Metric3Dv2 depth (Hu et al., 2024), ORB correspondences (Rublee et al., 2011), and EPNP/RANSAC pose estimation. Poses that pass held-out geometric checks are used to reproject depth into the next frame.

The reprojected geometry provides relative depth-change and local-expansion descriptors. Let g_{ikj} and r_{ikj} denote generated and logged descriptors for source i, interval k, and tile descriptor j. For scale and expansion classes $q \in \{s, e\}$, the class similarity is

$$s_{i,q} = \exp\left[- \operatorname{median}_{(k,j)\in\mathcal{E}_{i,q}} \frac{|g_{ikj} - r_{ikj}|}{\sigma_j} \right], \tag{9}$$

where $\mathcal{E}_{i,q}$ pools common supported tile-interval entries across retained intervals. Both classes are mandatory. A supported source requires at least two retained intervals and sufficient support in each class (Appendix A). We combine the classes geometrically:

$$\begin{aligned} \mathrm{GS}_{\mathrm{cond}} &= \operatorname{median}_{i\in\mathcal{I}_{\mathrm{GS}}} \sqrt{s_{i,s} s_{i,e}}, \\ \mathrm{GS}_{\mathrm{obs}} &= |\mathcal{I}_{\mathrm{GS}}| / N_{\mathrm{GS}}, \end{aligned} \tag{10}$$

with supported-source set $\mathcal{I}_{\mathrm{GS}}$ and declared count N_{GS}. Descriptor scales σ_j are calibrated on logged real futures, fixed before confirmation, and shared across evaluated models. Sources failing either class support requirement remain unavailable. Matched and shuffled references test whether the descriptors preserve source-specific geometric information. A different valid continuation can receive a lower score against the logged future; GS does not certify all aspects of physical validity.

Reporting. We report scores with measurement coverage, distinguishing unavailable evidence from measured disagreement or nonresponse. Resampling units are sources for AS and GS and intervention source groups for RCS (Appendix A). All evaluation parameters are fixed across models, except for coordinate-orientation signs.

3.5 Consistency-Guided Policy Fine-Tuning

We use supported visual motion estimates as auxiliary supervision for policy learning. The frozen visual estimator extracts motion targets $\mathbf{z}_c^F$ from generated futures, while the action branch produces $\mathbf{z}_c^A$. The original action objective is retained and augmented with:

$$\mathcal{L} = \mathcal{L}_{\mathrm{act}} + \lambda^\star \mathcal{L}_{\mathrm{mot}}. \tag{11}$$

Let ρ denote the unit Huber loss, sg the stop-gradient operator, and $\bar{m}^p = \mathbf{1}[\sum_k m_k^p > 0]$. The auxiliary loss combines supported yaw supervision with progress penalties outside the visual tolerance:

$$\begin{aligned} \ell_\psi &= \rho\left(\frac{\operatorname{wrap}(\psi^A - \operatorname{sg}(\psi^F))}{\tau_\psi} \right), \qquad \tau_{p,k} = \max\big(3\,\mathrm{m},\, 0.5|\operatorname{sg}(p_k^F)|\big), \\ e_{p,k} &= \frac{[|p_k^A - \operatorname{sg}(p_k^F)| - \tau_{p,k}]_+}{\tau_{p,k}}, \qquad \ell_p = \frac{\sum_k m_k^p \rho(e_{p,k})}{\sum_k m_k^p + \epsilon}, \\ \mathcal{L}_{\mathrm{mot}} &= \frac{m^\psi \ell_\psi + \bar{m}^p \ell_p}{m^\psi + \bar{m}^p + \epsilon}. \end{aligned} \tag{12}$$

Here, $[x]_+ = \max(x, 0)$ and τ_ψ is the fixed yaw deadband. The masks m^ψ and m_k^p follow the AS support criteria; the progress term penalizes deviations only outside the visual tolerance. Only action-side parameters are updated; the video generator and visual motion estimator remain fixed. We evaluate the resulting policies with PDMS (Dauner et al., 2024) to measure downstream driving performance. The auxiliary weight $\lambda^\star$ is selected on source-disjoint development samples and fixed for confirmation (Appendix A.7).

4 EXPERIMENTS

We evaluate IAC from three perspectives: the reliability of its candidate-blind visual motion measurements, the validity of the resulting consistency scores for world-action models, and the utility of consistency-guided supervision for downstream policy learning.

4.1 EXPERIMENTAL SETUP

Data and models. We use NAVSIM (Dauner et al., 2024) for motion validation and open-source WAM evaluation. The benchmark provides logged videos and synchronized vehicle poses as independent motion references, covering 1,000 source windows from 675 scenes and 64 collection logs. Following a fixed frame protocol, each sample contains four historical frames and four future frames at $t = 1, 2, 3, 4$ s. We evaluate publicly available WAMs including DriveWAM (Shi et al., 2026), Epona (Zhang et al., 2025), WorldDrive (Gui et al., 2026), and DriveVA (Liu et al., 2026). AS is evaluated on the same 30 NAVSIM source identities across all four models, whereas RCS follows each model's supported intervention protocol.

4.2 ACCURACY OF VISUAL MOTION MEASUREMENTS

Motion estimation accuracy. We first validate the visual motion evidence used by IAC against synchronized vehicle poses on held-out NAVSIM videos. The yaw estimator achieves 91.7% accuracy on 120 clips and 97.5% direction accuracy on turning clips, as reported in Table 2. Across 380 intervals from 95 held-out sequences, progress estimation achieves a mean absolute error of 1.686 m and 90.1% tolerance accuracy at 88.2% coverage. The exact-bin accuracy is 58.5%, indicating that the recovered motion supports coarse progress comparison rather than precise trajectory recovery.

Measurement selection. At matched coverage (88.2%), geometric support selection yields progress MAE of 1.686 m, compared with 1.745 m for random selection over 10,000 subsets. Tolerance accuracy is 90.1% and 90.2%, respectively. This control shows a modest error reduction at fixed coverage without increasing acceptance.

4.3 VALIDITY OF MOTION-BASED IAC SCORES

We further validate whether IAC scores preserve the intended video-action relationships. AS is evaluated against independent reference motion, while RCS is tested through controlled intervention pairs and negative controls.

AS validation. We compare conditional agreement $\mathrm{AS}_{\mathrm{cond}}$ from recovered visual motion and logged vehicle motion, keeping actions and scoring rules unchanged.

Across 95 logged-realized sequences, $\mathrm{AS}_{\mathrm{cond}}$ from recovered motion reaches 94.9 on supported measurements, compared with 100 using reference motion. The agreement decisions are preserved for all applicable yaw clips and 90.1 percent of supported progress intervals.

Controlled mismatch validation. On DriveWAM, we test whether AS can identify inconsistent video-action pairs by introducing controlled deviations to the action side while keeping the visual future fixed. AS decreases from 64.0 on matched pairs to 59.2 under large yaw deviations, 54.8 under large progress deviations, and 41.6 under random pairing. These controls show sensitivity to the tested motion and pairing perturbations. The controlled-mismatch results are summarized in Appendix Table 10.

Negative controls. Swapping visual branches yields 7.6% response-direction agreement as a control for response pairing. Repeated-frame inputs produce no false turns in 32 trials. For GS, we compare matched and shuffled references (Table 3) and use blurred and repeated futures to probe geometric support and static-scene shortcuts.

Table 2: Validation of candidate-blind visual motion estimation against logged vehicle motion. Yaw is evaluated per clip and progress over 1s intervals.

Measurement	Result
Yaw acc. (L/S/R)	91.7/97.5%
Progress coverage	88.2%
Progress acc. (coarse / exact)	90.1/58.5%
Abs. error (mean / median, m)	1.686/0.865
Spearman ρ	0.665

Table 3: **Reference-based geometric agreement.** GS denotes $\mathrm{GS}_{\mathrm{cond}}$ and Cov. denotes $100\,\mathrm{GS}_{\mathrm{obs}}$. $q_{.95}$ is the shuffled-reference 95th percentile; Margin $= \mathrm{GS} - q_{.95}$.

Model	GS	Cov.	$q_{.95}$	Margin
DriveWAM	0.3728	90%	0.2766	+.0962
WorldDrive	0.5587	100%	0.4069	+.1518
Epona	0.6184	100%	0.4799	+.1385
DriveVA	0.4846	70%	0.4509	+.0337

Table 4: **Alignment on native video-action pairs.** Yaw and Prog. are conditional agreement percentages; Cond. is $\mathrm{AS}_{\mathrm{cond}}$. Obs. $= 100\sqrt{C_p C_\psi}$ and AS = Cond. $\times$ Obs./100. N counts evaluated video-action pairs from 30 sources.

Model	N	Yaw	Prog.	Cond.	Obs.	AS
DriveWAM	60	94.6	65.4	78.7	81.4	64.0
Epona	30	100.0	83.3	91.3	97.5	89.0
DriveVA	30	88.9	63.2	75.0	85.1	63.8
WorldDrive	30	100.0	77.4	88.0	97.9	86.1

Table 5: **Paired yaw responses.** N counts contract-valid pairs. Act. is the percentage with material action responses; Vis. is the percentage of those with detected visual responses; Dir. is direction agreement among pairs with both responses. RCS uses all N pairs.

Model	N	Act.	Vis.	Dir.	RCS	95% CI
DriveWAM	745	49.5	83.2	92.5	38.1	$[34.8, 41.6]$
Epona	40	100.0	95.0	92.1	87.5	$[77.5, 97.5]$
DriveVA	50	74.0	43.2	62.5	20.0	$[10.0, 32.0]$
WorldDrive	50	60.0	66.7	100.0	40.0	–

4.4 Evaluation of World-Action Models

Within-pair agreement. The four models share 30 source identities. DriveWAM contributes 60 video-action pairs and the other models 30 each. AS separates conditional agreement from coverage. DriveWAM scores 64.0, with conditional yaw and progress agreement of 94.6% and 65.4%, respectively.

Response availability and direction. Under the tested DriveWAM intervention protocol, only 369 of 745 pairs produce material action responses. When both outputs respond, their directions agree in 92.5% of cases, yielding an overall RCS of 38.1 (Table 5). The main reduction in score occurs before direction agreement is assessed. For DriveVA, 37 of 50 pairs produce action responses, but only 16 of those also produce detectable visual responses. These counts reflect both response occurrence and visual measurement availability. Figures 3 and 4 illustrate the alignment and response measurements.

Model-specific intervention pools preclude a paired cross-model ranking.

Reference geometry. Matched-reference GS exceeds the shuffled-reference 95th percentile for all four models in Table 3, with margins from 0.0337 to 0.1518 and coverage from 70% to 100%. These descriptors retain source-specific geometric information. DriveVA's 70% coverage falls below the 90% eligibility threshold, so its GS is a conditional diagnostic. The model-specific pools preclude a paired cross-model ranking.

4.5 Consistency-Guided Policy Fine-Tuning

We further test whether IAC measurements provide useful supervision beyond evaluation. We use the candidate-blind visual motion targets extracted by the frozen IAC estimator as an auxiliary signal and evaluate the resulting policies with an independent driving metric. We fine-tune DriveWAM and WorldDrive under the same training budget. Each model is evaluated with action-only supervision, GT motion targets, shuffled visual targets, and paired IAC targets. All settings share the

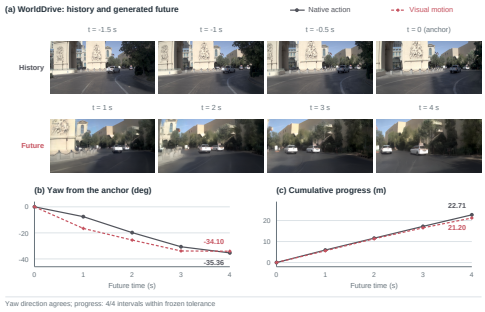


Figure 3: **AS example.** Recovered visual motion and the native action agree in turn direction and interval progress.

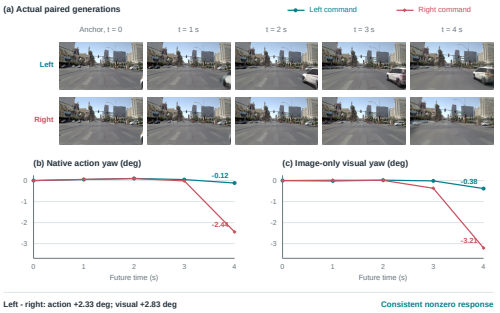


Figure 4: **RCS example.** A condition change produces action and visual yaw changes in the same direction.

Table 6: **Action learning with IAC-derived motion targets.** Fine-tuned PDMS values are means over three training seeds under matched budgets; PDMS is evaluated independently. All Δ values are relative to matched-budget action-only fine-tuning.

Training strategy	**DriveWAM**		**WorldDrive**	
	PDMS ↑	Δ	PDMS ↑	Δ
Original checkpoint	92.0827	–	92.2326	–
Action-only fine-tuning	92.0597	0	92.5086	0
GT motion targets	92.5078	0.4481	92.5081	-0.0005
Shuffled visual targets	92.5156	0.4559	92.5067	-0.0019
Paired IAC targets	**93.5953**	**1.5356**	**93.5081**	**0.9995**

same initialization and optimization schedule, while the video generator and motion estimator remain frozen. Table 6 shows that paired IAC targets improve PDMS on both models. Relative to action-only fine-tuning, GT motion targets raise mean PDMS by 0.4481 points on DriveWAM and leave it essentially unchanged on WorldDrive (-0.0005 points). Shuffled visual targets do not reproduce the full gain from paired supervision. For DriveWAM, paired supervision raises PDMS to 93.5953, and WorldDrive shows the same trend with a PDMS of 93.5081. These results support the utility of paired IAC targets as auxiliary supervision for action learning: under matched budgets, the PDMS gains are substantially larger than those from GT motion targets or shuffled visual targets. This pattern indicates that the benefit depends on pairing the action objective with the model's own generated future, rather than on motion-label accuracy alone. PDMS gains alone do not establish improved image-action consistency.

5 Conclusion

We presented Image-Action Consistency (IAC) for evaluating and learning from the relationship between imagined futures and native actions in driving WAMs. Through candidate-blind motion recovery and three complementary metrics, IAC quantifies motion alignment, response consistency under controlled interventions, and geometric agreement with logged futures. Our experiments show that strong agreement among responding outputs can mask limited action responsiveness. They also demonstrate that recovered visual motion provides useful supervision for planning, improving PDMS on both DriveWAM and WorldDrive under matched training budgets. The PDMS gains do not, by themselves, establish improved image-action consistency. Together, these findings establish the value of evaluating joint predictions through their motion relationships and using generated futures to support action learning.

Limitations and future directions are discussed in Appendix B.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting, language editing, and revision of scientific wording, as well as to help cross-check consistency among reported numerical quantities and their surrounding prose. The tools were not used to generate experimental results, datasets, or scientific conclusions. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the content of the final manuscript.

REFERENCES

Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. Act-bench: Towards action controllable world models for autonomous driving, 2024. URL https://arxiv.org/abs/2412.05337.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization, 2025. URL https://arxiv.org/abs/2412.08376.

Xingtai Gui, Meijie Zhang, Tianyi Yan, Wencheng Han, Jiahao Gong, Feiyang Tan, Cheng-zhong Xu, and Jianbing Shen. Bridging scene generation and planning: Driving with world model via unifying vision and motion representation. *arXiv preprint arXiv:2603.14948*, 2026.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, December 2024. ISSN 1939-3539. doi: 10.1109/tpami.2024.3444912. URL http://dx.doi.org/10.1109/TPAMI.2024.3444912.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024.

Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Michael Ying Yang, Francesco Nex, and Hao Cheng. Driveva: Video action models are zero-shot drivers. *arXiv preprint arXiv:2604.04198*, 2026.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. Orb: An efficient alternative to sift or surf. *2011 International Conference on Computer Vision*, pp. 2564–2571, 2011. URL https://api.semanticscholar.org/CorpusID:206769866.

Chen Shi, Jinrui Xu, Shaoshuai Shi, Kehua Sheng, Bo Zhang, and Li Jiang. Drivewam: Video generative priors enable scalable world-action modeling for autonomous driving. *arXiv preprint arXiv:2605.28544*, 2026.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, et al. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*, 2026.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Drivelaw:unifying planning and video generation in a latent driving world, 2026. URL https://arxiv.org/abs/2512.23421.

Yibo Yuan, Jiacheng Fu, Jiangtong Zhu, Yi Li, Jianhua Han, Meng Tian, Zhuohan Liu, Zhiwei Xiong, Hang Xu, Jianwu Fang, and Jianru Xue. Suv: Future scene understanding as video generation for end-to-end driving, 2026. URL https://arxiv.org/abs/2608.03084.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27220–27230. IEEE, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. From forecasting to planning: Policy world model for collaborative state-action prediction. *Advances in Neural Information Processing Systems*, 38:134585–134611, 2026.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L. Waslander. Driv-inggen: A comprehensive benchmark for generative video world models in autonomous driving, 2026. URL https://arxiv.org/abs/2601.01528.

A Evaluation Protocol and Additional Evidence

A.1 Sampling and Output Lineage

The frozen NAVSIM benchmark contains 1,000 non-overlapping windows from 675 scenes and 64 collection logs. Each AS input consists of observed frames at $t = -1.5, -1.0, -0.5, 0$ s, generated frames at $t = 1, 2, 3, 4$ s, and four native trajectory states aligned to the latter timestamps. The first scored progress interval runs from the last observed frame to the first generated frame. The common-source AS comparison uses the same 30 model-output-blind NAVSIM source identities for DriveWAM, Epona, DriveVA, and WorldDrive. Log-level independence is not claimed for this pool. Observation coverage is reported separately from the source count.

The model adapter preserves each WAM's native action provenance and aligns its coordinates and timestamps to the common evaluation convention. Model-consumed history images and resized camera intrinsics are audited together. Epona's native 5 Hz autoregression supplies 20 steps from which the four evaluation times are selected. WorldDrive supplies native 2 Hz outputs over eight future steps. DriveVA uses its official three-step inference and the calibrated 832×480 model input. Generated frames, native actions, camera intrinsics, and the model-consumed history must refer to the same run. A predicted action is never used as the logged reference. The common-pool source identity and per-model lineage checks are recorded in the public AS audit. Formal AS also requires at least 20% readable history-future boundary geometry and no more than 10% large boundary yaw discontinuities above 30°; failed lineage remains diagnostic.

Formal yaw RCS uses one left-right pair per source. Within each `counterfactual_group_id`, the branch roles must be exactly {left, right}. The two branches share the history fingerprint, nuisance seed, WAM identity, and an allowed intervention type (a native navigation command or action-trajectory perturbation), and both futures are regenerated. The declared branch roles and regenerated-video lineage must agree, and both future paths must exist and differ. Validation-only image degradations are excluded. Pairs failing this contract are invalid inputs; contract-valid pairs with no material action response or no visual response receive no score credit and remain in the denominator. The four models use different supported intervention pools, with $745/40/50/50$ valid pairs for DriveWAM/Epona/DriveVA/WorldDrive. Consequently, these RCS results characterize responses within each model and are not a paired cross-model comparison. The formal RCS score is yaw-only: longitudinal response and speed changes are retained as diagnostics and do not enter the published RCS number. An intervention must be applied before generation to both output branches; editing a generated video or injecting an action after video generation is invalid. RCS therefore measures paired response correspondence under the declared intervention, not causal mediation from video to action.

The separate RCS-progress protocol uses four native longitudinal conditions: `stop`, `slow`, `normal`, and `fast`, which multiply the original trajectory's longitudinal displacement by 0, 0.5, 1.0, and 1.5, respectively, while preserving lateral motion and yaw. Within each group, scoring selects the branches with the lowest and highest realized action progress. A group is eligible when their endpoint difference is at least 0.5 m, and its visual response is correct when the high-action branch has higher visual progress. Formal reporting requires at most one group per independent source/log/scene, at least 30 independent eligible sources, coverage of at least 0.90, endpoint accuracy of at least 0.75, and a Wilson 95% confidence lower bound above 0.5. This protocol does not enter the yaw RCS values in Table 5.

Native conditioning interfaces. The released DriveWAM interface conditions both output streams on the route prompt, command identifier, and ego state (Shi et al., 2026). WorldDrive feeds the driving command, velocity, and acceleration to its planner; the selected trajectory then conditions scene generation (Gui et al., 2026). Our WorldDrive pairs use the official FAR visualization and its final predicted trajectory. For Epona, the native left/right control uses lateral offsets of $-0.25/+0.25$ and yaw offsets of $+2^\circ/-2^\circ$, respectively; RCS scores the realized output difference rather than these nominal offsets.

GS uses model-specific pools of 30 sources: 15 turn and 15 non-turn, with each source from a distinct scene/log. The turn-stratification threshold is 0.12 rad. Endpoint matching permits at most 0.1 rad yaw error, 8 m longitudinal error, and 4 m lateral error between native action and logged trajec-

Table 7: Evaluation units for the main-text results.

Metric	Main-text result	Evaluation units
AS	Table 4	30 shared source identities; 60 video-action branches for DriveWAM and 30 for each other model; conditional agreement and coverage are reported.
RCS	Table 5	745/40/50/50 valid pairs for DriveWAM/Epona/DriveVA/WorldDrive, respectively.
GS	Table 3	30 condition-matched sources per model; supported-source counts and control outcomes are given in Table 11.

tory. Sources are fixed before generated pixels or visual-reader outputs are inspected. GS reports geometric agreement and coverage on each pool; its medians do not establish a paired ranking.

Split and leakage boundary. The visual layer is calibrated on real logged NAVSIM motion only. The source-disjoint requirement is part of the frozen AS configuration, and the common 30-source AS pool is fixed before model outputs or visual reader results are inspected. The 95-row visual holdout is used to report measurement validity, a different evaluation purpose from the 30-source WAM comparison. GS uses model-specific condition-matched pools. The predeclared GS condition matching described above does not fit the visual reader to candidate actions. Visual-reader calibration does not use generated pixels or candidate actions, and source selection based on generated pixels or visual-reader scores is diagnostic only.

What is and is not comparable. The AS table compares the same source identities with disclosed branch counts. RCS is computed on model-specific intervention pools, and GS is computed on model-specific condition-matched pools; their rows therefore share a protocol and score semantics but not identical statistical units. RCS and GS are reported side by side for diagnosis, not combined into a single three-metric rank. A GS number is eligible as a single-number result only when its status is `Valid`; coverage-limited rows retain their conditional score and status but do not receive an imputed rank.

A.2 Evaluation Units and Implementation Details

Table 7 links each metric to its main-text result and evaluation unit. The declared sample count is distinct from the number of supported measurements contributing a conditional score.

The AS pool contains 30 source identities per model. In Table 4, N counts video-action branch rows: 60 for DriveWAM and 30 for each other model. Four one-second intervals per row give 240 and 120 declared progress intervals, respectively. Progress coverage uses these interval denominators. Yaw coverage uses eligible turning rows; all four yaw intervals are available, so $C_\psi = 1$ on eligible turns. All AS ratios use counts pooled over this single evaluation. Table 8 gives the raw counts. Branches sharing a source are not independent source samples. Where reported, RCS confidence intervals use the frozen counterfactual-group bootstrap specified by `configs/response_consistency_v1.json`: 20,000 draws with seed 20260915. GS bootstrap resampling uses sources rather than tiles or individual frames.

GS descriptor construction and aggregation. The reported GS-v4 number is produced by `tools/probe_gs_orb_confirmation.py`, `src/iac_new/gs_pnp_geometry.py`, and `src/iac_new/static_geometry_probe.py`, and finalized by `tools/finalize_gs_model_confirmation.py`. It is not produced by the older RAFT/flow or five-dimensional structural-flow scorer. Those artifacts remain historical diagnostics and are not part of the frozen v4 result.

For an interval, the dense probe uses a 3×4 spatial grid covering the lower 75% of the image. For tile (r, c), with $r \in \{0, 1, 2\}$ and $c \in \{0, 1, 2, 3\}$, descriptor indices are

$$j = 2(4r + c) \quad \text{(log-depth scale)}, \qquad j + 1 \quad \text{(local expansion/divergence)}.$$

Table 8: AS counts on the shared 30-source pool. Progress support is shown as supported/declared intervals. Turns and same-direction counts refer to eligible branch rows.

Model	Sources	Branches	Progress support	Agree	Turns	Same dir.
DriveWAM	30	60	159/240	104	37	35
Epona	30	30	114/120	95	18	18
DriveVA	30	30	87/120	55	18	16
WorldDrive	30	30	115/120	89	23	23

The scale descriptor is the median of $\log(z_t/z_{t+1})$ over valid tile pixels. The expansion descriptor is the median of the divergence of the pose-induced dense reprojection field. Pixels enter the mask only after ORB correspondence, full EPNP/RANSAC pose fitting, held-out reprojection (at most 3 pixels), and log-depth residual (at most 0.25) checks. A tile requires at least 32 valid pixels and 15% tile support; a source requires at least two supported intervals. Missing descriptors are omitted, never zero-filled.

Let g_{ikj} and r_{ikj} denote generated and logged descriptors for source i, interval k, and descriptor j. Their normalized residual is $e_{ikj} = |g_{ikj} - r_{ikj}|/\sigma_j$, with real-only frozen scale σ_j. An interval is retained only if at least four descriptors have finite common support. Let $\mathcal{K}_i$ contain these intervals. For class $q \in \{s, e\}$ (scale or expansion), $\mathcal{E}_{i,q}$ pools all common finite entries (k, j) of that class with $k \in \mathcal{K}_i$. The supported-source set is

$$\mathcal{I}_{\mathrm{GS}} = \{i : |\mathcal{K}_i| \geq 2,\ |\mathcal{E}_{i,s}| \geq 4,\ |\mathcal{E}_{i,e}| \geq 4\}.$$

Aggregation follows Eq. (10):

$$\begin{aligned} s_{i,q} &= \exp\big(-\operatorname{median}_{(k,j)\in\mathcal{E}_{i,q}} e_{ikj}\big), & s_i &= \sqrt{s_{i,s}s_{i,e}}, \\ \mathrm{GS}_{\mathrm{cond}} &= \operatorname{median}_{i\in\mathcal{I}_{\mathrm{GS}}} s_i, & \mathrm{GS}_{\mathrm{obs}} &= \frac{|\mathcal{I}_{\mathrm{GS}}|}{N_{\mathrm{GS}}}. \end{aligned} \tag{13}$$

Here $N_{\mathrm{GS}} = 30$ is the declared source count per model. Class medians pool entries across retained intervals and tiles; no intermediate interval-score median is used. Fewer than two retained intervals or fewer than four common entries in either class makes the whole source unavailable. Both classes are mandatory: there is no scale-only or expansion-only fallback, and missing values are not zero-filled. Specificity compares the same-support score with fixed-point-free, support-preserving derangements of logged references. A single-number GS result is eligible only when coverage is at least 0.90, the matched score exceeds the shuffled 95th percentile, and the repeated-frame negative control passes.

The support mask is the intersection of finite generated and finite logged values at the same (k, j) location; generated-only or logged-only values are never imputed. For each descriptor index, calibration uses only the separate real-video reference split:

$$\sigma_j = \max(1.4826\ \operatorname{median}_u |r_{uj} - \operatorname{median}_v r_{vj}|,\ 0.01),$$

where u, v range over reference descriptors and unsupported calibration features are excluded. These scales are frozen before any WAM generated frame is inspected. GS v4 confirmation rows are one generated/logged pair per source, so no left/right branch averaging enters the reported GS number. The GS bootstrap resamples source rows, not tiles or individual frames.

A.3 Complete WAM Result Audit

The following tables give the raw AS counts and the main-text RCS evaluation counts. AS uses the identical 30-source pool for every model. RCS instead uses each model's valid intervention pool, so its rows are not a pooled cross-model ranking. All percentages are on a 0–100 scale; "Cond." is conditional agreement on supported measurements and "Obs." is observation coverage.

The RCS response columns distinguish response occurrence from conditional direction agreement. Act. is the material action-response rate among valid pairs; Vis. is the material visual-response rate among action-responsive pairs; Dir. is direction agreement among pairs with both responses. For example, DriveWAM's RCS of 38.1 combines a 49.5% action-response rate, an 83.2% conditional visual-response rate, and 92.5% conditional direction agreement. A detected response does not by itself establish correct execution of the condition.

Table 9: RCS audit on model-specific intervention pools.

Model	Pairs	Act.	Vis.	Dir.	RCS	95% CI
DriveWAM	745	49.5	83.2	92.5	38.1	$[34.8, 41.6]$
Epona	40	100.0	95.0	92.1	87.5	$[77.5, 97.5]$
DriveVA	50	74.0	43.2	62.5	20.0	$[10.0, 32.0]$
WorldDrive	50	60.0	66.7	100.0	40.0	–

A.4 Frozen Readers, Thresholds, and Missing Evidence

Reloc3r-512 supplies the candidate-blind signed visual yaw used by AS and RCS. The formal longitudinal AS channel uses Metric3Dv2-v2-S, static cross-frame correspondences, and known-rotation PnP. Its support gate requires at least 40 correspondences, an inlier fraction of at least 0.10, and median reprojection error at most 3 pixels. The yaw deadband is 0.0128 rad. Progress bins have edges at $0.5, 3, 6, 9$ m; accepted comparisons use a tolerance of $\max(3\,\mathrm{m}, 0.5p^A)$ or an adjacent bin, subject to the frozen stop-bin rule. Raw metric distance and speed are diagnostic outputs, not validated exact-motion estimates. AS calibration and GS descriptor scales use real logged video. The GS correspondence and rigid-support configuration was fixed after development and before the independent WAM confirmations.

The GS v4 reader uses Metric3Dv2-v2-S depth and ORB matches (ratio 0.80; 5,000 features; fast threshold 7; edge threshold 15). Full EPNP/RANSAC fits camera pose on one half of the matches and checks held-out rigid reprojection on the other half: at least six held-out points, at most 3 pixels reprojection error, and at most 0.25 log-depth residual. A supported pose reprojects dense depth into the next frame. The fixed 3×4 grid provides log-depth-scale and local-expansion descriptors, compared with real-only calibrated scales as in Eq. (9). Each tile requires at least 32 valid pixels and 15% support. A scored source requires at least two retained intervals and at least four common entries in each descriptor class. The reader and support thresholds are shared across the four models.

AS separates conditional yaw and progress agreement from measurement coverage. The tables display the geometric mean of the two channel coverages as Obs.; raw counts appear in Table 8. The coverage-aware summary is

$$\mathrm{AS}_{\mathrm{overall}} = \mathrm{AS}_{\mathrm{cond}} \sqrt{C_p C_\psi}. \tag{14}$$

It is interpreted together with the conditional score and both channel coverages, whose denominators differ. The RCS action and visual response deadbands are 0.0128 and 0.013317287876106008 rad, respectively. For yaw-response diagnostics, realized action differences are stratified as $[0, 0.0128)$, $[0.0128, 0.03)$, $[0.03, 0.06)$, $[0.06, 0.12)$, and $[0.12, \infty)$ rad. The first interval is the zero-response bin; the remaining bins describe response magnitude and do not replace the formal material-response threshold. RCS divides directionally matched material responses by all contract-valid pairs, including nonresponsive and visually unavailable pairs in the denominator. GS combines the two class similarities geometrically before taking the supported-source median and reports source coverage separately. A missing visual measurement does not become a zero-valued motion estimate. The GS single-number eligibility gate requires at least 90% source coverage, a matched-reference score above the support-preserving shuffled reference 95th percentile, and a passed repeated-frame control.

For clarity, the AS table's observation value is $C_{\mathrm{obs}} = \sqrt{C_p C_\psi}$. Its component coverages C_p and C_ψ use different denominators, so Obs. alone does not provide their raw support counts. Thus $\mathrm{AS} = \mathrm{AS}_{\mathrm{cond}} C_{\mathrm{obs}}$ and the observation value is not an additional agreement score.

For bin lookup, visual and action progress use $p^+ = \max(p, 0)$; the stop bin matches only stop or very-short bins. The signed metric-tolerance test and the bin-agreement test are combined by OR as in Eq. (5), so either can accept an interval. Clipping affects bin lookup only and does not add an exclusion gate for negative progress.

The frozen visual layer has separate readers for separate claims: Reloc3r-512 supplies candidate-blind yaw for AS and RCS; Metric3Dv2 with static SIFT correspondences and known-rotation translation supplies the coarse AS progress channel; and the GS v4 reader uses Metric3Dv2 with ORB,

Table 10: **AS sensitivity to controlled video-action mismatches on DriveWAM.** AS decreases under both yaw and progress perturbations; random pairing yields the lowest score among the tested controls.

Mismatch	Clean	Small	Medium	Large
Yaw deviation	64.0	63.9	60.8	59.2
Progress deviation	64.0	56.3	56.2	54.8
Random pairing	–	–	–	41.6

full EPNP/RANSAC, held-out rigid reprojection, and dense depth-reprojection descriptors. The earlier RAFT and SegFormer experiments are not used by any formal score.

Reader-only validation. On the held-out NAVSIM videos used for reader validation, the longitudinal channel contains 95 rows and 380 declared intervals. Its measurement coverage is 88.16%, mean absolute error is 1.686 m, median absolute error is 0.865 m, and Spearman correlation is 0.665, exact-bin accuracy is 58.5%, and adjacent-bin accuracy is 89.6%; supported progress agreement under the main-text AS rule is 90.1%. On the 120-clip yaw-validation set, Reloc3r-512 achieves 97.5% direction accuracy on applicable turns and 91.7% three-state accuracy. In the 95-sequence AS validation, all applicable yaw agreement decisions are preserved, giving $\mathrm{AS}_{\mathrm{cond}} \approx 100\sqrt{1 \times 0.901} \approx 94.9$, compared with 100.0 using reference motion. These checks assess the measurement layer rather than WAM performance.

A.5 Confirmation Uncertainty and Controls

Statistical comparison of AS on the common 30-source pool requires resampling sources rather than frames, keeping each source's branches together. No cross-model significance claim is made here. RCS confidence intervals, where reported, resample intervention source groups using the frozen 20,000-draw protocol. Results without confidence intervals are interpreted descriptively. The GS bootstrap protocol uses 20,000 source-level draws for its conditional score. Specificity is tested with fixed-point-free, support-preserving derangements of logged references. Generated branches retain their original support masks during each shuffle. Repeated future frames are an additional negative control: a motionless future must not obtain a positive specificity margin merely from scene appearance. These targeted controls assess sensitivity to lost pair correspondence and static-scene shortcuts under the specified perturbations.

The DriveWAM mismatch control keeps the visual future fixed while perturbing the action-side motion. AS decreases across the tested yaw and progress perturbation levels. The Small/Medium/Large labels denote ordered perturbation levels in this AS control; they are separate from the RCS yaw-response bins and the RCS-progress multipliers defined above. The large yaw and progress settings reduce AS by 4.8 and 9.2 points, respectively; these drops do not compare sensitivity per unit perturbation. Random pairing is reported as a separate negative control.

DriveVA's score remains a conditional diagnostic because its source coverage is below the single-number eligibility threshold. Passing the coverage threshold alone does not establish that all controls pass. Repeated frames need not have zero similarity: static geometry can retain nonzero agreement. The control requires a nonpositive specificity margin and no significant excess over shuffled references. DriveWAM's signed identity margin is $0.2581 - 0.2766 = -0.0185$; all four models pass the repeated-frame control. On a separate 30-source real-video development control, the ORB reader scores 30/30 sources on clean and $\sigma = 2$ Gaussian-blurred futures and 27/30 on $\sigma = 4$. This blur experiment measures one reader sensitivity; it cannot assign every generated-domain failure to either the reader or the WAM.

A.6 Reproducibility Boundary

The evaluation repository contains protocol configurations, source-pool hashes, aggregate reports, and a byte-level freeze manifest for an independent audit of score definitions and retained results. The AS comparison uses the same 30 source identities for all four models. Full image-level reruns

Table 11: GS support and repeated-frame audit. Scores are conditional on supported sources; coverage is reported separately. DriveVA falls below the 90% coverage eligibility gate.

Model	Supported/declared	GS	Coverage	Shuffled $q_{.95}$	Margin
DriveWAM	27/30	0.3728	90%	0.2766	+0.0962
WorldDrive	30/30	0.5587	100%	0.4069	+0.1518
Epona	30/30	0.6184	100%	0.4799	+0.1385
DriveVA	21/30	0.4846	70%	0.4509	+0.0337

Model	Identity score	p-value	Repeated-frame control
DriveWAM	0.2581	1.000	Pass
WorldDrive	0.2225	1.000	Pass
Epona	0.2346	1.000	Pass
DriveVA	0.2456	1.000	Pass

additionally require licensed NAVSIM data, the corresponding WAM checkpoints and generated videos, and the registered visual-reader weights. Private paths and source identities are excluded from the aggregate reports prepared for release. Historical RAFT-based GS artifacts and earlier diagnostic model pools do not enter the v4 confirmation results.

A.7 Policy Fine-Tuning Protocol

Table 12: Policy-training audit. Fine-tuning results are mean ± standard deviation across three training seeds. The original checkpoint is reported separately. All Δ values are relative to matched-budget action-only fine-tuning.

Training strategy	DriveWAM PDMS	Δ	WorldDrive PDMS	Δ
Original checkpoint	92.0827	–	92.2326	–
Action-only fine-tuning	92.0597 ± 0.0319	0	92.5086 ± 0.0194	0
GT motion targets	92.5078 ± 0.0191	0.4481	92.5081 ± 0.0149	−0.0005
Shuffled visual targets	92.5156 ± 0.0786	0.4559	92.5067 ± 0.0494	−0.0019
Paired IAC targets	93.5953 ± 0.1378	1.5356	93.5081 ± 0.2190	0.9995

The policy experiment uses the same frozen visual estimator and four-second temporal alignment protocol as the evaluation. Training, development, and PDMS evaluation use non-overlapping source samples. A future-motion target is extracted from the initial checkpoint's generated video by the frozen visual reader and is detached from the policy computation graph. Only action-side parameters are updated; the video generator and the visual reader remain frozen throughout fine-tuning. The auxiliary loss uses both yaw and progress in the IAC motion representation and follows its support criteria; it does not directly optimize the reported AS. For WorldDrive, the trainable action-side component is the action adapter; the pretrained WorldDrive model remains frozen. The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda^{\star}\mathcal{L}_{\text{mot}}, \tag{15}$$

where $\mathcal{L}_{\text{mot}}$ is the robust yaw-progress consistency loss defined in Section 3.5, and $\lambda^{\star}$ is selected on source-disjoint development samples and then held fixed for confirmation. The comparison uses the same initialization, data budget, optimizer budget, and evaluation split for action-only, GT-motion-target, shuffled-target, and paired-IAC-target runs. The paired target is extracted from the model's own frozen generated future, whereas the GT-motion control derives its yaw and progress target from the logged pose trajectory. The public code release will specify the learning rates, optimization steps, batch sizes, auxiliary-loss weights, and training seeds for all fine-tuning settings.

Table 6 reports the resulting independently evaluated PDMS values. On DriveWAM, paired IAC supervision improves over the matched action-only run by 1.5356 points; on WorldDrive the im-

provement is 0.9995 points. These numbers are downstream evidence for the usefulness of the detached target and are reported separately from the three frozen IAC evaluation metrics.

B Limitations and Future Work

The reported fine-tuning results establish downstream planning utility, while changes in AS and RCS after fine-tuning remain unreported.

IAC evaluates coarse yaw and longitudinal progress, leaving lateral displacement, detailed path shape, and interactions unmeasured. The tolerance-based progress criterion limits sensitivity to longitudinal discrepancies. The open-ended upper progress bin does not bound metric disagreement, and the action-conditioned yaw component does not evaluate visual turns when the action is straight.

Visual estimation is validated on logged videos, but its accuracy under generation artifacts remains uncertain; candidate-blindness and support checks do not establish unbiased measurement in that domain. RCS measures correspondence between yaw responses rather than correct execution of a condition, and its score also depends on response occurrence and observability.

GS compares geometry with one logged continuation, so an alternative valid future may receive a lower score. Model-specific RCS interventions and GS reference pools further limit direct cross-model ranking. Future work should strengthen generated-domain validation and extend evaluation to finer motion and interaction structure.

C Extended Related Work

Prediction and planning in driving world models. Driving world models connect future prediction with action generation through several forms of coupling. Epona (Zhang et al., 2025) combines modular trajectory and video prediction within an autoregressive diffusion framework. PWM (Zhao et al., 2026) uses action-free future-state forecasting to inform collaborative state-action prediction, while DriveLaW (Xia et al., 2026) supplies its planner with latent representations from a video generator. WorldDrive (Gui et al., 2026) transfers shared vision and motion representations from trajectory-conditioned scene generation to planning. DriveVA (Liu et al., 2026) and DriveWAM (Shi et al., 2026) bring video and action streams into shared generative processes. These architectures motivate an evaluation interface defined by the exported predictions. IAC requires a future video and its native action from the same inference instance and compares their expressed motion over a common horizon. Its applicability depends on this output interface rather than on a particular internal coupling architecture.

Visual quality and driving performance. Video and planning benchmarks assess different properties of a driving model. FVD (Unterthiner et al., 2018) evaluates generated video distributions through learned video features. NAVSIM (Dauner et al., 2024) uses non-reactive simulation to assess driving policies, while Bench2Drive (Jia et al., 2024) evaluates closed-loop behavior across interactive driving scenarios. These evaluations establish evidence about video quality or driving performance. They do not directly compare the motion expressed by a particular generated video with that of its paired native action. IAC adds this relation as an explicit evaluation target. Agreement in its coarse motion channels does not certify image quality, complete trajectory equivalence, or successful driving, so IAC scores should be interpreted alongside the corresponding video and planning evaluations.

Action fidelity and video-action agreement. Existing work already evaluates motion recovered from generated driving videos. ACT-Bench (Arai et al., 2024) estimates action categories and trajectories from generated scenes and compares them with the supplied instructions and trajectory conditions. DrivingGen (Zhou et al., 2026) evaluates ego-conditioned generation using recovered trajectories, with ADE for temporally corresponding positions and DTW for path-shape comparison. DriveVA (Liu et al., 2026) uses DPVO to recover video motion and evaluates trajectory error after fitting a two-dimensional similarity transform to the corresponding reference. These protocols provide direct evidence about action fidelity or video-trajectory alignment; visual motion recovery itself is therefore not unique to IAC.

IAC focuses on the relation between jointly produced video and native action under a declared, candidate-blind measurement protocol. Its visual calibration does not fit recovered motion to the action being assessed. AS summarizes supported agreement in coarse yaw and progress channels, while RCS asks whether the two outputs exhibit material, directionally corresponding yaw changes across a controlled condition pair. This paired-response question differs from checking a single generated video's fidelity to its input condition. GS supplies a separate comparison with logged future geometry. The distinctions concern the evaluated relation, calibration, and reporting of measurement support; they do not establish superiority to other evaluators without a direct comparison. Table 13 summarizes these evaluation objects.

Visual motion as measurement. The measurement layer builds on established geometric estimators. Reloc3r (Dong et al., 2025) predicts relative camera rotation and translation direction; the norm of its relative translation output is not a metric-distance estimate. Metric3Dv2 (Hu et al., 2024) addresses metric depth estimation across camera models. IAC uses relative rotation for signed yaw and combines metric depth with static image correspondences and camera geometry for coarse longitudinal progress. These components provide the basis for the frozen visual reader. Candidate-blindness prevents fitting the visual estimate to the evaluated action, while support checks and coverage reporting expose where measurements are available. Neither property alone establishes measurement accuracy under generation artifacts. The logged-video validation and generated-domain limitations must therefore remain explicit.

Generated futures as supervision. Methods such as PWM (Zhao et al., 2026) and DriveLaW (Xia et al., 2026) already use predicted future information to support planning. IAC studies a specific

supervisory use of an existing model's generated future: recover visual motion, detach it as an auxiliary target, and update the action side while keeping video generation and visual estimation frozen. This connects the evaluation representation to action learning without requiring a new world-model backbone. The reported PDMS comparison evaluates the planning utility of this supervision. It does not by itself establish that fine-tuning improves AS or RCS.

Table 13: Evaluation objects in related protocols. Each row summarizes the cited protocol; it does not claim that other evaluation designs are incompatible with that method.

Work / protocol	Quantities compared	Reported comparison	Question addressed
ACT-Bench (Arai et al., 2024)	Video-estimated action/trajectory and supplied condition	Instruction-execution consistency and trajectory alignment	Does the generated video follow the action condition?
DrivingGen, ego-conditioned track	Video-recovered trajectory and conditioning ego trajectory	ADE and DTW	Does the generated trajectory follow the supplied path?
DriveVA consistency analysis	Video-reconstructed trajectory and predicted trajectory	DPVO reconstruction and L2 error after 2D similarity alignment	Do the aligned trajectory shapes agree?
IAC	Native video-action pair; changes across conditions; generated/logged geometry	AS, paired yaw RCS, and separate reference-based GS with support reporting	Do coarse motions agree and respond together, and how does future geometry compare with the logged continuation?