ATTENDINGSCORER: AN END-TO-END DRIVING EVALUATOR THAT READ OUT TRAJECTORY EVALUATION SCORE FROM LATENT OF VISION-LANGUAGE-MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving vision-language models are now capable of acting as effective evaluators between competing trajectories, and providing accurate explanations for their choice. However, they still struggle to provide accurate quantitative scores to these on a non-comparative scale. Existing learned evaluators address this gap by either relying on predefined heuristic rules, or through techniques that involve finetuning the existing comparison-based evaluator to also provide a numerical score. The former approach struggles with unforeseen, long tail events, while the latter involves repeated modification of model weights which can have unintentional knock on effects on existing evaluator capability. We propose TrajSteering, a learned approach that aims to read out the trajectory score from the evaluator, AttendingScorer, without touching its weights. It trains only a small set of query embeddings, including a designated score query that aims to get a score through a sum of the context values, weighted by attention, as dictated by the other query embeddings. Our key architectural observation is that attention weights are thus analagous to classical filter taps, with the key distinction being that with filter taps the weights can take on any values, attention weights are restricted to what the frozen model can actually realize. Thus they can be approached with similar principles. TrajSteering combines two objectives. An automatic pseudo-score loss intends to train a scalar head to extract a numeric score from the score-query output, to match targets generated offline by a rule-gated teacher. A deflection-criterion, conditioned on the same teacher, identifies which query embeddings best isolate score-relevant signal from irrelevant noise, with the aim of selecting a readout capable of generalizing among the ones that fit. The backbone remains unchanged; the teacher is used only for one-time offline supervision. Using extensive experiments, AttendingScorer perform at a high level of evaluation quality.

1 INTRODUCTION

Autonomous driving is converging on a world-model view, in which one learned representation supports several downstream roles at once: a Driver that plans, a Simulator that imagines, and a Critic that judges (Li et al., 2025a). Two of the three roles have already become neural and end-to-end. The Critic has not: production validation still rests on hand-written rule tables that score a rollout by combining collision, drivable-area, traffic-light, etc.(Dauner et al., 2024; Cao et al., 2025). Because such a table is fixed in advance, it cannot exchange representations or explanations with the planner it is meant to improve, and performs poorly in *Out-of-Distribution* Scenarios that have not been pre-considered. The natural response, now visible across the field, is to hand judgment to a vision-language model (VLM) that can read the scene and output *semantic context* about why a maneuver is unreasonable. However, even though VLMs have been shown to be able to produce semantic output that correctly reflects the evaluation (for example, unnecessary lane change), they rely on pairwise comparison between trajectories. They are not yet able to produce quantitative evaluation scores comparable to *rule-based scores* Song et al. (2026)

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.

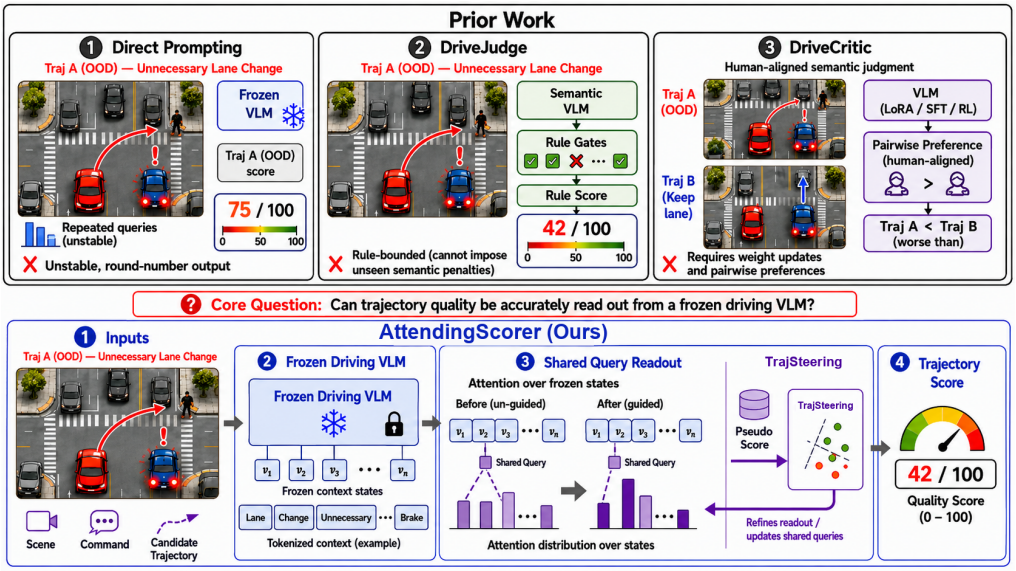


Figure 1: **High-level comparison of trajectory-evaluation paradigms.** Prior methods either prompt a frozen VLM directly, route VLM judgments through fixed rules, or modify model weights for task-specific evaluation. AttendingScorer instead keeps the driving VLM frozen and uses TrajSteering to learn a lightweight shared query readout that extracts a scalar trajectory-quality score from its existing representations.

That response is currently approached in one specific way: by rewriting the evaluator (Sun et al., 2025; Song et al., 2026; Chen et al., 2026). The gains are real: on a human-aligned test set built from simulation scenarios, the rule score agrees with expert preference far less often than the same backbone after task-specific adaptation (Song et al., 2026). But every such gain re-opens the weights, and re-opening the weights forfeits the guarantee a deployed evaluator most needs: that the model's other driving abilities are untouched with a strong reproducibility.

The problem worsens as evaluation expands to safety, comfort, assertiveness, human likeness, and reasoning efficiency: retraining the model and recollecting human preferences for every new dimension does not scale.

Existing work approaches this from three directions. Rule-based simulation metrics compute the score from a fixed set of sub-metrics and treat semantic context as something to be anticipated in advance (Dauner et al., 2024; Cao et al., 2025; Hao et al., 2026). Rule-gate evaluator let a vision-language model decide which sub-rules apply in a given scene, but return scoring authority to the rule set, so the model can excuse a penalty and never impose one (Sun et al., 2026). Weight-modified evaluators instead train the model to judge, through pairwise preference data, reward-model supervision, or an external scoring head attached to an adapted backbone (Song et al., 2026; Chen et al., 2026; Sun et al., 2025; Jiang et al., 2025). Despite their effectiveness, these methods share an implicit assumption that trajectory-scoring ability must be installed into an evaluator: either by rewriting its weights or by routing its judgment back through a fixed rule set; rather than read out through the evaluator's own native interface. That assumption is what makes adaptation cost, semantic drift, and per-dimension retraining feel unavoidable. However, since the VLM are able to produce convincing semantic output Song et al. (2026) , the core problem is *how to reflect the understanding on the score*. Our preliminary analysis (C) is designed around three observations. First, the judgment is already present: probing frozen VLM hidden states separates human-preferred from dispreferred trajectories far better than rule-based score method. Second, directly prompting THE VLM for the rating collapses ratings to a few round values, an issue that persists across prompt and decoding controls, meaning the failure lies in score readout rather than representation. Third, the automatic target does not determine how the score is read: among readouts trained to fit the same automatic scores equally well, held-out and long-tail behaviour diverges, and a deflection statistic computed inside the attention-weighted latent tracks which of them generalizes. Thus, fitting the target does not specify the readout; something must select among the readouts that fit it.

These observation suggest that core question is not whether a frozen driving model contains trajectory quality, but which readouts its own causal attention interface can realize, and which one should be selected. Appended trailing queries leave the original context untouched in a causal model, and their context contribution is exactly a weighted aggregation of frozen values: the defining form of a tap-weighted filter (Capon, 1969; Van Trees, 2002): positions as samples, attention weights as taps, trajectory quality as the signal, and formatting effects and irrelevant semantics as the interference. The classical criterion - signal deflection over nuisance variance at the output - has a closed-form maximum once a signal direction and a nuisance covariance are available, which collapses the design criterion into a single differentiable scalar. But the transfer is not a copy, and the difference is the contribution. Classical taps are free variables; here the taps are generated by a frozen network from shared query embeddings, not assigned to it, and one readout must serve the entire scene distribution rather than being re-optimized per scene. Directly designing per-instance taps is therefore unavailable; we relax the design into a distribution-level problem over the realizable readout family of the frozen model, whose coordinates are the shares of readout weight a shared query bank places on each context position. In that family, the classical optimum is a reference solution, not an algorithm. This motivates our central research question: *How much of the trajectory-quality information decodable from a frozen driving VLM's representation can be recovered through the readouts its own causal attention interface realizes?* **Hypothesis (H):** scoring ability need not be installed: a shared readout trained on the frozen interface can recover (reflect) human-aligned scoring approaching the probe ceiling of the representation.

To address this gap, we propose evaluation framework AttendingScorer, and our concrete method instantiated within it TrajSteering. We answer the question with a single frozen readout trained under two objectives. A pseudo-score objective fixes what to predict: since the judgment is already present (observation i), no weight moves (a scalar head reads the score query, bypassing the numeric-token channel (observation ii indicated)), with its target generated automatically by a *rule-gated teacher* Sun et al. (2026). A teacher-conditioned deflection objective fixes how to read (observation iii): inside the attention-weighted context latent, it estimates the direction along which the automatic scores vary, whitens the variation they do not explain, and rewards readouts whose signal survives.

Our contributions are summarized as follows.

- To the best of our knowledge, we are among the first to realise an *end-to-end self-driving scoring evaluator* that emits the evaluation score under the frozen VLM.
- We introduce a new perspective of trajectory scoring: given a vision-language model with strong (modality) contextual understanding, eliciting a trustworthy score is a problem of *guiding and filtering its latent attention*, rather than teaching the model semantic understanding from scratch.
- Thus, we present TrajSteering, a lightweight post-training design that guides an existing driving vision-language model to emit reasonable end-to-end evaluation score for trajectory evaluation.
- Extensive experiments show that this design produces reliable scores with only lightweight post-training adjustment, with particularly strong gains on out-of-distribution scenarios.

2 Related Work

From rule metrics to semantic evaluators. Driving evaluation is built on closed-loop and data-driven protocols: nuPlan provides a real-world closed-loop benchmark (Karnchanachari et al., 2024), while NAVSIM, PseudoSim, and StyleDrive develop scalable sub-metric scoring and its synthetic- and style-conditioned variants (Dauner et al., 2024; Cao et al., 2025; Hao et al., 2026). InterPlan further exposes long-tail cases in which rule penalties require manual correction (Hallgarten et al., 2024). Learned scorers replace explicit aggregation with reward-world-model, imagined-future, or dense-candidate scoring, but still inherit supervision from rule metrics (Jiang et al., 2025; Li et al., 2025a;b). Semantic evaluators instead use VLMs to adapt on rule–expert disagreements, gate applicable rules, emit multidimensional ratings, or read a scoring token from an adapted backbone (Song et al., 2026; Sun et al., 2026; Chen et al., 2026; Sun et al., 2025); related multimodal judges likewise rely on full finetuning and remain more reliable for pairwise than absolute ratings (Xiong et al., 2025; Chen et al., 2024). Thus, existing methods obtain semantic scoring either by changing model weights or by routing judgment back to a fixed rule set. We instead freeze the backbone and use the rule-gated score only as a scalable, structurally limited supervisor whose ceiling is measured (Sec. H).

Method	No weight update	Absolute scalar score	Evidence-grounded rationale	Range not rule-capped
Rule-based metrics				
EPDMS / NAVSIM	✓	✓	✗	✗
SM-PDMS / StyleDrive	✓	✓	✗	△
Learned scorers / reward models				
RWM	✗	✓	✗	✗
WoTE	✗	✓	✗	△
GTRS	✗	✓	✗	△
Vision-language evaluators				
DriveJudge	✗	✓	△	✗
DriveCritic	✗	✗	✓	✓
DriveReward	✗	✓	✓	△
VLoE	✗	✓	△	✓
TrajSteering (ours)	✓	✓	✓	✓

Table 1: Capability comparison across evaluation paradigms. The columns assess whether each method preserves backbone weights, outputs an absolute scalar score, provides evidence-grounded rationales, and is not limited by a fixed rule range. Together, these properties characterize a scalable and semantically grounded validator. ✓ denotes direct support, △ partial support, and ✗ no direct support.

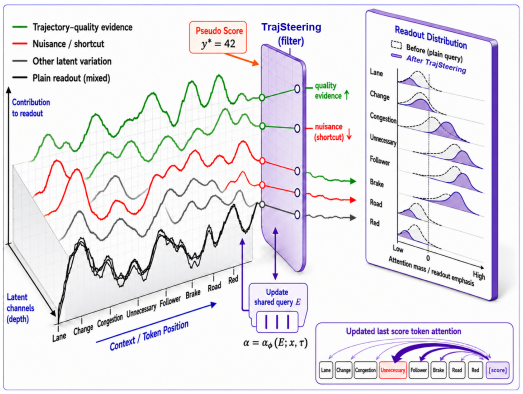


Figure 2: **Constrained filtering view of TrajSteering.** Teacher-conditioned guidance reshapes the frozen model's realizable attention readout toward trajectory-quality evidence.

Frozen representations and native readout. Probing work shows that information may be present yet difficult to decode: linear probes measure simple accessibility, control tasks distinguish probe learning from represented information, and recent studies find reward signals in frozen features (Alain & Bengio, 2016; Hewitt & Liang, 2019; Guo et al., 2026; Li et al., 2025c). Learned queries and prompts provide a native aggregation mechanism over frozen encoders (Li et al., 2023; Li & Liang, 2021; Zhou et al., 2022; Jia et al., 2022); most directly, Lou et al. (2026) append trainable queries to a frozen multimodal model and report competitive readout without changing base abilities. Our question is narrower than general probing: rather than ask what an arbitrary external decoder can extract, we optimize only the readouts realizable by the frozen model's own causal interface, using the external probe solely to estimate the representation ceiling.

Constrained filtering and weak supervision. The selection criterion builds on classical constrained filtering: MVDR, generalized sidelobe cancellation, and deflection-ratio maximization suppress nuisance under an explicit signal constraint (Capon, 1969; Griffiths & Jim, 1982; Van Trees, 2002). We transfer this principle to a frozen nonlinear front end whose attention taps are generated rather than freely assigned, and distinguish it experimentally from plain distillation, target-agnostic whitening, and hard position selection (Sec. D.2). This also connects to weak-to-strong generalization, where gains can depend on competence already present in the student (Burns et al., 2023). Here the rule-gated teacher has an identifiable blind spot: it cannot penalize behavior absent from its rule set. We test whether the constrained readout can exceed that teacher only as a secondary result; failure would reduce the claim to honest pseudo-label distillation, not refute the readout formulation.

3 PRELIMINARIES

Figure 2 previews the constrained-filtering formulation instantiated in Sec. 4.3.

Frozen causal readout Let x denote the multimodal driving context—the rendered scene, driving command, and ego state—and let τ denote a candidate trajectory rendered in the same input format. A frozen driving vision–language model Φ maps the assembled prompt to contextual states $H(x,\tau) = (h_1, \ldots, h_n)$. We append K auxiliary query embeddings and one score query, $E = (e_1, \ldots, e_K, e_s) \in \mathbb{R}^{(K+1)\times d}$, while keeping every parameter of Φ fixed.

Because the appended queries trail the original context in a causal model, they may attend to all preceding positions while leaving every original context state unchanged. For final-layer attention head r, the contribution of the original context to the score-query attention branch is exactly

$$o_E^{(r)}(x,\tau) = \sum_{i \in \mathcal{C}(x,\tau)} \alpha_{i,\Phi}^{(r)}(E; x, \tau)\, v_i^{(r)}(x,\tau), \qquad \alpha_\Phi^{(r)}(E; x, \tau) \in \Delta^{|\mathcal{C}|-1}, \tag{2}$$

where $\mathcal{C}$ indexes the original multimodal context positions and $v_i^{(r)}$ is the frozen value projection at position i. Combining attention heads through the frozen output projections gives the attention-filtered context latent

$$z_E(x,\tau) = \sum_{r=1}^{R_h} W_O^{(r)}\, o_E^{(r)}(x,\tau). \tag{3}$$

When the input is fixed we suppress the arguments and write α_Φ and $\mathcal{C}$ for $\alpha_\Phi(E;x,\tau)$ and $\mathcal{C}(x,\tau)$. Equation 2 is the exact filtering object: the score query's context-attention contribution is a weighted aggregation of frozen context values. We do not identify the complete score-query state with a linear filter; the reporting loss reads the full state h_s, whereas the filtering criterion is applied only to z_E.

Classical Reference and Realizable Readout Space For a signal template θ and positive-definite nuisance covariance Σ, classical free-tap filtering maximizes

$$\rho_{\text{free}}(\alpha) = (\alpha^\top\theta)^2 \,/\, (\alpha^\top\Sigma\alpha), \qquad \alpha^\star_{\text{free}} \propto \Sigma^{-1}\theta, \tag{4}$$

which whitens nuisance before correlating with signal (Capon, 1969; Van Trees, 2002); see App. C.1. Equation 4 is only a reference here: the corresponding θ and Σ are unobserved behind the frozen nonlinear model, and the taps are generated by one shared query bank rather than assigned freely per input. These failures respectively motivate teacher-conditioned statistics and a distribution-shared optimization over the frozen model's realizable readout family:

$$\mathcal{A}_\Phi(x,\tau) = \{\alpha_\Phi(E;x,\tau) : E \in \mathcal{E}\} \subseteq \prod_{r=1}^{R_h} \Delta^{|C|-1}, \qquad E^\star = \arg\min_{E\in\mathcal{E}} \mathbb{E}_{(x,\tau)\sim\mathcal{D}}[\mathcal{L}(x,\tau;E)]\,. \tag{5}$$

Proposition 1 (Frozen-model realizability constraint)**.** *With Φ fixed, equation 5 restricts every trainable score-query context-attention pattern to $\mathcal{A}_\Phi(x,\tau)$ while sharing one query bank E across the data distribution. Training therefore selects a distribution-level readout rather than per-instance filter taps, and the unconstrained classical optimum executable when realizable through $\mathcal{A}_\Phi(x,\tau)$.*

4 Method: TrajSteering

4.1 Problem Statement and Method Overview

Given a frozen driving vision–language model Φ, a scenario context x, and a candidate trajectory τ, we seek a scalar trajectory-quality score. Training uses only offline Scheme-B pseudo scores $y^\star(x,\tau) = T_B(x,\tau) \in [0,1]$; no human rating, preference pair, or anchoring set enters training or model selection. We train one query bank E shared across the data distribution and scalar-head parameters ω with Φ fixed, so the goal is not unconstrained score regression but fitting $y^\star$ with a distribution-shared readout $\alpha_\Phi(E;x,\tau) \in \mathcal{A}_\Phi(x,\tau)$. Endpoint fit alone does not identify which realizable readout should produce the score; Sec. 4.3 derives the additional selection criterion.

Figure 7 summarizes the workflow. Following SLQ's trailing-query construction over frozen multimodal models (Lou et al., 2026), whose appended queries preserve the original context under causal masking, the frozen backbone processes (x,τ) once with the shared query bank E, yielding the full score-query state $h_s(E;x,\tau)$ for scalar prediction and the context-attention latent z_E of equation 3 for readout selection. Unlike SLQ, where the trailing state serves as a retrieval embedding, here E parameterizes the realizable attention taps $\alpha_\Phi(E) \in \mathcal{A}_\Phi$. Thus, the pseudo-score loss fixes what to predict, while teacher-conditioned deflection selects how to read; only (E,ω) are trained, and the deflection computation is discarded at inference.

4.2 Query Parameterization and Scalar Scoring

A lightweight scalar head predicts $\widehat{y} = g_\omega\big(h_s(E;x,\tau)\big)$, trained with a Huber loss on the teacher scale (line 3 of Alg. 1).

A plain query readout stops here. Although its score loss updates E, identical training predictions may be realized by different attention readouts: $\alpha_\Phi(E_1;x,\tau) \neq \alpha_\Phi(E_2;x,\tau)$ for some inputs, because the scalar head can compensate internal differences. Endpoint supervision therefore fixes what to predict but not how to read; observation (O3) motivates a criterion on z_E itself.

Algorithm 1 One TrajSteering training step.

Require: pseudo-labelled set $\mathcal{D}^\star$; frozen Φ; trainable (E, ω); loss ℓ; weight λ; batch size B; stabilizers $\varepsilon_s, \varepsilon_D$;
1: Sample $\mathcal{B} = \{(x_b, \tau_b, y_b^\star)\}_{b=1}^B \subset \mathcal{D}^\star$
2: Run frozen Φ on $\{(x_b, \tau_b)\}_{b=1}^B$ with appended E; obtain $\{(h_{s,b}, z_b)\}_{b=1}^B$
3: $\widehat{y}_b \leftarrow g_\omega(h_{s,b})$ for all b; $\mathcal{L}_{\text{score}} \leftarrow B^{-1} \sum_b \ell(\widehat{y}_b, y_b^\star)$
4: $(\widehat{s}_E, \widehat{D}_E, \rho_{\text{TCD}}) \leftarrow$ TEACHERDEFLECTION$(\{(z_b, y_b^\star)\}_{b=1}^B, \varepsilon_s, \varepsilon_D)$
5: $\mathcal{J} \leftarrow \mathcal{L}_{\text{score}} - \lambda \rho_{\text{TCD}}$; $(E, \omega) \leftarrow \text{OptStep}(E, \omega; \nabla_{E,\omega} \mathcal{J})$, with Φ fixed
6: **return** updated (E, ω)

Algorithm 2 Teacher-conditioned deflection. The batch size B doubles as the sample size

Require: filtered latents and pseudo scores $\{(z_b, y_b^\star)\}_{b=1}^B$; numerical stabilizers $\varepsilon_s, \varepsilon_D > 0$
1: $\bar{y}^\star \leftarrow B^{-1} \sum_b y_b^\star$, $\bar{z} \leftarrow B^{-1} \sum_b z_b$;
2: $\widetilde{y}_b \leftarrow y_b^\star - \bar{y}^\star$, $\widetilde{z}_b \leftarrow z_b - \bar{z}$ for all b
3: $\widehat{s}_E \leftarrow (\sum_b \widetilde{y}_b \widetilde{z}_b) \,/\, (\sum_b \widetilde{y}_b^2 + \varepsilon_s)$
4: $r_b \leftarrow \widetilde{z}_b - \widehat{s}_E \widetilde{y}_b$ for all b
5: $\widehat{D}_E \leftarrow \text{diag}\big(B^{-1} \sum_b r_b \odot r_b\big) + \varepsilon_D I$
6: $\rho_{\text{TCD}} \leftarrow \widehat{s}_E^\top \widehat{D}_E^{-1} \widehat{s}_E$
7: **return** $(\widehat{s}_E, \widehat{D}_E, \rho_{\text{TCD}})$

4.3 TEACHER-CONDITIONED DEFLECTION

We therefore apply the selection criterion to the exact context-attention branch z_E (Sec. 3), while the reporting loss still reads h_s.

Along any score-relevant direction w, we model the per-position contribution to equation 3 as

$$w^\top W_O^{(r)} v_i^{(r)} = \theta_i^{(r)} + \epsilon_i^{(r)}, \tag{7}$$

where $\theta_i^{(r)}$ is the trajectory-quality component carried by position i in head r, and $\epsilon_i^{(r)}$ collects irrelevant semantics, formatting effects, and other nuisance variation. Stacking over positions and heads gives vectors $\theta, \epsilon \in \mathbb{R}^{|\mathcal{C}|R_h}$, indexed exactly as the taps α_Φ of Eq. equation 5.

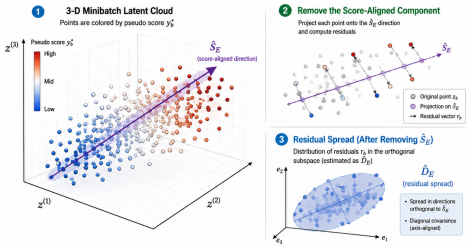


Figure 3: **Geometry of teacher deflection:** strong pseudo-score-aligned variation relative to residual nuisance.

Summed through equation 2 and equation 3, the readout along w is $w^\top z_E = \alpha_\Phi^\top \theta + \alpha_\Phi^\top \epsilon$: exactly the tap-weighted signal-plus-interference form of Sec. 1, so Algorithm 2 can execute its free-tap optimum rather than invent a criterion.

Algorithm 2 executes it under the substitution $(\theta, \Sigma, \alpha) \mapsto (\widehat{s}_E, \widehat{D}_E, w)$: estimable latent statistics replace the unobserved signal and nuisance statistics, and the closed-form solve runs over a latent direction w, while the attention taps remain generated by the frozen model and are selected only indirectly through $\nabla_E \rho_{\text{TCD}}$ inside $\mathcal{A}_\Phi$. The classical improvement guarantee is therefore not inherited, and the benefit is evaluated empirically (App. D.2).

For the minibatch estimates returned by Algorithm 2, the teacher-conditioned deflection is

$$\rho_{\text{TCD}}(E; \mathcal{B}) = \max_{w \neq 0} \big(w^\top \widehat{s}_E\big)^2 \,/\, \big(w^\top \widehat{D}_E\, w\big) = \widehat{s}_E^\top \widehat{D}_E^{-1} \widehat{s}_E. \tag{9}$$

with maximizing direction $w^\star \propto \widehat{D}_E^{-1} \widehat{s}_E$ (Fig. 3). The complete training objective is

$$(E^\star, \omega^\star) = \arg\min_{E,\omega} \left[\mathcal{L}_{\text{score}}(E, \omega) - \lambda \rho_{\text{TCD}}(E; \mathcal{B})\right]. \tag{10}$$

with $\lambda > 0$ weighting the criterion; $\lambda = 0$ recovers the plain query readout used as the matched baseline (Table 3). Hyperparameter roles and sweeps are listed in App. E.

Algorithm 1 separates reporting (line 3) from readout selection (line 4, via Algorithm 2); line 5 updates only (E, ω). Appending empirically outperforms prepending (Lou et al., 2026).

5 EXPERIMENTS

5.1 SETUP

Backbone and supervision. All VLM-based methods start from the same Qwen2.5-VL-7B checkpoint used as the base model of DriveCritic, before its task-specific trajectory-evaluation adaptation (Song et al., 2026). Direct prompting, fixed-query readout, plain query readout, and TrajSteering keep every backbone parameter frozen. The LoRA and full-backbone SFT baselines start from

Table 2: Public human-aligned trajectory scoring on WOD-E2E. All learned methods use zero task-specific human labels during training and model selection.

Method	Pairwise Acc. (%) ↑	Kendall τ_b ↑	NDCG@3 ↑	Fixed-scale MAE ↓
Imitation score (negative ADE/FDE)	64.5	0.28	0.87	–
Trajectory-only MLP scorer	61.0	0.21	0.85	2.64
Same-backbone direct prompting	54.2	0.07	0.81	2.92
Plain query readout	66.8	0.33	0.89	2.31
Same-backbone LoRA scorer	71.5	0.42	0.92	2.02
TrajSteering (Ours)	70.3	0.40	0.91	2.10

Table 3: Matched pseudo-score fit and public cross-domain human generalization. Plain query readout and TrajSteering share the same frozen checkpoint.

Variant	Pseudo-val. RMSE ↓	Pseudo-val. SRCC ↑	WOD pairwise Acc. (%) ↑	WOD NDCG@3 ↑	Held-out ρ_{TCD} ↑
Fixed/random queries, scalar head only	0.142	0.712	61.5	0.86	3.1
Plain query readout ($\lambda = 0$)	0.096	0.888	66.8	0.89	6.4
TCD with permuted pseudo scores	0.097	0.885	66.3	0.89	6.1
TCD on full score-query state	0.095	0.889	67.9	0.90	6.9
Same-backbone LoRA scorer	0.089	0.901	71.5	0.92	–
TrajSteering (Ours)	0.095 ±0.002	0.891 ±0.001	70.3 ±0.003	0.91 ±0.002	8.7 ±0.001

the identical checkpoint but update low-rank adapters or all backbone parameters, respectively. Full teacher and implementation details appear in App. F.

Benchmarks. Pseudo-score training and validation use public NAVSIM and Pseudo-Simulation scene–trajectory pools (Dauner et al., 2024; Cao et al., 2025). External human-alignment evaluation uses the public WOD-E2E validation set (Xu et al., 2025), which provides three candidate trajectories per long-tail driving segment together with existing expert ratings on a 0–10 scale. These ratings form a locked external test and are never used for training, prompt selection, hyperparameter tuning, checkpoint selection, seed selection, or output calibration.

The semantic-regime analysis uses pre-registered subsets defined automatically from the fixed teacher score, available rule outputs, and existing human ratings; no item is manually selected according to model performance. No new human annotation is collected for any experiment.

Metrics and protocol. On WOD-E2E, we report pairwise accuracy, Kendall's τ_b, NDCG@3, and fixed-scale MAE. Native-channel analysis additionally reports the effective number of score levels, prompt sensitivity, and valid-output rate. Semantic-regime analysis reports subgroup MAE, TranscendRate, and worst-group MAE. Causal interventions report paired performance changes and held-out ρ_{TCD}. On held-out pseudo-labelled data, we report RMSE and Spearman rank correlation coefficient (SRCC) against the automatic teacher. Adaptation comparisons report trainable parameters, GPU-hours, peak training memory, inference time per candidate, and changes in the backbone's original driving and multimodal abilities.

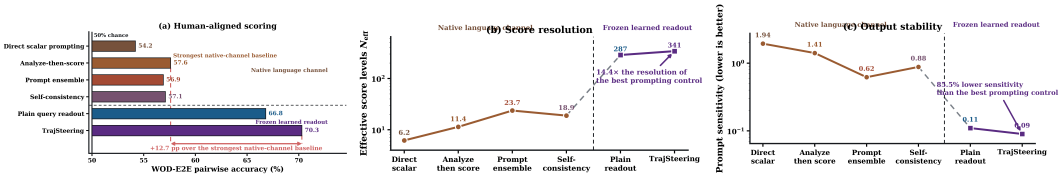
5.2 Main Result

Public Human-Aligned Trajectory Scoring Table 2 evaluates whether scores learned only from automatic supervision agree with the existing WOD-E2E human ratings. The plain-readout versus direct-prompting comparison tests whether the frozen attention interface is a better scoring channel, while the LoRA comparison measures the remaining gap to weight modification.

Matched-Fit Readout Comparison Table 3 isolates the contribution of teacher-conditioned deflection. Every trainable row uses the same correct pseudo scores in its endpoint scoring loss. In the permutation control, only the targets used to construct the deflection criterion are permuted.

As shown in Table 2, TrajSteering raises pairwise accuracy from 54.2% with direct prompting to 70.3% and NDCG@3 from 0.81 to 0.91; it also improves over the plain query readout by 3.5 percentage points and reduces MAE from 2.31 to 2.10. Its pairwise accuracy is only 1.2 points below the same-backbone LoRA scorer (70.3% vs. 71.5%) without updating the backbone, showing that most of the human-aligned scoring ability can be recovered through the frozen query interface.

Table 3 shows nearly matched pseudo-score fit between TrajSteering and the plain readout (0.095 vs. 0.096 RMSE and 0.891 vs. 0.888 SRCC), while TrajSteering improves WOD pairwise accuracy from 66.8% to 70.3% and increases held-out ρ_{TCD} from 6.4 to 8.7. The permuted-target and full-state controls reach only 66.3% and 67.9% pairwise accuracy, respectively, indicating that the gain depends on the correctly conditioned criterion acting on the exact attention branch rather than on generic latent regularization.



(a) WOD-E2E pairwise accuracy (%). Higher is better.

(b) Effective score levels N_{eff} (log scale). Higher is better.

(c) Prompt sensitivity across fixed paraphrases. Lower is better.

Figure 4: Native textual scoring versus frozen learned readouts. Stronger elicitation improves the native channel but remains below the learned query interface in human alignment, score resolution, and stability.

Table 4: Human-alignment error across semantic disagreement regimes. blind spots are cases where the rules pass but the teacher substantially overestimates the human score.

Metric	EPDMS	Teacher	Prompt	Plain	LoRA	**Ours**
Overall ↓	2.95	2.78	2.92	2.31	2.02	**2.10**
Agreement ↓	1.72	1.58	2.60	1.66	1.49	**1.61**
Over-penalty ↓	4.31	3.42	3.05	2.88	2.51	**2.42**
Blind spot ↓	3.88	3.95	3.30	3.41	3.02	**2.79**
Transcend (%) ↑	–	–	49.8	52.4	56.1	**60.7**
Worst group ↓	4.31	3.95	3.30	3.41	3.02	**2.79**

Table 5: Causal interventions on the learned context-attention readout. Top-attended and random suppression remove equal attention mass; ΔAcc. is relative to the unchanged readout.

Metric	**Ours**	Uniform	Top-supp.	Rand.-supp.	Mod.-shuffle
RMSE ↓	0.095	0.151	0.139	0.101	0.118
PairAcc. (%) ↑	70.3	60.9	61.8	68.9	65.1
NDCG@3 ↑	0.91	0.85	0.86	0.90	0.88
ΔAcc. ↑	–	-9.4	-8.5	-1.4	-5.2
ρ_{TCD} ↑	8.7	1.2	2.6	7.9	4.8

5.3 Inner Comparison

The cross-method results establish whether TrajSteering is externally competitive. We next hold the backbone, rendered inputs, pseudo labels, and evaluation protocol fixed, and test three narrower claims: whether the native numeric channel remains limiting under stronger prompting controls; whether the gain concentrates in semantically difficult teacher–human disagreement regimes; and whether the learned context-attention pattern causally contributes to the score. Detailed constructions and metric definitions are provided in App. F.7.

Native scoring channel. Figures 4a–4c compare native textual scoring with frozen learned readouts. The strongest prompting control reaches 57.6% pairwise accuracy, 23.7 effective score levels, and 0.62 prompt sensitivity, compared with 66.8%, 287, and 0.11 for the plain readout and 70.3%, 341, and 0.09 for TrajSteering. Thus, stronger elicitation does not close the gap in human alignment, score resolution, or stability; the learned query interface extracts a better scalar score from the same frozen backbone.

Semantic Disagreement Regimes Overall performance can conceal whether an evaluator succeeds on the cases that motivate semantic scoring. We therefore partition WOD-E2E automatically using the fixed Scheme-B teacher score, public rule outputs, and existing human ratings. No additional annotation or manual case selection is introduced.

Table 4 shows that TrajSteering obtains the lowest MAE in both rule over-penalty (2.42) and rule-pass blind-spot (2.79) regimes, improving over the zero-shot teacher's 3.42 and 3.95 and also over the LoRA scorer's 2.51 and 3.02. Its 60.7% TranscendRate and 2.79 worst-group MAE are the best results, while its agreement-subset MAE is not the lowest, indicating that the benefit concentrates on teacher–human disagreement cases rather than merely reproducing the teacher.

Causal Readout Interventions Table 5 intervenes on the exact context-attention branch used by the filtering criterion. A single trained TrajSteering checkpoint and scalar head are held fixed, and no intervention is followed by retraining.

Table 5 shows that suppressing top-attended mass reduces pairwise accuracy by 8.5 points, compared with only 1.4 points for equal-mass random suppression; uniform attention similarly causes a 9.4-

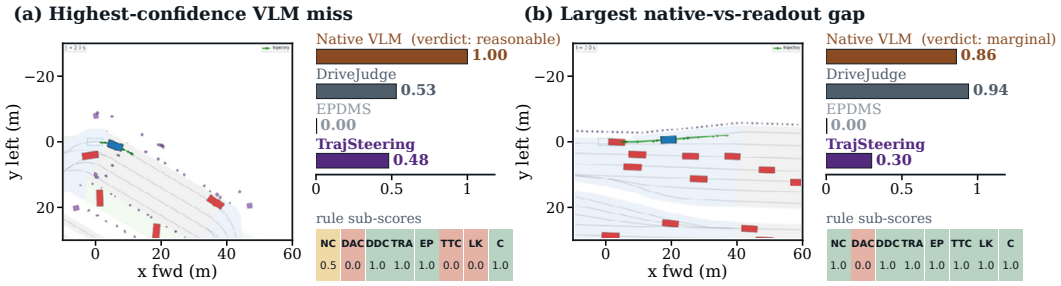


Figure 5: **Visualization of selected scenarios.** Each panel shows the BEV frame of the candidate trajectory, the scalar score from each evaluator, and the rule sub-scores shared by DriveJudge eq. (8) and EPDMS. The VLM semantic output is available to TrajSteering, but using the VLM's own scalar score directly is unreliable: in (a), the VLM describes the trajectory as safe, legal, and hard-gate compliant even though `drivable_area_compliance` (DAC) is zero. In (b), the same semantic failure appears again: the VLM assigns a high score to a real hard-rule violation, while TrajSteering gives a substantially lower score.

point drop and reduces ρ_{TCD} from 8.7 to 1.2. Shuffling attention within each modality also lowers accuracy by 5.2 points, providing causal evidence that the learned attention pattern is functionally responsible for the score rather than merely descriptive.

5.4 Visualization

We complement the aggregate metrics with two visualizations. Figure 5 shows that the native VLM can produce a plausible semantic assessment yet retain a high scalar score despite zero `drivable_area_compliance` or a hard-rule violation. TrajSteering lowers both scores, suggesting that relevant quality cues are present in the frozen model but poorly exposed by its native scoring channel.

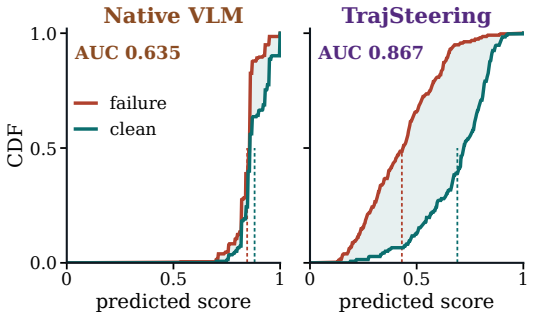


Figure 6: **Score CDFs of failure and clean scenes.** Dotted lines mark group medians.

Figure 6 confirms the same pattern at the distribution level: the native VLM's failure and clean scores overlap, whereas TrajSteering separates them more clearly and raises AUC from 0.635 to 0.867. Together, Figures 5 and 6 show that the learned readout turns latent semantic judgment into a more discriminative scalar channel. Both of which have proven the effectiveness of AttendingScorer.

6 Conclusion

We presented AttendingScorer, a framework for turning a frozen driving VLM into an end-to-end trajectory evaluator by reading out, rather than installing, its scoring ability. Our instantiation, TrajSteering, appends a small shared query bank and scalar head: automatic rule-gated scores determine what to predict, while teacher-conditioned deflection selects how the frozen attention interface should read the scene. The resulting evaluator leaves the backbone unchanged, requires no task-specific human labels, and removes rule execution from inference. More broadly, this work separates information stored in a model from information accessible through its native scoring channel: reliable evaluation can be a constrained readout problem rather than a finetuning problem. Its boundary is equally clear—the frozen representation must contain the relevant quality signal, and the query interface must be able to realize an effective readout.

AI USE STATEMENT

Generative AI tools were used only to polish the presentation of this paper, namely to edit wording, grammar, and LaTeX layout and to refine figures. They were not used for research ideation, theoretical derivations, experiment design, implementation, data analysis, or the generation of results or scientific claims. All AI-assisted edits were reviewed by the authors, who take full responsibility for the content of this paper.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 4709–4722, 2025.

Jack Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969. doi:10.1109/PROC.1969.7278.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2406.15349.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2501.11561.

Xinglong Sun, Kevin Xie, Jenny Schmalfuss, Despoina Paschalidou, Xiuming Zhang, Sanja Fidler, Kashyap Chitta, and Jose M. Alvarez. DriveJudge: Rethinking autonomous driving evaluation with vision-language models. *arXiv preprint arXiv:2606.17362*, 2026.

Qimao Chen, Fang Li, Yuechen Luo, Zehan Zhang, Haiyang Sun, Fangzhen Li, Bing Wang, Guang Chen, Yang Ji, Jiong Deng, Hongwei Xie, Hangjun Ye, Long Chen, and Yi Zhang. DriveReward: A comprehensive dataset and generative vision-language reward model for autonomous driving. *arXiv preprint arXiv:2606.08525*, 2026.

Yi-Chen Li, Tian Xu, Yang Yu, Xuqin Zhang, Xiong-Hui Chen, Zhongxiang Ling, Ningjing Chao, Lei Yuan, and Zhi-Hua Zhou. Generalist reward models: Found inside large language models. *arXiv preprint arXiv:2506.23235*, 2025.

Lloyd J. Griffiths and Charles W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982. doi:10.1109/TAP.1982.1142739.

Marcel Hallgarten, Julián Zapata, Martin Stoll, Katrin Renz, and Andreas Zell. Can vehicle motion planning generalize to realistic long-tail scenarios? In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5388–5395, 2024. doi:10.1109/IROS58592.2024.10803052.

Ruiyang Hao, Bowen Jing, Haibao Yu, and Zaiqing Nie. StyleDrive: Towards driving-style aware benchmarking of end-to-end autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4627–4635, 2026. doi:10.1609/aaai.v40i6.42463.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743, 2019. doi:10.18653/v1/D19-1275.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, pages 709–727, 2022.

Anqing Jiang, Yu Gao, Yiru Wang, Zhigang Sun, Shuo Wang, Yuwen Heng, Hao Sun, Shichen Tang, Lijuan Zhu, Jinhao Chai, Jijun Wang, Zichong Gu, Hao Jiang, and Li Sun. IRL-VLA: Training an vision-language-action policy via reward world model. *arXiv preprint arXiv:2508.06571*, 2025.

Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 629–636, 2024. doi:10.1109/ICRA57147.2024.10610077.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597, 2021. doi:10.18653/v1/2021.acl-long.353.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742, 2023.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via BEV world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. arXiv:2504.01941.

Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and Jose M. Alvarez. Generalized trajectory scoring for end-to-end multimodal planning. *arXiv preprint arXiv:2506.06664*, 2025.

Bin Sun, Yaoguang Cao, Yan Wang, Rui Wang, Jiachen Shang, Xiejie Feng, Jiayi Lu, Jia Shi, Shichun Yang, Xiaoyu Yan, and Ziying Song. MindDrive: An all-in-one framework bridging world models and vision-language model for end-to-end autonomous driving. *arXiv preprint arXiv:2512.04441*, 2025.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6562–6595, 2024.

Haoran Lou, Ziyan Liu, Chunxiao Fan, Yuexin Wu, Yue Ming, Hao Wu, Kai Zuo, Yibo Chen, and Xu Tang. SLQ: Bridging modalities via shared latent queries for retrieval with frozen MLLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2604.13710.

Jingyu Song, Zhenxin Li, Shiyi Lan, Xinglong Sun, Nadine Chang, Maying Shen, Joshua Chen, Katherine A. Skinner, and Jose M. Alvarez. DriveCritic: Towards context-aware, human-aligned evaluation for autonomous driving with vision-language models. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2026. arXiv:2510.13108.

Jizhou Guo, Zhaomin Wu, Hanchen Yang, and Philip S. Yu. Mining intrinsic rewards from LLM hidden states for efficient best-of-N sampling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi:10.1145/3770854.3780302.

Harry L. Van Trees. *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. Wiley-Interscience, 2002. doi:10.1002/0471221104.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54015–54029, 2024.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. LLaVA-Critic: Learning to evaluate multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2410.02712.

Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Ekaterina Tolstaya, Sarah Tang, Brandyn White, Ben Sapp, Mingxing Tan, Jyh-Jing Hwang, and Dragomir Anguelov. WOD-E2E: Waymo Open Dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. doi:10.1007/s11263-022-01653-1.

APPENDIX

A SUMMARY OF APPENDIX

We include the following supplementary materials to provide additional details on the formulation, experimental protocol, validation criteria, and scope of AttendingScorer.

- B **Backbone provenance.** Defines the frozen-backbone boundary and rules out prior trajectory-scoring or teacher-specific adaptation.
- C **Preliminary study and filtering reference.** Details the motivating observations, falsification conditions, classical deflection reference, and realizability argument.
- D **More method details.** Discusses estimator approximations, deployment cost, and when constrained filtering constitutes an algorithmic contribution rather than only an interpretation.
- E **Hyperparameters.** Records the roles and sweeps of the principal optimization and estimator hyperparameters.
- F **More experiment settings.** Specifies the pseudo-label teacher, datasets, rendering, baselines, metrics, matched-fit tests, and statistical protocol.
- G **Implementation details.** Lists the readout configuration, teacher pipeline, evaluation protocol, and compute information required for reproducibility.
- H **Structurally limited teacher stress test.** Tests whether the learned readout can improve on its automatic teacher in teacher–human disagreement regimes.

B BACKBONE PROVENANCE

The claim that our evaluator performs no backbone weight update refers to the readout stage: every parameter of Φ is fixed from the moment the readout is introduced, and only the appended embeddings and the scalar head are trained. We reuse a published driving backbone so that the comparison isolates the adaptation mechanism rather than the model, and the distinction this rests on is sharper in v7 than before, because there are now two ways the claim could become circular. The checkpoint may have been domain-adapted for scene understanding — driving question answering, scene description, object and intent recognition — but must not have been trained on trajectory scoring or preference supervision, and must not have been distilled from the same rule-gated teacher we use for supervision.

C PRELIMINARY STUDY DETAILS

Protocol. All panels are measured on the same frozen backbone, over scenario–trajectory pairs drawn from the public hard-case split and from preference pairs where rule scores and expert judgment disagree (Cao et al., 2025; Song et al., 2026); no component of our method is involved in (a) and (b). Panel (c) is the one panel that uses our own training loop, and it is stated as such: it trains only the plain readout variant ($\lambda = 0$) and reports the spread across seeds, so that it measures underdetermination rather than the merit of our method. Setup — backbone, split, sample count, prompt text, decoding configuration, probe capacity, seeds, and repeated samples per item — is listed here so that every panel is reproducible.

Binding falsification conditions. *(a)* If the frozen probe does not land clearly above both the rule score and direct prompting, the representation premise fails and the framing must be abandoned rather than repaired. *(b)* If prompt paraphrase or a change of decoding configuration closes the gap to the probe, observation (ii) must be narrowed to the deterministic decoding configuration used for evaluation, and the channel argument weakened accordingly. *(c)* If matched-fit readouts do not diverge on held-out and long-tail data, the selection principle of Sec. 4.3 loses its motivation and must be re-justified as a regularizer; if they diverge but held-out ρ_{TCD} does not track which one generalizes, the criterion is refuted before training and Eq. equation 10 is withdrawn.

Ruling out trivial explanations. Each panel is accompanied by checks required to exclude alternative causes: whether the effect survives a second backbone; whether it survives paraphrased prompts; whether it is an artefact of the split; whether it disappears under repeated sampling; whether a simple heuristic on the rendered geometry reproduces it; and, for (c), whether the divergence is explained by teacher-fit differences too small to see rather than by readout choice.

C.1 Classical deflection reference and realizability

Free-tap optimum. Let $\theta \in \mathbb{R}^m$ be a signal template and let $\Sigma \in \mathbb{R}^{m\times m}$ be positive definite. For any nonzero tap vector α, define

$$\rho_{\text{free}}(\alpha) = \frac{(\alpha^\top \theta)^2}{\alpha^\top \Sigma \alpha}. \tag{11}$$

Set

$$b = \Sigma^{1/2}\alpha, \qquad q = \Sigma^{-1/2}\theta.$$

Then

$$\rho_{\text{free}}(\alpha) = \frac{(b^\top q)^2}{b^\top b} \le q^\top q = \theta^\top \Sigma^{-1}\theta,$$

where the inequality follows from Cauchy–Schwarz. Equality holds exactly when $b \propto q$, equivalently

$$\alpha^\star_{\text{free}} \propto \Sigma^{-1}\theta, \qquad \max_{\alpha\neq 0} \rho_{\text{free}}(\alpha) = \theta^\top \Sigma^{-1}\theta. \tag{12}$$

Thus the classical solution whitens the nuisance covariance before correlating with the signal. When Σ contains cross-position correlations, the optimum need not concentrate on a single position; it may combine multiple observations to suppress correlated nuisance. The criterion is invariant to global rescaling of α, and the analogous scale invariance of the learned latent criterion is recorded in Remark 1.

Why this is only a reference for AttendingScorer. Equations equation 11–equation 12 assume that the signal template and nuisance covariance are known in the coordinates in which the taps are chosen, and that the taps may be assigned freely. None of these assumptions is made by TrajSteering: the frozen VLM is nonlinear, the corresponding statistics are estimated only after the realizable attention aggregation, and $\alpha_\Phi(E; x, \tau)$ is generated by the frozen network from the shared query bank. Therefore $\alpha^\star_{\text{free}}$ defines an unconstrained reference, not a directly executable update.

Proof of Proposition 1. For every $E \in \mathcal{E}$ and every fixed (x, τ), Eq. equation 5 defines $\alpha_\Phi(E; x, \tau)$ as an element of $\mathcal{A}_\Phi(x, \tau)$; hence every attention pattern reached by optimizing E is feasible by construction. Equation equation 5 contains one decision variable E outside the expectation over $(x, \tau) \sim \mathcal{D}$, so the same learned query bank is applied to every example rather than being re-optimized per instance. Finally, an unconstrained filter solution can be implemented by this frozen interface only if it belongs to the image $\mathcal{A}_\Phi(x, \tau)$ on the relevant inputs. □

D More Method Details

Remark 1 (Scale invariance and its consequence)**.** *Both numerator and denominator of equation 9 are quadratic in z_E, so the criterion is invariant to a global rescaling of the filtered latent and cannot be satisfied by inflating hidden norms. What it can still be satisfied by is generic variance shaping that ignores the target; the label-permutation control of Sec. 5.2 is the test that separates the two, and it is a required, not optional, experiment.*

Stated approximations, with their account. Three choices are computational rather than theoretical and are ablated as such: the diagonal nuisance covariance in line 5 of Alg. 2, against shrinkage, full-covariance and label-bin variants; the minibatch estimator, against exponential moving averages; and the ridge ε, whose only role is numerical. The criterion is also batch-dependent by construction, so its estimator variance is reported alongside its value rather than assumed away.

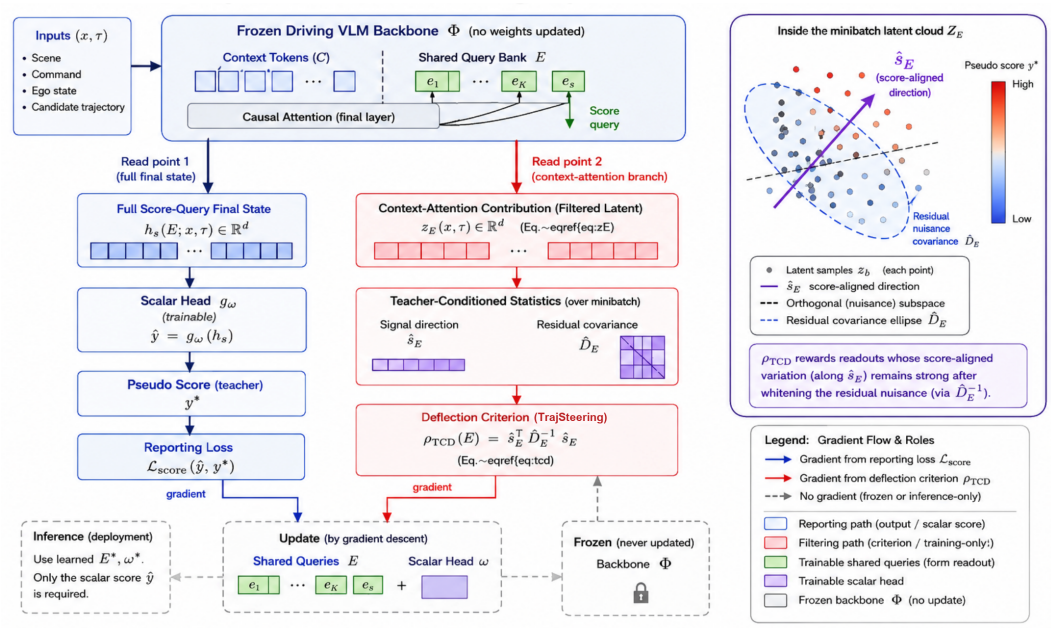


Figure 7: Visualized Method Workflow

D.1 Cost and Deployment

Teacher scores are generated once and stored, so no teacher call occurs during student optimization. The original multimodal context is frozen and causally precedes the readout queries, so its key/value states may be cached; training adds only the incremental query computation and gradients with respect to (E, ω). For hidden width d and $K+1$ queries with a linear head, the trainable count is approximately $(K+1)d + d + 1$, and the added attention cost per layer is $O\big((K+1)n + (K+1)^2\big)$ rather than recomputing or differentiating all context–context attention. At inference the filter statistic is dropped entirely. Relative to the design this paper replaces, three costs also disappear: there is no generation pass before scoring, no second forward pass for calibration, and no human preference pipeline to maintain per evaluation dimension.

D.2 Why Filtering is an Algorithmic Contribution Rather than an Analogy

The filtering view earns its place only if it changes the learned method beyond plain pseudo-label distillation. The two terms of Eq. equation 10 have different objects: $\mathcal{L}_{\mathrm{score}}$ constrains the endpoint and can be minimised by any readout–head pair that reproduces the teacher scores, whereas ρ_{TCD} acts on the internal context-attention output before the head and rewards a readout whose teacher-aligned variation survives nuisance whitening. The criterion is therefore a *selection principle over output-compatible solutions*. It does not guarantee human alignment, does not make attention uniquely identifiable, and does not claim that the realized taps are optimal in the classical sense. Its falsifiable prediction is narrower and is the one the experiments test: at matched teacher-score fit, adding Eq. equation 9 should increase held-out deflection *and* held-out scoring performance, most visibly under distribution shift, where readouts that fit the target through shortcut positions are least stable. If deflection rises without any held-out benefit, or if the same benefit appears under permuted labels, constrained filtering remains an interpretive lens and we say so.

Registered prediction. At matched teacher-score fit, query count, head capacity, optimization budget, and frozen checkpoint, adding ρ_{TCD} should

1. increase held-out teacher-conditioned deflection; and
2. improve held-out trajectory scoring, with the largest gain under distribution shift, where shortcut-based readouts should be least stable.

Table 6: Hyperparameter ledger.

Symbol	Role	Swept?	Bound to claim	Where
λ	criterion weight; λ=0 = plain readout	yes	filtering verdict	Tab. 3
K	capacity; $(K+1)d+d+1$ trainable params	yes	capacity sweep	–
B	estimator sample size (Alg. 2 moments)	yes	d/B bias	App. C
τ_{gate}	teacher label distribution, $\mathrm{Var}(y^\star)$	yes	teacher-strength axis	App. F
$\varepsilon_s, \varepsilon_D$	numerical stabilizers	no	–	–
EMA momentum	moment-estimate variant	ablated	–	–
δ_{Huber}; lr; schedule	training details	no	–	–

The decisive comparison is therefore

$$\text{plain query readout} \;<\; \text{TrajSteering} \tag{13}$$

under the matched protocol of Sec. 5.2.

This prediction is withdrawn if deflection rises without an external held-out benefit, or if comparable gains appear when the pseudo scores used to construct the criterion are permuted. In either case, constrained filtering remains an interpretive formulation rather than an established algorithmic contribution.

E Hyperparameters

See table. E

F More Experiment Settings

This appendix records the details omitted from Sec. 5.1: backbone provenance, Scheme-B pseudo-label generation, data construction, input rendering, baseline implementation, metric definitions, matched-fit criteria, and statistical reporting.

F.1 Backbone and Trainable Boundary

All VLM-based experiments start from the same Qwen2.5-VL-7B base checkpoint used by DriveCritic (Song et al., 2026), before its task-specific trajectory-evaluation SFT and reinforcement-learning stages. We record the exact model repository, checkpoint revision, tokenizer revision, multimodal processor revision, image resolution, and temporal sampling configuration.

For direct prompting, fixed-query readout, plain query readout, and TrajSteering, every parameter of the vision encoder, multimodal projector, language model, attention projections, and original embedding and output layers is frozen. Only the appended query embeddings E and scalar head g_ω are trainable. The LoRA baseline starts from the same checkpoint and input pipeline but updates only its declared low-rank adapters and scoring head.

We verify the frozen boundary by comparing the original context states before and after appending the readout queries. Because the queries are appended after the context under causal masking, the original context states must be bit-identical.

F.2 Scheme-B Automatic Pseudo Labels

Zero-shot rule-gated teacher. Pseudo labels are generated by a separately instantiated public Qwen-VL checkpoint used without task-specific finetuning. Given a scene x_i and candidate trajectory τ_i, the teacher receives a fixed DriveJudge-style rule-applicability prompt (Sun et al., 2026). It returns one label for each criterion in the fixed ordered set

$$\mathcal{M} = \{\mathrm{LK}, \mathrm{DDC}, \mathrm{DAC}, \mathrm{TTC}, \mathrm{NC}, \mathrm{EP}, \mathrm{Comfort}\}.$$

Each label must be exactly `Useful` or `Questionable`. The result is converted to a binary gate,

$$\hat{g}_i^m = \mathbf{1}\left[a_i^m = \texttt{Useful}\right], \qquad m \in \mathcal{M}.$$

The teacher is asked only whether a criterion is applicable in the current context. It does not estimate geometry-dependent rule values and does not directly output the scalar pseudo score.

Rule execution and aggregation. For every criterion, the public NAVSIM/Pseudo-Simulation implementation computes

$$s_i^m = f_{\mathrm{EPDMS}}^m(x_i, \tau_i), \qquad s_i^m \in [0,1].$$

Following the rule-gated aggregation form, we partition the metrics into

$$\mathcal{M}_{\mathrm{pen}} = \{\mathrm{NC}, \mathrm{DAC}, \mathrm{DDC}\}, \qquad \mathcal{M}_{\mathrm{avg}} = \mathcal{M} \setminus \mathcal{M}_{\mathrm{pen}},$$

and define the automatic target as

$$y_i^\star = \left[\prod_{m \in \mathcal{M}_{\mathrm{pen}}} (\hat{g}_i^m s_i^m + 1 - \hat{g}_i^m)\right] \frac{\sum_{m \in \mathcal{M}_{\mathrm{avg}}} \hat{g}_i^m w^m s_i^m}{\sum_{m \in \mathcal{M}_{\mathrm{avg}}} \hat{g}_i^m w^m}, \qquad y_i^\star \in [0,1],$$

where w^m and the rule definitions are inherited from the public metric implementation and are not tuned using WOD-E2E ratings.

F.3 Data Splits and Input Rendering

Pseudo-score data. Training and pseudo-validation examples are constructed from public NAVSIM and Pseudo-Simulation scene–trajectory pools (Dauner et al., 2024; Cao et al., 2025). Splits are formed at the driving-log or scene level. Alternative trajectories, neighboring frames, and synthetic variants derived from the same scene remain in the same split.

F.4 Baseline Implementation

Imitation score. The imitation baseline ranks trajectories by negative ADE/FDE to the logged future. It uses no learned parameters and no semantic scene representation.

Trajectory-only scorer. The trajectory-only MLP receives trajectory and ego-motion features but no camera or language input. It is trained on exactly the same pseudo-score examples as the VLM-based learned scorers.

Direct prompting. The frozen backbone receives the same multimodal input as TrajSteering and a fixed instruction requesting one scalar quality score in $[0,1]$. Decoding is deterministic. Invalid outputs are handled using the pre-registered automatic parsing policy and are never manually repaired.

Plain query readout. This model uses the same query bank and scalar-head architecture as TrajSteering, but sets $\lambda = 0$ and is trained only with the pseudo-score loss.

Fixed/random-query control. The query embeddings are initialized using the same procedure as the learned readout and remain frozen. Only the scalar head is optimized.

Permuted-target control. The endpoint score loss continues to use the correct $y^\star$. Only the pseudo scores used to estimate the teacher-conditioned deflection direction are randomly permuted within each training batch.

Full-state criterion control. This variant applies the same deflection calculation to the complete score-query state h_s rather than to the exact context-attention output z_E. All other components remain unchanged.

LoRA scorer. The LoRA baseline starts from the identical backbone and uses the same pseudo-labelled examples, rendering, scalar target, optimizer family, and number of updates. Its rank and adapter placement are selected using only the pseudo-validation split.

F.5 Metric Definitions

Pairwise accuracy. For WOD-E2E scene i, let $y_{i,a}^{\text{human}}$ and $y_{i,b}^{\text{human}}$ be the human scores of two candidates. We evaluate only pairs with unequal human ratings:

$$\mathcal{P} = \left\{(i,a,b) : y_{i,a}^{\text{human}} \neq y_{i,b}^{\text{human}}\right\}.$$

Pairwise accuracy is

$$\text{PairAcc} = \frac{1}{|\mathcal{P}|} \sum_{(i,a,b)\in\mathcal{P}} \mathbf{1}\left[(\hat{y}_{i,a} - \hat{y}_{i,b})\left(y_{i,a}^{\text{human}} - y_{i,b}^{\text{human}}\right) > 0\right].$$

Within-scene ranking. Kendall's τ_b is computed over the three candidates in each scene and macro-averaged across scenes; τ_b accounts for ties in the human ratings. NDCG@3 uses the three 0–10 human ratings directly as graded relevance and is also macro-averaged across scenes.

Fixed-scale MAE. Because the student is trained on a $[0,1]$ teacher scale, we evaluate absolute scoring without fitting a transformation on WOD-E2E:

$$\text{MAE}_{0\text{–}10} = \frac{1}{N}\sum_{i=1}^{N}\left|10\hat{y}_i - y_i^{\text{human}}\right|.$$

No affine, isotonic, or nonlinear calibration function is estimated from WOD-E2E labels.

Pseudo-score fit. On the held-out automatic pseudo-validation split, we report RMSE and Spearman rank correlation coefficient against $y^\star$. Held-out ρ_{TCD} is always recomputed using the correct, unpermuted pseudo scores and the attention-filtered context branch z_E, including for controls trained with a different criterion target or location.

F.6 Matched-Fit and Statistical Protocol

Plain query readout and TrajSteering use the same:

- pseudo-labelled training and validation examples;
- query count and query initialization;
- scalar-head architecture and initialization;
- optimizer, learning-rate schedule, batch size, and number of updates;
- checkpoint-selection rule and random seeds.

Their only algorithmic difference is the $-\lambda\rho_{\text{TCD}}$ term. We declare pseudo-score fit to be matched when

$$|\text{RMSE}_{\text{ours}} - \text{RMSE}_{\text{plain}}| \leq \epsilon_{\text{RMSE}}$$

and

$$|\text{SRCC}_{\text{ours}} - \text{SRCC}_{\text{plain}}| \leq \epsilon_{\text{SRCC}}.$$

All learned methods are run with five random seeds. We report the mean and a 95% confidence interval across seeds. For WOD-E2E comparisons, paired bootstrap resampling is performed over complete driving segments so that the three trajectories from one segment are never treated as independent bootstrap units. Pairwise accuracy and ranking metrics are recomputed within each bootstrap sample.

F.7 Additional Inner-Comparison Protocols

F.7.1 Native-Channel Controls

All native-channel controls use the same frozen checkpoint, multimodal input, trajectory rendering, and scoring rubric as the learned readouts. Direct scalar prompting requests one score in $[0, 1]$. Analyze-then-score prompting first requests a concise assessment and then a scalar score. The prompt-paraphrase ensemble averages predictions from P fixed, semantically equivalent prompts. Self-consistency averages M stochastic samples from one fixed prompt. The values of P, M, temperature, and all prompt texts are fixed using only the automatic pseudo-validation split.

Scores are placed into a fixed set of bins with empirical probabilities $\{p_b\}_{b=1}^{B_s}$. Score diversity is measured by the effective number of levels

$$N_{\text{eff}} = \exp\left(-\sum_{b=1}^{B_s} p_b \log p_b\right).$$

For P fixed prompt paraphrases, prompt sensitivity is

$$\text{Sens}_{\text{prompt}} = \frac{1}{N}\sum_{i=1}^{N} \text{Std}_{p\in\{1,\dots,P\}}\left[\hat{y}_i^{(p)}\right].$$

Format-valid rate is the fraction of examples for which a scalar in the specified range is recovered without manual correction. Invalid responses receive the pre-registered automatic fallback and are not manually repaired.

Pairwise prompting is reported separately in the appendix because it does not produce an absolute scalar score. Candidate order is reversed to measure order-flip robustness, and the prompt receives the same candidate renderings used by all pointwise scorers.

F.7.2 Automatic Disagreement-Regime Construction

Let

$$\tilde{y}_i^\star = 10y_i^\star, \qquad \tilde{y}_i = 10\hat{y}_i,$$

denote the automatic teacher score and model score on the WOD-E2E human 0–10 scale. Given a fixed disagreement margin δ, we define

$$\mathcal{D}_{\text{agree}} = \left\{i : \left|\tilde{y}_i^\star - y_i^{\text{human}}\right| \le \delta\right\},$$

$$\mathcal{D}_{\text{over}} = \left\{i : y_i^{\text{human}} - \tilde{y}_i^\star > \delta\right\},$$

where $\mathcal{D}_{\text{over}}$ contains trajectories that the automatic teacher over-penalizes relative to the human score.

The strict rule-pass blind-spot subset is

$$\mathcal{D}_{\text{blind}} = \left\{i : \tilde{y}_i^\star - y_i^{\text{human}} > \delta, \quad s_i^m \ge \tau_m \;\forall m \in \mathcal{M}\right\},$$

where s_i^m is the public rule score and τ_m is its pre-registered pass threshold. Thus, every represented rule passes, while the existing human score still indicates poor trajectory quality.

The complete disagreement set is

$$\mathcal{D}_{\text{dis}} = \left\{i : \left|\tilde{y}_i^\star - y_i^{\text{human}}\right| > \delta\right\}.$$

We report

$$\text{TranscendRate} = \frac{1}{|\mathcal{D}_{\text{dis}}|}\sum_{i\in\mathcal{D}_{\text{dis}}} \mathbf{1}\left[\left|\tilde{y}_i - y_i^{\text{human}}\right| < \left|\tilde{y}_i^\star - y_i^{\text{human}}\right|\right].$$

Worst-group MAE is the maximum MAE over $\mathcal{D}_{\text{agree}}$, $\mathcal{D}_{\text{over}}$, and $\mathcal{D}_{\text{blind}}$. The margin δ and all rule-pass thresholds are frozen before WOD-E2E results are inspected.

F.7.3 ATTENTION-INTERVENTION CONSTRUCTION

All interventions operate on the score query's normalized context-attention weights α_E while keeping the trained queries, scalar head, residual paths, and all backbone parameters fixed.

Uniformization. For n context positions, replace

$$\alpha_{E,i} \leftarrow \frac{1}{n}.$$

Top-mass suppression. Let $\mathcal{S}_{\text{top}}$ be the smallest set of highest-weight positions whose total attention mass reaches a fixed fraction q:

$$\sum_{i \in \mathcal{S}_{\text{top}}} \alpha_{E,i} \geq q.$$

Set their weights to zero and renormalize the remaining weights.

Equal-mass random suppression. Randomly select context positions until their original attention mass matches that of $\mathcal{S}_{\text{top}}$, suppress them, and renormalize. Results are averaged over R random draws for each example. This equal-mass construction, rather than equal token count, ensures that the intervention strength is matched.

Within-modality shuffling. Permute attention weights only among positions belonging to the same modality, preserving the total mass assigned to image, trajectory, ego-state, and text tokens while destroying position-specific selection.

The mass fraction q, number of random draws R, intervention layer, and head-aggregation rule are fixed using only pseudo-validation data.

G IMPLEMENTATION DETAILS

Input rendering. How the scenario context and candidate trajectory are rendered into the backbone's input format; overlay conventions during teacher labelling versus clean inputs at training time; and the fixed evaluation instruction.

Readout configuration. Query count K, embedding initialisation, scalar head architecture, optimizer, schedule, λ, ridge ε, batch size (which is also the estimator's sample size for lines 3–5 of Alg. 2), and the exact count of trainable parameters.

Teacher pipeline. Pool construction, the teacher's configuration, the one-time offline labelling pass, storage format, and the audit of teacher score distribution over the pool (including how many items fall in the regime where the teacher is structurally unable to penalise, which defines $\mathcal{D}_{\text{dis}}$ candidates in Sec. H).

Evaluation protocol. Splits, seeds, number of runs, and the matched-protocol statement for each literature baseline.

H SECONDARY STRESS TEST: LEARNING FROM A STRUCTURALLY LIMITED TEACHER

The rule-gated teacher is structurally limited in exactly the regime this paper cares about: a gate that selects which rules apply can remove a penalty but cannot introduce one that no rule expresses, so on scenes where every rule check passes and humans still object, the teacher cannot be right. Pseudo-label training does not by itself prevent the student from inheriting that bias, and a frozen backbone with a small readout offers no guarantee against it. Our narrower hypothesis is that the teacher's labels underdetermine which frozen features are used to reproduce them — the same underdetermination that motivates Sec. 4.3 — while the deflection criterion biases that choice toward

stable score-aligned structure already represented by the backbone. Whether that suffices to cross the teacher's blind spot is a question, not a premise.

We test it after training, on held-out human-labelled items where the teacher and humans disagree, and compare the student to both:

$$\mathrm{TranscendRate} = \frac{1}{|\mathcal{D}_{\mathrm{dis}}|} \sum_{i \in \mathcal{D}_{\mathrm{dis}}} \mathbf{1}\big[\, |\hat{y}_i - y_i^{\mathrm{human}}| < |y_i^{\star} - y_i^{\mathrm{human}}| \,\big]. \tag{14}$$

Disagreement regimes are identified after training; exact evaluation items never enter training, although the same scenario types may; and no human label is used for training or model selection anywhere in this test. A rate above the plain-readout and calibration baselines supports weak-to-strong behaviour. Full inheritance does not invalidate constrained filtering: it limits the paper to honest pseudo-label distillation with a frozen readout, which is the downgrade we accept in advance rather than argue about afterwards.