CHAIN-OF-ATTENTION: FIELD-VERIFIABLE REASONING FOR TRAFFIC VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Traffic vision–language models can produce plausible answers without revealing when an event occurs, which actors are involved, or what visual evidence supports the explanation. We introduce Chain-of-Attention (COA), a phase-indexed record that binds five traffic phases to selected frames, pedestrian and vehicle boxes, and an event-level account. These explicit fields support separate training, scoring, reward optimization, and stress testing. On 48 fixed-camera WTS validation scenes, supervised training raises pedestrian IoU from .014 to .406 and vehicle IoU from .035 to .570 under the field-level evaluation with uniform sampling and no phase labels. Reward optimization preserves this grounding and improves blinded ratings of visual factuality, causal support, and evidence consistency. A phase-free control shows that supervision balance strongly affects reliable structured output. Balancing full-record and auxiliary examples raises strict schema completion from 3/48 to 15/48; COA reaches 16/48 under its corresponding strict schema. This experiment identifies task mixture as a major source of format failure, while COA provides an explicit interface for training and auditing evidence-bound predictions. We also expose a temporal shortcut: equal-per-phase sampling makes phase identity recoverable from input rank. Replacing images across 47 paired scenes collapses actor IoU while leaving the temporal score essentially unchanged, showing that temporal structure alone does not establish visual grounding.

1 INTRODUCTION

Pedestrian–vehicle conflicts unfold through a short sequence: a road user approaches, recognizes another actor, chooses an action, and either avoids or enters a collision. A useful model must therefore say *when* risk changes, *who and where* the relevant actors are, and *why* the visible sequence supports its conclusion. A fluent answer alone does not reveal whether these commitments agree with one another or with the images.

This problem has two forms. First, one output field may be correct while the joint record is incomplete: a model can answer questions yet fail to serialize the frames and actors needed to check those answers. Second, an aggregate temporal score may reward input structure rather than visual evidence. Both failures concern *evidence binding*: whether temporal, spatial, and explanatory claims form a complete, image-bound record.

We introduce Chain-of-Attention (COA), a structured interface for five-phase traffic interactions. Each phase names a displayed frame and pedestrian and vehicle boxes; an event-level account connects the phases to the outcome. Named fields make errors local and testable. We use supervised fine-tuning (SFT) to learn the record, then group-relative policy optimization (GRPO) with an additive field-based reward to refine it. An additive advantage decomposition identifies which fields supply within-group preference signal, without claiming separate gradients or causal attribution.

Four complementary evaluations test this account. First, matched automatic metrics separate the large Base-to-SFT grounding gain from GRPO's smaller automatic changes, and blinded ratings test GRPO's role in explanation quality. Second, a phase-free Flat target preserves the historical inputs and atomic supervision. Balancing increases strict schema completion from 3/48 to 15/48, compared with 16/48 for COA, while field recovery identifies the remaining omissions. Third, a three-seed

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.

reward intervention tests whether direct IoU terms change the intended spatial fields. Fourth, rank, order, cross-scene shuffling, and cross-backbone probes test whether temporal scores require visual evidence.

Our contributions are:

- We introduce COA, a phase-indexed record that binds selected frames, actor boxes, and an event account into separately trainable and auditable fields. Under uniform sampling without phase labels, SFT provides the main spatial gain, while GRPO preserves grounding and improves explanation quality in blinded evaluation.
- We use phase-free Flat controls to separate target organization from Phase-B task sampling. Balancing raises strict schema completion from 3/48 to 15/48, showing that task mixture drives much of the initial format collapse.
- We formalize the exact rank shortcut created by equal-per-phase sampling and use order, cross-scene replacement, and cross-backbone probes to separate temporal structure from visual grounding. Under cross-scene replacement, actor IoU collapses while the GRPO temporal score changes by only .001.

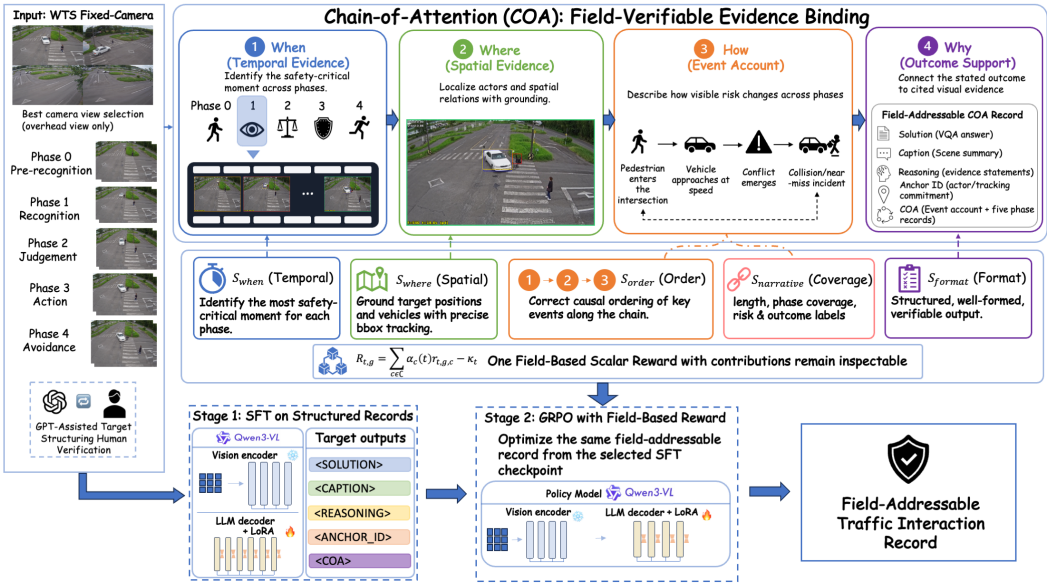


Figure 1: **Conceptual overview of COA training and auditing.** The left column shows WTS phase annotations used to construct training targets. Headline evaluation instead uses ten uniformly sampled frames with identifiers and questions but no phase labels. COA links selected moments, actor boxes, and an event account to named fields for training, scoring, and reward optimization. The labels summarize the design; Section 3 defines the exact output contract and scalar reward. GPT-assisted structuring is used only during target construction.

2 Related Work

Traffic interaction understanding. Pedestrian-safety research spans detection, trajectory and intention prediction, and interaction analysis (Ismail et al., 2009; Rasouli et al., 2019; Kotseruba et al., 2021). WTS adds synchronized roadway video, five interaction phases, boxes, captions, and question-answer labels (Kong et al., 2024); traffic VLMs study captioning and broader scene understanding (Dinh et al., 2024; Wang et al., 2024b). These tasks provide rich labels, but do not by themselves require one output to bind phase, frame, actors, and explanation.

Grounded and structured multimodal reasoning. Video captioning and QA benchmarks test event description and temporal reasoning (Krishna et al., 2017; Xiao et al., 2021; Mangalam et al.,

2023). Video VLMs extend image instruction tuning to multiple frames (Liu et al., 2023; Maaz et al., 2024), and Grounded-VideoLLM adds explicit timestamps (Wang et al., 2024a). Chain-of-thought exposes intermediate text (Wei et al., 2022), while verifiable-reward methods optimize rule-based outcomes (Shao et al., 2024; DeepSeek-AI, 2025; Shen et al., 2025; Zhang et al., 2025). COA instead makes frame identity and actor localization part of a single output contract, so intermediate claims can be checked against pixels.

Shortcut-aware evaluation. Language-only, moment-prior, single-frame, and atemporal controls expose shortcuts in video-language benchmarks (Otani et al., 2020; Yuan et al., 2021; Buch et al., 2022; Lei et al., 2023). Such effects are part of the broader shortcut-learning problem (Geirhos et al., 2020; Agrawal et al., 2018; Lapuschkin et al., 2019). Our artifact arises from the sampler: drawing the same number of frames from every labeled phase makes phase identity an exact function of input rank. We remove this code without changing the data split, then intervene separately on order and pixels.

3 Problem Setting and Chain-of-Attention

3.1 Why a field-verifiable record?

For an event, the model receives ten displayed overhead frames with identifiers and multiple-choice questions. The standard evaluation prompt does not reveal ground-truth phase labels. A useful output must answer the questions and link the visible interaction to five WTS phases: pre-recognition, recognition, judgment, action, and avoidance.

COA represents phase p by $z_p = (i_p, b_p^{\text{ped}}, b_p^{\text{veh}})$, where i_p is one displayed image identifier and each b is a normalized $xyxy$ box. The evaluation-time contract has five ordered blocks: answer JSON (`SOLUTION`), a short scene summary (`CAPTION`), two reasoning steps (`REASONING`), an actor and tracking commitment (`ANCHOR_ID`), and `COA`. The last block contains a risk/outcome account followed by exactly five ordered phase records. In the automatic evaluator, phase 4 is the model-selected critical or avoidance frame: anchor IoU scores both phase-4 boxes, whereas AFA@15 asks whether its identifier lies within 15 frames of the annotated avoidance anchor. This differs from an ordinary answer or caption because temporal and spatial claims are explicit and separately parseable. A wrong frame, actor, box, order, or outcome can be identified instead of hidden inside one fluent paragraph. We call the record *field-verifiable*: identifiers, boxes, answers, and syntax have automatic checks, while narrative factuality and causal support require rubric-based human judgment. The interface exposes commitments for checking; it does not certify that the narrative is true. Figure 1 summarizes the design. Figure 2 instantiates the field bindings and previews why phase-aware input order requires a shortcut audit.

3.2 Learning and auditing the same fields

We initialize Qwen3-VL-8B-Instruct (Bai et al., 2025) with LoRA (Hu et al., 2022). SFT first emphasizes detection spans, then trains the complete assistant record. At update t, GRPO starts from the selected SFT checkpoint and scores completion g with

$$R_{t,g} = \sum_{c \in \mathcal{C}} \alpha_c(t) r_{t,g,c} - \kappa_t, \tag{1}$$

where $\mathcal{C}$ contains every completion-dependent term in the implementation: temporal membership, pedestrian and vehicle localization, order, narrative coverage, format, VQA, zero-box and copy penalties, and a tracking bonus. Coefficients $\alpha_c(t)$ include the signed weights and spatial curriculum. The additional logged penalty is $\kappa_t = .02 \max(0, \widehat{\mathrm{KL}}_{t-1} - 2)$, computed from the previous trainer log and therefore shared by the current reward group. The reward remains one scalar; field separation makes its sources inspectable.

Let bars denote means within a rollout group and let $s_t > 0$ be its shared normalization scale.

Observation 1 (Additive advantage decomposition)**.**

$$A_{t,g} = \frac{R_{t,g} - \bar{R}_t}{s_t} = \sum_{c \in \mathcal{C}} \frac{\alpha_c(t)(r_{t,g,c} - \bar{r}_{t,c})}{s_t}. \tag{2}$$

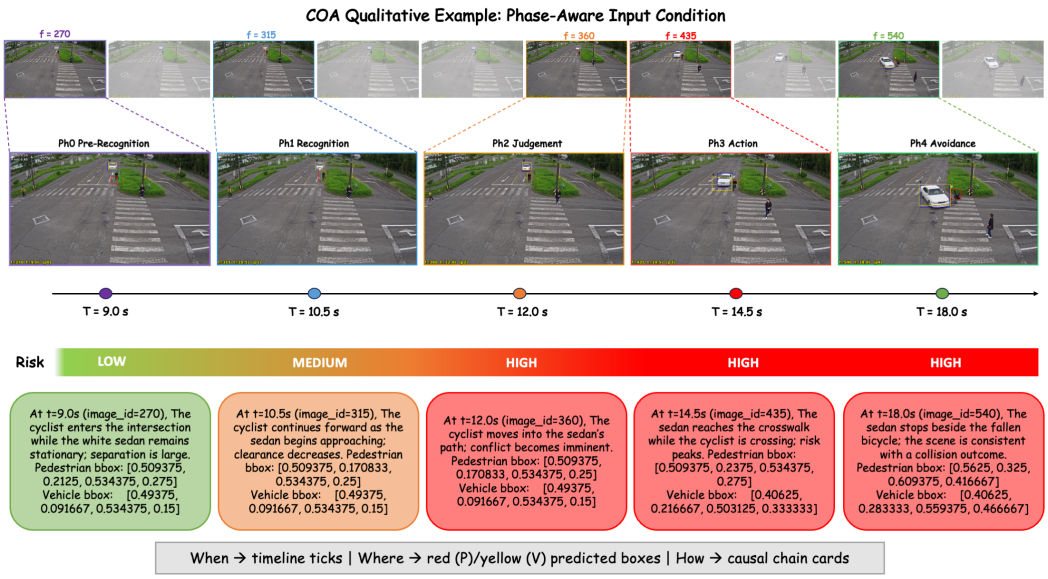


Figure 2: **COA field binding under the phase-aware input condition.** The top strip shows the equal-per-phase ordered input, whose rank structure makes phase identity recoverable without pixels. The selected frames below bind temporal position (*when*), predicted actor boxes (*where*), and a phase-wise event account (*how*). Frames and boxes come from a saved model output; the prose cards come from a separate prompt to the same checkpoint and serve as visualization. Headline evaluation uses uniformly sampled frames without phase labels. Red and yellow denote predicted pedestrian-category and vehicle boxes, respectively.

The group-shared κ_t cancels, as does any component constant within the group. Separately, the saved trainer configuration applies a per-token reference-model KL term with coefficient .001 in the policy loss; it is outside Equations 1–2. The observation is diagnostic bookkeeping for the scalar advantage, not an independent gradient or a causal effect of a reward term.

3.3 An Exact Rank Shortcut

Suppose a phase-aware sampler draws exactly k frames from each of N_ϕ ground-truth phases and concatenates them in phase order. If J is the one-indexed input rank and $\Phi(J)$ is the zero-indexed phase label, then:

Observation 2 (Rank identifiability)**.** *The phase label is* $\Phi(J) = \lfloor (J-1)/k \rfloor$ *for* $1 \le J \le kN_\phi$. *Hence* $H(\Phi \mid J) = 0$ *and* $I(\Phi; J) = H(\Phi)$.

The temporal label is therefore recoverable without pixels. Removing printed phase names does not remove the code. This result concerns protocol identifiability; Section 5 tests model behavior empirically.

4 Experimental Protocol

Data and models. We use the fixed-camera overhead subset of WTS (Kong et al., 2024): 95 training and 48 validation scenes containing all five phases. The fixed export has 1,624 training rows and 797 validation rows: 95/48 full-scene examples and 1,529/749 auxiliary grounding examples. GPT-4o (OpenAI, 2024) structures source captions into event-target fields. We compare the base VLM, COA-SFT at step 1,200, and Round-1 SFT+GRPO at step 800. The saved SFT state records no best metric or best checkpoint, so its exact selection rule is not recoverable. For GRPO, steps 600/800/1,000 were compared on these same validation scenes; step 800 had the highest pedestrian and vehicle anchor-IoU point estimates. All model comparisons are therefore validation-set evidence

rather than independent-test estimates. Round 2 is a separate run from the same SFT checkpoint, not a continuation; it is reported only as a reward trade-off in Appendix F.

Primary input and metrics. All headline results use the same 48 scene IDs, ten frames sampled uniformly across the event, chronological presentation, displayed image identifiers, and no phase labels. We report micro-averaged VQA over 749 questions, mean pedestrian and vehicle IoU, actor IoU at the model-selected avoidance frame (anchor IoU), and anchor-frame accuracy within 15 frames (AFA@15). For parsed phases, the historical temporal score is the mean of an order score and a phase-pool score: adjacent selected IDs receive order credit when increasing, while a frame receives pool credit 1 in its annotated phase, .4 in an adjacent phase, and 0 otherwise. Anchor IoU and AFA measure localization and moment selection, respectively. Paired intervals use 20,000 scene-bootstrap samples; paired randomization tests use 50,000 sign flips (Efron & Tibshirani, 1994).

Two evaluation regimes must remain distinct. Table 1 and Appendix D preserve the historical field-level evaluator, which scores recoverable answer and phase fields and records a permissive parse flag; it does not gate every metric on whole-record validity. Table 3 instead uses a post-hoc, representation-comparable *strict schema-complete validity* audit. That rule was specified after an evaluator defect was discovered, but frozen before full-corpus strict reaggregation. It requires each representation's complete contract with no repair and assigns zero after failure on the fixed 48-scene/749-question denominator. The historical results are neither overwritten nor relabeled as strict.

Controlled Flat targets. Natural Flat preserves the historical COA-SFT inputs, image order, displayed identifiers, questions, and atomic answer/narrative/anchor/detection supervision, but serializes the assistant target without phase records as four fields: `Answers`, `Narrative`, `Anchor`, and `Detections`. Task-balanced Flat changes only Phase-B sampling, from 219/3,581 to 1,900/1,900 full/auxiliary microbatches; Phase A and the audited model, loss, data, and decoding controls remain fixed. This is microbatch balance, not token balance. Strict Flat parsing was frozen before either Flat evaluation. The comparable strict COA rule was frozen before its full-corpus reaggregation; both reject incomplete or malformed records without repair, retry, or fuzzy matching.

Human evaluation and stress tests. Two primary raters who did not generate the model outputs score COA-SFT and SFT+GRPO on 48 scenes. Model identity is hidden and A/B order is independently randomized per scene. For each 1–5 criterion, we average the two ratings within scene before computing the mean and scene-bootstrap interval. We map each preference vote to GRPO, SFT, or tie and sum the two votes; a zero sum is a scene-level tie. A third rater resolves only the single factual-validity disagreement. We also test a rank-only predictor, shuffle frame presentation while retaining identifiers, and replace image bytes across 47 equal-length scene pairs while holding the target event's questions, identifiers, and annotations fixed.

5 Results and Audits

5.1 RQ1: Can COA training learn spatially grounded records?

Table 1 shows the primary uniform/no-label evaluation. SFT supplies most of the observed spatial gain: pedestrian IoU rises from .014 to .406 and vehicle IoU from .035 to .570. The paired confidence intervals for Base-to-SFT spatial gains exclude zero (Appendix D). GRPO adds smaller changes in the same direction, while its clearest improvement appears in the blinded explanation ratings reported in RQ3; its paired intervals against SFT include zero.

Table 1: Historical field-level matched evaluation on 48 uniform/no-label scenes. Recoverable answer and phase fields are scored without a strict whole-record gate; these values are not schema-complete validity results. VQA is micro-averaged over 749 questions; other entries are scene means. A-IoU uses the phase-4 selected frame; CPP is pedestrian box copy rate. Bold marks the best point estimate, not significance.

Model	VQA	Temp.	Ped	Veh	P A-IoU	V A-IoU	CPP↓	AFA@15
Base	.628	.853	.014	.035	.013	.010	.689	**.188**
COA-SFT	.669	**.897**	.406	.570	.257	.563	**.250**	.146
SFT+GRPO	.665	.887	**.434**	**.627**	**.316**	**.588**	.297	**.188**

The spatial gain does not extend to critical-moment selection. AFA@15 changes from .188 for Base to .146 for SFT, while GRPO returns only to the Base point estimate of .188. The uniform sampler's oracle AFA@15 is .979, meaning that an eligible frame is usually present. Box localization and anchor selection are therefore empirically distinct commitments.

We also intervene on the two direct actor-IoU reward terms across three paired training seeds, 42, 2027, and 2028 (Table 2). Each condition is evaluated at optimizer step 800 on the same 48 scenes. The full reward has positive mean differences on all four spatial metrics, while temporal score changes by $-.012$. Across three seeds, the direct-IoU terms produce consistently positive spatial differences; the small seed count makes this directional rather than population-level evidence.

Table 2: Three-seed intervention on direct spatial-IoU reward terms (mean $\pm$ sample standard deviation). Full and no-direct runs are paired by seed at step 800; the no-direct condition retains zero-box, tracking, and copy regularizers. Δ is the paired seed-wise full-minus-no-direct difference.

Variant	Temp.	Ped	Veh	P A-IoU	V A-IoU
Full reward	$.874 \pm .024$	$.432 \pm .011$	$.614 \pm .012$	$.298 \pm .023$	$.566 \pm .025$
No direct IoU	$.886 \pm .031$	$.405 \pm .012$	$.579 \pm .017$	$.275 \pm .011$	$.505 \pm .051$
Δ	$-.012 \pm .023$	$+.027 \pm .017$	$+.035 \pm .029$	$+.023 \pm .034$	$+.061 \pm .073$

Takeaway 1. SFT provides the main spatial gain; GRPO preserves grounding and improves explanation quality in the blinded evaluation.

5.2 RQ2: How reliably is a complete evidence record produced?

Table 3 applies the post-hoc strict schema-complete rule to every scene and assigns zero after whole-record failure. Natural Flat produces 3/48 valid records; balancing the Phase-B task modes raises this to 15/48, while strict COA-SFT validity is 16/48. In one frozen three-test Holm family, exact paired tests give adjusted $p = .00836$ for Balanced versus Natural Flat, $.00705$ for COA versus Natural Flat, and $p = 1$ for COA versus Balanced Flat. Balancing sharply narrows the earlier reliability gap. The paired test detects no difference between the 16/48 and 15/48 validity outcomes, although this is not an equivalence test.

Table 3: **Primary all-48 representation comparison under the post-hoc strict audit.** Each model uses its frozen schema-complete whole-record rule; failure gives zero on every metric, and VQA keeps all 749 questions. The COA row is separate from the historical field-level values in Table 1.

Model	Strict valid	VQA	Ped	Veh	P A-IoU	V A-IoU	AFA@15
COA-SFT	16/48	185/749	.146	.233	.075	.221	.000
Natural Flat	3/48	39/749	.045	.050	.038	.039	.021
Task-balanced Flat	15/48	194/749	.224	.275	.183	.239	.125

The failure progression shows what balancing repairs. Natural Flat begins with `Detections:` on 45/48 scenes and reaches the 2,600-token cap on 28/48. Task-balanced Flat instead begins with `Answers:` on 48/48 and reaches the cap on 8/48. Balancing largely fixes initial task-mode selection, but whole-record validity remains 15/48 because later fields still fail to serialize or bind reliably.

The strict score can hide fields that appeared before a later failure. We therefore froze a separate field parser before running a *post-hoc diagnostic* (Table 4); it never repairs output or replaces Table 3. Balanced Flat has extractable Answers on all 48 scenes and scores 618/749 (.825) over the full denominator. Narrative, Anchor, and Detections are extractable on 47/48, 16/48, and 35/48, respectively; narrative coverage is syntactic, not a quality score. The recovered Answers score demonstrates substantial scene-level answer competence and localizes later omissions to serialization and evidence-binding fields. Grammar-constrained decoding might reduce syntax errors, but these outputs do not show whether it would recover correct anchors, boxes, or cross-field consistency. The 15 fully valid Balanced scenes are model-selected and easier under several independent historical COA metrics, so their conditional performance is confined to Appendix E.

Table 4: **Post-hoc Flat field-recovery diagnostic.** Coverage uses exact frozen headers and no repair. Answers VQA retains all 749 questions, assigning zero when Answers are not extractable. These values do not replace the primary invalid-as-zero results.

Model	Answers cov.	Answers VQA	Narrative cov.	Anchor cov.	Detect. cov.
Natural Flat	10/48	97/749 (.130)	4/48	3/48	4/48
Task-balanced Flat	48/48	618/749 (.825)	47/48	16/48	35/48

Balanced Flat's strict all-scene point estimate exceeds strict COA on VQA, both actor IoUs, both anchor IoUs, and AFA@15. These are descriptive comparisons; no semantic-performance significance test was frozen for this post-hoc audit.

Takeaway 2. Balancing recovers much of Flat's schema completion, showing that task mixture is a major determinant of output reliability. COA contributes a phase-indexed interface for training and auditing evidence-bound records.

5.3 RQ3: What does reward optimization change beyond automatic scores?

The blinded ratings in Table 5 favor GRPO for visual factuality, causal support, and evidence consistency. Temporal coherence is unchanged near the ceiling. Raters prefer GRPO on 35 scenes, SFT on one, and tie on 12 (exact sign test on non-ties, $p = 1.08 \times 10^{-9}$). After third-rater adjudication of one disagreement, factual validity is 83.3% for GRPO and 75.0% for SFT; the paired difference has a 95% interval of $[0.0, 18.8]$ percentage points.

Table 5: Two-rater blinded evaluation on 48 scenes (1–5; paired scene-bootstrap 95% CIs).

Criterion	COA-SFT	SFT+GRPO	Δ	95% CI
Visual factuality	4.30	4.67	+.36	[.23, .50]
Temporal coherence	4.99	4.98	-.01	[-.06, .03]
Causal support	4.36	4.71	+.34	[.19, .50]
Evidence consistency	4.30	4.68	+.38	[.23, .52]

Takeaway 3. GRPO leaves automatic grounding broadly unchanged while improving visual factuality, causal support, and evidence consistency in the blinded evaluation.

5.4 RQ4: Do temporal scores require visual evidence?

A rank-only predictor reaches a temporal score of 1.000 on the 38 scenes with exactly two frames from each phase and .958 across all 48 phase-aware samples. On uniform ten-frame samples, rank-only phase accuracy falls from 1.000 to .783. The learned models also depend on presentation order: shuffling uniformly sampled frames reduces temporal score from .897 to .659 for SFT and from .887 to .647 for GRPO.

The stronger test changes pixels while preserving textual and positional cues. In the 47 equal-length pairs in Table 6, replacing one event's images with another's reduces GRPO pedestrian/vehicle IoU from .425/.621 to .013/.042, yet temporal score moves from .887 to .888 ($\Delta = .001$, $p = .978$). The spatial collapses have paired $p < 10^{-4}$. This cross-scene replacement test establishes the visual dependence of localization; generated narrative correctness is evaluated separately.

Table 6: Cross-scene pixel replacement on 47 paired scenes. Questions, displayed identifiers, and annotations stay fixed; only image bytes are replaced. Differences are replaced minus aligned.

Model	Pixels	Temporal	Ped IoU	Veh IoU
COA-SFT	aligned	.916	.414	.582
COA-SFT	replaced	.897	.012	.035
SFT+GRPO	aligned	.887	.425	.621
SFT+GRPO	replaced	.888	.013	.042

Sampling sensitivity also appears outside Qwen (Table 7). Pixtral-12B (Mistral AI, 2024) receives ten images natively, whereas DeepSeek-VL2-Tiny (Wu et al., 2024) receives the same frames in a labeled 2×5 contact sheet because of its interface limit. Both score higher under uniform sampling. These frozen open-VLM probes broaden the scope of the sensitivity result, but they are temporal-selection probes rather than trained grounding baselines. Because the input interfaces differ, these probes establish sampling sensitivity across backbones rather than a universal causal mechanism. The deterministic construction and rank-only baseline establish the exact shortcut for the equal-per-phase protocol.

Table 7: Cross-backbone sampling probes on 48 scenes. Entries are temporal-selection scores; intervals are paired scene-bootstrap 95% CIs for uniform minus phase-aware sampling.

Probe	Input interface	Phase	Uniform	Δ [95% CI]
Pixtral-12B	ten native images	.723	.839	+.116 [.081, .150]
DeepSeek-VL2-Tiny	labeled 2×5 sheet	.696	.853	+.157 [.124, .188]

Takeaway 4. Actor localization depends on event-specific pixels, whereas the temporal score can remain high through input rank and output order. Temporal structure and visual grounding therefore require separate audits.

6 Discussion

COA turns moment selection, actor localization, and explanation into separate commitments that can be scored and perturbed. The Flat controls show that task balance has a large effect on complete-record generation, while COA retains an explicit phase-indexed structure for temporal and spatial audit.

The diagnostics also separate question answering from complete evidence binding. Balanced Flat produces extractable Answers on all 48 scenes but less often completes Anchor and Detections, indicating that answering questions and emitting a complete evidence record are different capabilities. Constrained decoding may improve grammar, but its effect on anchors, boxes, and cross-field consistency remains open. The shortcut tests reinforce the same point: temporal structure and pixel-grounded localization must be evaluated separately.

7 Limitations

Our evidence comes from Qwen3-VL-8B and 48 WTS validation scenes rather than an independent test set. The Flat controls use one training seed, and bootstrap intervals quantify scene variation rather than training-seed variation. GRPO checkpoint selection uses the same validation scenes as headline evaluation, and the historical selection rule for the SFT step-1,200 checkpoint was not retained.

The strict COA audit, field-recovery analyses, and paired validity tests were specified after earlier outputs had been inspected and were then frozen before full-corpus aggregation. The 15 Balanced-valid scenes form a model-selected subset. Target length, the 2,600-token cap, and microbatch rather than token balance leave multiple explanations for the remaining Flat failures; constrained decoding is untested. The cross-backbone probes evaluate temporal selection through different input interfaces, and exact rank identifiability applies to equal-per-phase sampling.

8 CONCLUSION

COA makes traffic-VLM predictions auditable by linking each traffic phase to a displayed frame, actor boxes, and an event account. SFT provides the main grounding gain, while GRPO preserves grounding and improves explanation quality in blinded human evaluation. Across three seeds, direct spatial rewards provide directional evidence that the intended fields respond to reward design. The Flat controls show that task mixture strongly affects complete-record generation, and the shortcut tests show that temporal scores can remain high even when event-specific visual grounding falls almost to zero. COA therefore provides a practical interface for training, scoring, and stress testing evidence-bound traffic reasoning.

AI USE STATEMENT

Generative AI tools were used to structure existing WTS captions into event-target fields, provide feedback on experimental design, assist with implementation and debugging of analysis code, help formulate and check mathematical claims, support interpretation of results, and assist with manuscript organization, drafting, editing, bibliography verification, formatting, and revision of scientific figures and captions. Generative AI was not used as a rater in the blinded human evaluation. AI-assisted code, derivations, statistics, citations, and manuscript claims reported here were checked against primary sources and saved artifacts. Responsibility for the final content remains with the authors.

ETHICS STATEMENT

This work is intended for human-led traffic-safety research, not automated findings of fault or liability. It uses roadway video from the published WTS dataset and does not perform identity recognition or infer sensitive personal traits. The human study evaluates model outputs, and the paper reports no rater-identifying information. Models, annotations, and explanations can be wrong and may not transfer across locations, weather, road design, or demographic groups. Use on live roadway feeds would require independent safety validation, privacy controls, and human oversight.

REPRODUCIBILITY STATEMENT

Sections 4–5 state the split, comparison groups, frozen scoring rules, denominators, interventions, and paired analysis. The appendix gives output grammars, training settings, reward components, seeds, statistical details, and extended results. Reproduction additionally requires the samplers, evaluators, frozen parsers, analysis code, split manifests, and saved scene-level outputs used for the reported tables. Raw WTS media remain governed by the dataset's terms.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4971–4980, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2917–2927, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Buch_Revisiting_the_Video_in_Video-Language_Understanding_CVPR_2022_paper.html.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948. URL https://arxiv.org/abs/2501.12948.

Quang Minh Dinh, Minh Khoi Ho, Anh Quan Dang, and Hung Phong Tran. TrafficVLM: A controllable visual language model for traffic video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7134–7143, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/AICity/html/Dinh_TrafficVLM_A_Controllable_Visual_Language_Model_for_Traffic_Video_Captioning_CVPRW_2024_paper.html.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1994. ISBN 9780412042317. URL https://www.routledge.com/link/link/p/book/9780412042317.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Edward J. Hu, Yuanzhi Li, Yelong Shen, Shean Wang, Phillip Wallis, Lu Wang, Zeyuan Allen-Zhu, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Karim Ismail, Tarek Sayed, Nicolas Saunier, and Clark Lim. Automated analysis of pedestrian–vehicle conflicts using video data. *Transportation Research Record*, 2140(1):44–54, 2009. doi: 10.3141/2140-05. URL https://journals.sagepub.com/doi/10.3141/2140-05.

Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, Yoichi Sato, and Norimasa Kobori. WTS: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. In *Computer Vision – ECCV 2024*, volume 15134 of *Lecture Notes in Computer Science*, pp. 1–18, 2024. doi: 10.1007/978-3-031-73116-7_1. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/9667_ECCV_2024_paper.php.

Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Benchmark for evaluating pedestrian action prediction. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1258–1268, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Kotseruba_Benchmark_for_Evaluating_Pedestrian_Action_Prediction_WACV_2021_paper.html.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision*, pp. 706–715, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Krishna_Dense-Captioning_Events_in_ICCV_2017_paper.html.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4. URL https://www.nature.com/articles/s41467-019-08987-4.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487–507, 2023. doi: 10.18653/v1/2023.acl-long.29. URL https://aclanthology.org/2023.acl-long.29/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2004. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Mistral AI. Pixtral-12B-2409 model card, 2024. URL https://huggingface.co/mistralai/Pixtral-12B-2409.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. doi: 10.48550/arXiv.2410.21276. URL https://arxiv.org/abs/2410.21276.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. URL https://www.bmvc2020-conference.com/assets/papers/0306.pdf.

Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *IEEE/CVF International Conference on Computer Vision*, pp. 6262–6271, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Rasouli_PIE_A_Large-Scale_Dataset_and_Models_for_Pedestrian_Intention_Estimation_ICCV_2019_paper.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. doi: 10.48550/arXiv.2504.07615. URL https://arxiv.org/abs/2504.07615.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*, 2024a. doi: 10.48550/arXiv.2410.03290. URL https://arxiv.org/abs/2410.03290.

Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Fady Alnajjar, Ganzorig Batnasan, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7261–7272, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024W/AICity/html/Wang_The_8th_AI_City_Challenge_CVPRW_2024_paper.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models.

In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. doi: 10.48550/arXiv.2412.10302. URL https://arxiv.org/abs/2412.10302.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *Proceedings of the 2nd International Workshop on Human-centric Multimedia Analysis*, pp. 13–21, 2021. doi: 10.1145/3475723.3484247. URL https://arxiv.org/abs/2101.09028.

Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient VLMs. *arXiv preprint arXiv:2505.15436*, 2025. doi: 10.48550/arXiv.2505.15436. URL https://arxiv.org/abs/2505.15436.

A Proofs and Formal Details

Additive advantage decomposition. Within update t, κ_t is shared by every completion in the reward group. Therefore $R_{t,g} - \bar{R}_t = \sum_{c \in \mathcal{C}} \alpha_c(t)(r_{t,g,c} - \bar{r}_{t,c})$; the $-\kappa_t$ terms cancel. Dividing by the common positive scale s_t gives Equation 2. If a completion-dependent component is constant within the group, its centered contribution is zero. The trainer's separate token-level reference KL is part of the policy loss and is not included in this identity.

Rank identifiability. The phase-aware sampler fills ranks $1, \ldots, k$ from phase 0, ranks $k+1, \ldots, 2k$ from phase 1, and so on through phase $N_\phi - 1$. Thus $\Phi(J) = \lfloor (J-1)/k \rfloor$ is deterministic, $H(\Phi \mid J) = 0$, and $I(\Phi; J) = H(\Phi) - H(\Phi \mid J) = H(\Phi)$.

B Output Contracts and the Flat Control

The evaluation-time COA record has exactly five tagged blocks in order: `SOLUTION`, `CAPTION`, `REASONING`, `ANCHOR_ID`, and `COA`. `SOLUTION` is strict answer JSON; `CAPTION` is nonempty and at most 55 words; `REASONING` contains nonempty `Step 1` and `Step 2` spans; and `ANCHOR_ID` contains the four ordered atoms `TARGET_PED`, `TARGET_VEH`, `TRACK_START`, and `TRACK_NOTE`, with `TRACK_START` referring unambiguously to a displayed identifier. The `COA` block contains a nonempty event account ending in syntactically valid Risk and Outcome fields, followed by one strict JSON array. The array has phases 0–4 in order, one strictly increasing displayed identifier per phase, and normalized finite $xyxy$ boxes with positive area for both actors. The parser consumes the complete response apart from permitted structural whitespace and performs no repair, retry, truncation, or fuzzy match. Validity tests completeness and syntax/value domains, not answer correctness, phase correctness, IoU, anchor timing, or narrative truth.

This strict rule is a post-hoc evaluator audit, not a prospective metric from the original experiment. It was specified after discovery of a defect in the historical parser; historical flags and at least one illustrative response had already been inspected. The protocol and its three-test Holm family were then frozen before the full 48-response reaggregation. The historical evaluator in Tables 1 and D

remains a separate field-level path: it scores recoverable fields and its permissive `coa_parse_ok` flag is not whole-record validity.

The Flat converter preserves the original input messages, image references, image order, displayed IDs, and questions. A full assistant target has exactly four ordered fields: `Answers`, `Narrative`, `Anchor`, and `Detections`; an auxiliary target contains only `Detections`. It contains no `<COA>` or `<GROUNDING>` tag, phase label, or five per-phase records. Converter audits establish 1,624 training rows and a deterministic dataset digest. The Flat evaluator uses the existing field value parsers and official VQA/spatial scorers, but its strict whole-record parser follows the Flat grammar. No constrained decoding, prompt repair, output repair, retry, or replacement is used.

Natural Flat uses 1,000 Phase-A grounding-only microbatches followed by 3,800 Phase-B microbatches, of which 219 are full and 3,581 auxiliary. Task-balanced Flat preserves Phase A and deterministically interleaves 1,900 full and 1,900 auxiliary Phase-B microbatches. Exposure is not token-balanced: Phase B contains 3,315,960 supervised full-target tokens and 263,373 auxiliary-target tokens. Both Flat runs start fresh from the same 4-bit base, use seed 42, batch size 1 with four-step accumulation, LoRA $r = 8$, $\alpha = 16$, dropout .05, AdamW 8-bit, learning rate 8×10^{-6}, and a .05 warmup ratio. The configured maximum is 2,000 optimizer steps with a fixed stop at step 1,200. Steps 0–249 supervise only the exact `Detections` span; later steps supervise all and only assistant-target tokens with equal cross-entropy weight. Both runs use the same 2,600-token generation cap. Natural and Balanced Flat reach that cap on 28/48 and 8/48 scenes, respectively. Normalizing only a missing final newline leaves complete validity unchanged at 3/48 and 15/48, so the primary gap is not a trailing-newline artifact.

C Training and Data Details

Camera selection first requires all five phases, then chooses the fixed overhead view with the largest combined pedestrian and vehicle box area. Boxes are converted from WTS (x, y, w, h) pixels to normalized $xyxy$ and invalid boxes are removed.

COA-SFT uses 4-bit NF4 quantization, LoRA rank 8 with $\alpha = 16$ and dropout .05, learning rate 8×10^{-6}, batch size 1, gradient accumulation 4, seed 42, and AdamW 8-bit. Detection-only supervision is used for optimizer steps 0–249; all assistant-target tokens receive equal cross-entropy weight from step 250 onward. The reported fresh-base checkpoint is step 1200. Both visual and language LoRA tensors are present in the adapter.

GRPO starts from that checkpoint, uses two configured generations per prompt, LoRA rank 8, learning rate 3×10^{-7}, and 1,000 configured steps. The signed main reward weights are temporal .20, pedestrian spatial .30, vehicle spatial .20, order .10, narrative .10, format .05, and VQA .05. Spatial terms are off for 50 steps and ramp to full weight by step 200; zero-box, copy, and tracking terms follow the same ramp. The copy penalty is capped at $-.30$. A lagged, group-shared threshold penalty uses coefficient .02 and threshold 2.0 inside the logged scalar reward, while the trainer independently uses token-level reference KL coefficient .001 in the policy loss. Training used two RTX A6000 GPUs.

D Full Matched COA Evaluation

For uniform/no-label evaluation, paired scene-bootstrap 95% intervals for SFT minus Base are $[.305, .478]$ for pedestrian IoU, $[.441, .625]$ for vehicle IoU, $[.160, .333]$ for pedestrian anchor IoU, and $[.446, .657]$ for vehicle anchor IoU. The corresponding GRPO minus SFT intervals are $[-.022, .087]$, $[-.020, .139]$, $[-.011, .136]$, and $[-.085, .137]$, respectively.

For uniform sampling, the oracle AFA is .479/.979/1.000/1.000 at 5/15/30/60-frame tolerances. The raw GRPO AFA is .042/.188/.250/.500 at those tolerances. We report the oracle because a sampler may omit the annotated critical frame.

Table 8: Historical field-level matched evaluation on the same 48 scenes. Phase denotes ground-truth phase-aware sampling; labels says whether phase labels are printed. VQA is micro-averaged over 749 questions. Recoverable fields are scored without the post-hoc strict whole-record gate.

Model	Sampler	Labels	VQA	Temp.	Ped	Veh	P A-IoU	V A-IoU	AFA@15
Base	Phase	yes	.637	.927	.023	.033	.004	.011	.479
Base	Phase	no	.641	.887	.024	.057	.012	.023	.312
Base	Uniform	no	.628	.853	.014	.035	.013	.010	.188
COA-SFT	Phase	yes	.662	.843	.405	.526	.289	.488	.438
COA-SFT	Phase	no	.674	.800	.480	.655	.292	.574	.396
COA-SFT	Uniform	no	.669	.897	.406	.570	.257	.563	.146
SFT+GRPO	Phase	yes	.662	.883	.420	.528	.281	.473	.542
SFT+GRPO	Phase	no	.668	.771	.474	.658	.324	.599	.375
SFT+GRPO	Uniform	no	.665	.887	.434	.627	.316	.588	.188

E Post-hoc Whole-record and Flat Diagnostics

All analyses in this section were specified after seeing the primary outputs and do not replace the all-48 invalid-as-zero table. The field parser requires unique exact headers in canonical order and applies existing value rules without fuzzy matching, repair, truncation, or retry.

Table 9: Field-wise recovery under the frozen post-hoc protocol. Conditional Answers VQA uses only scenes with extractable Answers; all-48 VQA retains 749 questions. Narrative coverage is syntax only.

Model	Whole valid	Ans. cov.	Ans. VQA cond.	Ans. VQA all	Narr.	Anchor / Detect.
Natural Flat	3/48	10/48	97/149 (.651)	97/749 (.130)	4/48	3/48 / 4/48
Balanced Flat	15/48	48/48	618/749 (.825)	618/749 (.825)	47/48	16/48 / 35/48

The 15 scenes on which Balanced Flat is strictly valid were frozen before subset scoring. On this model-selected subset, Balanced Flat scores 194/223 (.870) VQA, .716 pedestrian IoU, and .880 vehicle IoU. COA-SFT scores 160/223 (.717), .524, and .592 on the same scenes. However, COA itself scores higher on this subset than on the complementary 33 scenes on six of seven reported rates or metrics (for example, VQA .717 versus .648 and pedestrian IoU .524 versus .352). The subset is therefore selection-biased evidence about successful Balanced outputs, not an all-scene comparison.

The frozen paired strict-validity tables are: COA/Natural both valid 1, COA only 15, Natural only 2, neither 30; COA/Balanced both valid 3, COA only 13, Balanced only 12, neither 20; and Balanced/Natural both valid 1, Balanced only 14, Natural only 2, neither 31. Two-sided exact McNemar/binomial raw p values are $.002350$, 1, and $.004181$. One Holm correction over all three gives $.007050$, 1, and $.008362$, respectively. The strict COA failures, recorded only as a post-hoc diagnostic, are malformed actor box (18), unknown top-level block (7), non-increasing image identifiers (3), reversed or zero-area box (2), out-of-range coordinate (1), and missing top-level block (1). Categories describe parser outcomes, not failure causes.

F Reward Interventions and Round 2

Table 2 reports the three-seed direct-IoU intervention in the main paper. For the original seed, Round-1 checkpoints 600/800/1,000 obtain pedestrian anchor IoU .276/.316/.300 and vehicle anchor IoU .553/.588/.545 on the same validation scenes. Step 800 is the reported checkpoint because both anchor-IoU point estimates are highest there. Checkpoint 1,000 has lower copy rates, so the trajectory contains a real objective trade-off rather than uniform late-stage improvement.

Round 2 starts independently from the same SFT checkpoint, relaxes the copy cap from $-.30$ to $-.10$, and adds a diversity bonus. Relative to Round 1, vehicle copy rate falls from .266 to .142, while pedestrian/vehicle IoU changes from .434/.627 to .411/.595 and pedestrian/vehicle anchor IoU

from .316/.588 to .267/.521. Round 1 is therefore the primary model; Round 2 is not a continued or uniformly improved checkpoint.

G Additional Shortcut and Cross-backbone Results

Under uniform order shuffling, SFT temporal score changes from .897 to .659 and GRPO from .887 to .647. Pedestrian/vehicle IoU changes from .406/.570 to .383/.570 for SFT and from .434/.627 to .402/.578 for GRPO. Presentation order changes while original image identifiers remain, so this is a sensitivity test rather than a natural deployment distribution.

For cross-scene replacement, paired differences (replaced minus aligned) are $-.402$ pedestrian IoU and $-.547$ vehicle IoU for SFT, and $-.412$ and $-.578$ for GRPO. Their 95% intervals are $[-.493, -.312]$, $[-.639, -.452]$, $[-.501, -.320]$, and $[-.663, -.489]$; all paired sign-flip values are below 10^{-4}. The GRPO temporal difference is $+.000957$ with interval $[-.073, .076]$ and $p = .9780$.

The cross-backbone values and scope qualification appear in Table 7. Their frozen prompts request one valid displayed identifier for each phase rather than the full COA schema. All outputs in both reported probes parse, contain five valid identifiers, and preserve order. The DeepSeek contact sheet is row-major and prints the identifier above each image.

H Human Evaluation Details

The interface shows ten frames and two anonymized outputs in randomized order. Visual factuality measures agreement with visible actions and objects; temporal coherence evaluates the stated sequence; causal support asks whether risk changes and outcomes follow from the evidence; evidence consistency checks agreement among frames, boxes, and prose. Exact rater agreement is .750, .979, .823, and .719 for these criteria. Linearly weighted Cohen's κ is .584, .000, .704, and .560; temporal $\kappa = 0$ results from near-constant ceiling ratings despite 97.9% exact agreement. Factual-validity agreement is 99.0%.

I Reproducibility Checklist

All automatic comparisons use frozen scene IDs and saved per-scene outputs. The historical COA regression reproduces 24/24 displayed field-level reference metrics after standard three-decimal formatting. Flat aggregation reproduces every integer and floating-point reference within 10^{-9}. The strict COA audit preserves all 48 scenes and the 749-question denominator, and hashes its frozen parser, tests, protocol, inputs, and outputs. Reproduction requires the exact commands, hashes, parsers, samplers, target converters, data manifests, human-evaluation rubric, and paired-analysis scripts.