FAITHFUL REASONING UNDER INTENT AMBIGUITY VIA TOULMIN-AUGMENTED CHAIN-OF-THOUGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) tend to give confident answers on driving scenes that even humans find ambiguous. We argue that free-form chain-of-thought (CoT) inherits a habit of human reasoning: the reasons are generated after the conclusion, and the premise that actually produced the conclusion is left out. Following Toulmin's model of argumentation, we build a multi-agent pipeline in which each agent has its own responsibility. Several agents describe what is visible in the video, others propose a warrant for each possible conclusion (in Toulmin's terms, the general rule that leads from the observations to the conclusion), and a judge weighs the two warrants against the video and reaches a conclusion. Because the observations are written before any conclusion exists, the trace cannot be a rationalization of an answer reached in advance. We evaluate on 200 ambiguous pedestrian intent samples with two expert human studies and standard classification metrics. First, expert annotators compared each method's reasoning output against the video. The Toulmin pipeline produces 39% fewer traces with a major hallucination than self-consistency CoT (31.4% against 51.5%). Second, in a study without the video, experts recover self-consistency CoT's conclusion from its observations alone in 94.6% of cases, and its reason changes their conclusion in only 5.0%, which is the signature of post-hoc rationalization. By contrast, the Toulmin warrant changes the reader's conclusion in 37.0% of cases. On the most ambiguous cases, self-consistency CoT is rated clearer, yet it is correct on only 8.9%, against 35.6% for Toulmin. On standard classification metrics, Toulmin also improves accuracy, by 6.5 points over the warrant-ablated pipeline (48.0% against 41.5%) and 21 points over self-consistency CoT (27.0%).

1 INTRODUCTION

Vision-Language Models (VLMs) and their extension to action, Vision-Language-Action (VLA) models, have moved quickly from answering questions about an image to reading video, predicting what the agents in it will do, and acting on the prediction (Brohan et al., 2023; NVIDIA Autonomous Vehicle Research Group et al., 2025). Their reasoning has improved with them. Trained with chain-of-thought (CoT) supervision and reinforcement learning (Wei et al., 2022; Shao et al., 2024), a current model writes a long, orderly trace before it answers. Yet the trace still leaves something out. A model that has watched thousands of doors being opened will say that it turned the handle and the door opened, so turning the handle opens the door. It will not say that the handle pulls back the latch. The rule that turns the observation into a reason, the piece of physics or social convention the model is relying on, is never written down. The trace tells you what the model saw and what it decided. The link between the two stays in the model's head.

Free-form CoT inherits a habit of human reasoning. People reach a conclusion first and generate the reasons afterwards (Nisbett & Wilson, 1977; Haidt, 2001; Johansson et al., 2005); the evidence is selected and read in the light of that conclusion (Kunda, 1990); and the premise that actually produced it is left out of the account (Habernal et al., 2018). CoT is trained on text written this way and writes the same way, so its stated reasons often do not contain what moved the answer (Turpin et al., 2023; Lanham et al., 2023; Arcuschin et al., 2025; Chen et al., 2025). In a vision-language model, the observations can be affected too, since the perception steps of a chain need not be grounded in the image (Li et al., 2026; Uppaal et al., 2025). A free-form chain gives the observations, the rule, and

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.

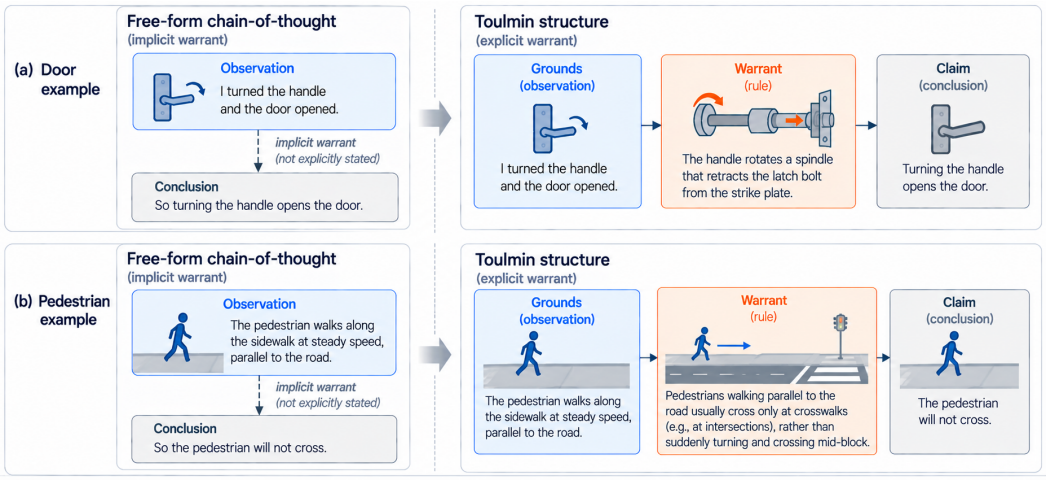

Figure 1: Free-form chain-of-thought and the Toulmin structure on two examples. Left: the chain moves from an observation to a conclusion, and the rule that licenses the step is not stated. Right: the same argument with the **warrant** written out as a separate component between the **grounds** and the **claim**. The warrant is the premise that human explanations leave out and that free-form CoT never states; our pipeline produces it as its own output before it is judged.

the conclusion in one piece of text, so a reader cannot tell whether the conclusion followed from the observations or the observations were written to fit the conclusion.

Whether this matters depends on the case. If the scene is clear, a trace that rationalizes and a trace that reasons both reach the correct result. If the scene is ambiguous, there may be no single correct result at all. The model then assumes a premise, argues from it, and produces a clear and convincing case for an answer that the scene does not settle. The argument is fluent because the premise was chosen to make it so. This is a hallucination in the form of a rationalization. No current reasoning model, on such a case, stops to consider which rule should apply and whether the evidence supports it. That is the gap this paper addresses. We study it on pedestrian crossing intent in the PSI dataset (Jing et al., 2025), restricted to clips on which the human annotators themselves disagree.

Toulmin's model of argument (Toulmin, 2003) names the missing part. An argument moves from *grounds*, the observations, to a *claim*, the conclusion, by way of a *warrant*, the general rule that licenses the step (Figure 1). With the aim of making the trace faithful, we build a multi-agent pipeline that produces the three in that order, so that the observations are written before any conclusion exists and the rule is written before it is judged (Section 3). Whether this changes the text cannot be read from a confusion matrix, since a rationalization and an argument can carry the same label. We therefore have expert annotators read the reasoning itself, once with the video to check the observations, and once without it to see what the rule adds (Section 3.3).

Contributions. **(C1) A multi-agent Toulmin pipeline that improves faithfulness.** Agents describe the video before any conclusion exists, two agents state opposing warrants, and a judge selects between them by checking their conditions against the video (Section 3.1). Checked against the video by expert annotators, the pipeline produces 39% fewer traces with a major hallucination than self-consistency CoT (31.4% against 51.5%). Read without the video, its warrant changes the reader's conclusion in 37.0% of cases against 5.0% for the CoT reason (Section 5.2).

(C2) Evidence that free-form CoT rationalizes, and that this reads well. The same readers find that self-consistency CoT's observations already carry its conclusion and its reason adds nothing, yet rate its traces as clearer and better supported and prefer them. On the most ambiguous clips it never answers `not_sure` and is correct on 8.9%, against 35.6% for Toulmin.

(C3) Consistent gains on standard classification metrics. On 200 ambiguous PSI windows the pipeline reaches 48.0% accuracy against 41.5% for a warrant-ablated variant and 27.0% for self-consistency CoT, with `yes`-recall of 0.681 against 0.159 and 0.116 (Section 5.1).

2 Related Work

2.1 Structured Reasoning

Chain-of-thought prompting (Wei et al., 2022) and self-consistency CoT (Wang et al., 2023b) leave the trace free-form and give no signal on its quality. Later work shapes the trace, either by changing its topology (Tree of Thoughts (Yao et al., 2023), Graph of Thoughts (Besta et al., 2024), Skeleton-of-Thought (Ning et al., 2024), Least-to-Most (Zhou et al., 2023), Plan-and-Solve (Wang et al., 2023a)) or by tying each step to a form that can be executed or checked (Chain-of-Table (Wang et al., 2024b), PAL (Gao et al., 2023), Faithful CoT (Lyu et al., 2023)). All of these split the trace into steps; none makes the general rule behind a step an object of its own. Toulmin's schema (Toulmin, 2003) does exactly that, and it has long served argument mining in NLP (Stab & Gurevych, 2017; Habernal & Gurevych, 2017), where the main finding is that the warrant is almost never stated in natural text and must be reconstructed (Habernal et al., 2018), a task on which models rely on surface cues (Niven & Kao, 2019). We use the schema as a constraint on generation rather than a target for extraction. The judge in our pipeline follows multi-agent debate (Irving et al., 2018; Du et al., 2024; Liang et al., 2024), but the debate is confined to one component, the warrant, since free-form debate does not reliably beat self-consistency (Smit et al., 2024; Wang et al., 2024a). The one prior debate system for VLMs uses a text-only judge (Adhikari & Lapata, 2025); ours re-watches the video.

2.2 Faithfulness of Chain-of-Thought

A chain of thought reads as the reason for an answer, and a line of work asks whether it is. Chains omit the feature that moved the answer (Turpin et al., 2023; Chen et al., 2025), and truncating or corrupting them often leaves the answer unchanged (Lanham et al., 2023; Arcuschin et al., 2025). Jacovi & Goldberg (2020) call this property faithfulness and separate it from plausibility, whether the explanation convinces a reader. Multimodal chains fail in a second way: their perception steps need not be grounded in the image, a property measured with automatic judges under the names perceptual and visual faithfulness (Li et al., 2026; Uppaal et al., 2025). We measure it on driving video with expert annotators, and we measure plausibility on the same traces so that the two can be compared. Two responses to unfaithful chains are close to ours: treating a structured argument as the explanation (Cakar & Kristensson, 2025), and separating perception from reasoning in post-training (Wu et al., 2026). Our pipeline separates them at inference time, in the output.

2.3 Driving VLMs and Pedestrian Intent

Driving is where VLM reasoning is being deployed fastest, and where a wrong but fluent trace costs the most. Recent systems attach reasoning to driving decisions in several forms: graph QA over perception, prediction, and planning (DriveLM (Sima et al., 2024)), CoT traces (Reason2Drive (Nie et al., 2024)), hierarchical planning (DriveVLM (Tian et al., 2024)), interpretable control (DriveGPT4 (Xu et al., 2024)), free-form VQA (LingoQA (Marcu et al., 2024), DrivingVQA (Corbière et al., 2026)), causal chains tied to the action (Alpamayo-R1 (NVIDIA Autonomous Vehicle Research Group et al., 2025)), and counterfactual reasoning over alternative actions (OmniDrive (Wang et al., 2025)). In every case the reasoning is one continuous text split along driving subtasks, so the rule linking observation to decision is never a field of its own; Social Genome (Mathur et al., 2025) comes closest by labelling external knowledge as a step, but does not test what writing it down changes. Pedestrian crossing intent is a natural downstream task for this question. Trajectory models such as SF-GRU (Rasouli et al., 2019b), bounding-box transformers (Achaji et al., 2022), and TrEP (Zhang et al., 2023) reach high accuracy on PSI (Jing et al., 2025), JAAD (Rasouli et al., 2017), and PIE (Rasouli et al., 2019a) but give no rationale, and PSI records annotator disagreement per clip, which lets us select the ambiguous cases on which a rationalization is hardest to tell from an argument.

3 Method

3.1 Multi-Agent Toulmin Orchestrator

The orchestrator (Figure 2) converts each PSI clip into a structured argument $(\text{grounds}, \text{warrants}, \text{answer})$ with an explicit `not_sure` option. Six vision-grounded agents

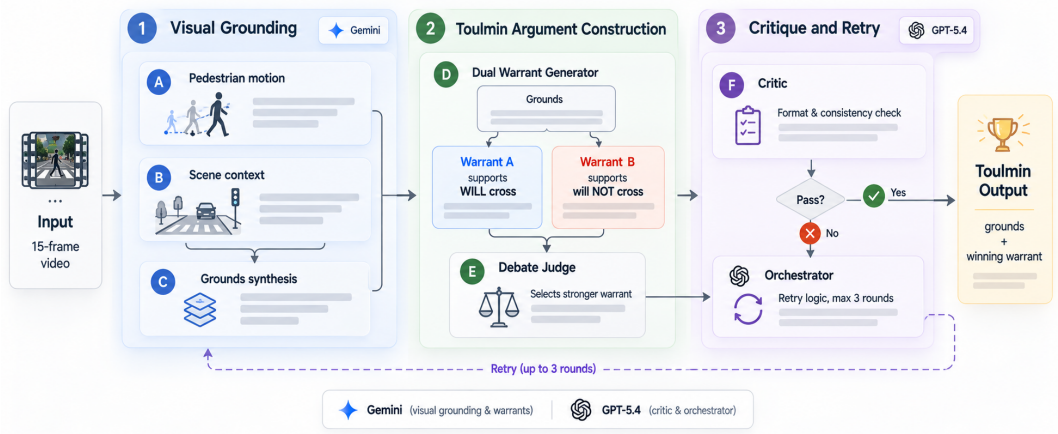

Figure 2: The multi-agent Toulmin orchestrator. Agents A and B extract pedestrian motion and scene context from the video and Agent C combines them into the *grounds*, with no conclusion available. Agents D-yes and D-no each write a warrant for one side. Agent E re-watches the video and picks the warrant whose visible condition holds, or returns `not_sure`. Agent F checks format and consistency and can override to `not_sure`. A controller retries failed stages up to three times.

run on Gemini; a text-only critic and a controller run on GPT-5.4, and the controller retries a failed stage up to three times. The agents run in three stages, and no first-stage agent sees a conclusion.

Grounds. Agent A reports pedestrian motion cues from the raw clip (gaze direction, body and torso orientation, gait phase, speed changes within the window); Agent B reports scene cues (road type, traffic control, distance to curb, vehicle proximity, crosswalks); Agent C fuses the two into a grounds string of at most three sentences, in present tense and visual-only language.

Warrants. Agents D_{yes} and D_{no} each receive the grounds and argue one side. Each writes a short free-form reasoning trace, then commits to two structured outputs: a *warrant*, one 12–25 word general rule drawn from {*social_norm*, *physical_law*, *traffic_rule*}, and an *applicability caveat*, one visible condition that must hold in the late frames for the warrant to apply.

Judgment. Agent E re-watches the video, lists the key late-frame features, and records which features confirm or falsify each side's caveat. A deterministic rule commits to `yes` or `no` only when one side has clearly stronger and uncontradicted evidence, and returns `not_sure` otherwise (balanced evidence, a behaviour reversal within the window, or occluded cues). Agent F, the critic, checks the output format and overrides the answer to `not_sure` when the judge's evidence is close or mixed.

3.2 Baselines

The pipeline differs from free-form CoT in two ways: it runs several agents in a debate, and it forces the warrant to be written out. The two baselines separate these (schematics in Appendix A).

Self-Consistency Chain-of-Thought (SC-CoT). The free-form reference, with neither change. We follow Wang et al. (2023b) with $K=6$ CoT samples per video at temperature 1.0. The prompt asks the model to reason through the same four cue categories the orchestrator decomposes (motion, scene, late-frame features, judgment), but in a single flat trace. The final answer is the majority vote; `not_sure` is returned when the top class leads the second by fewer than 2 votes. SC-CoT is also the method whose traces the human studies compare against, since it is the reasoning format the faithfulness literature examines.

Agentic-Debate (warrant-ablated). The first change without the second. It keeps the multi-agent debate and removes the warrant. Each side emits free-form `thinking` and `key_observations`; the judge counts `yes_evidence` and `no_evidence` from late-frame features and applies the same margin rule and the same `not_sure` triggers as the full pipeline; the critic has the same override authority. Agent count, compute, and dual-side argumentation are held fixed, so any difference between this baseline and the full pipeline is due to the warrant alone, and any difference between this baseline and SC-CoT is due to the debate alone. Both run at temperature 0.3.

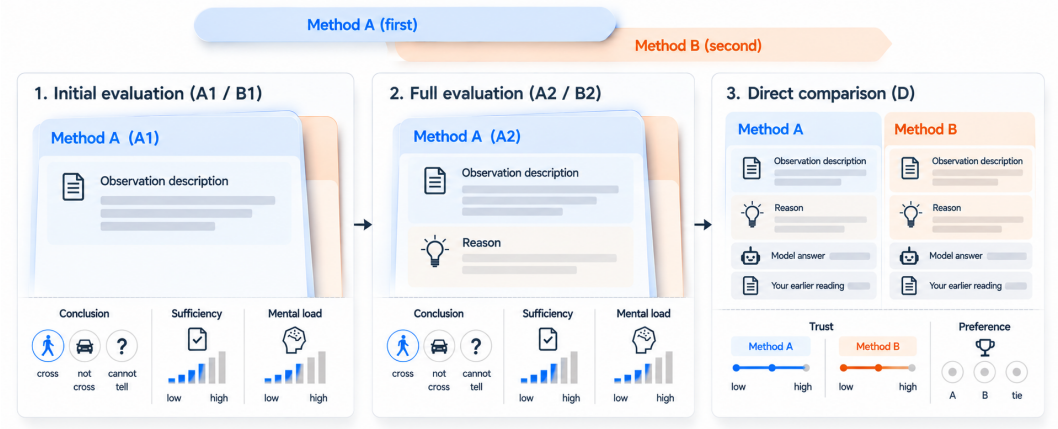

Figure 3: Strength-of-reasoning study. Each clip is shown without the video in three stages: the initial evaluation shows one method's observations alone (Part 1), the full evaluation adds its reason (Part 2), and the direct comparison reveals both methods' full traces with the model answers. The sequence runs for method A, then method B, in random order with names hidden. Each stage asks for the conclusion the text supports, its sufficiency, and mental load; the comparison asks for trust and preference.

3.3 Human Evaluation Design

Classification metrics score the answer. To score the trace we use two expert studies on the same 200 clips, one with the video and one without. For SC-CoT the displayed trace is one chain drawn at random from those that agree with the majority vote; for Toulmin it is the grounds, the winning warrant (both warrants when the answer is `not_sure`), and the answer. Procedures, interfaces, and question wording are in Appendix E.

Perception faithfulness asks whether the observations stated in a trace are present in the video (Li et al., 2026; Uppaal et al., 2025); it is distinct from behavioural faithfulness, whether the trace reflects the computation behind the answer (Jacovi & Goldberg, 2020). Prior work measures it on image benchmarks with automatic judges. We measure it on driving video with three expert annotators, who watch each clip and rate each trace as containing no hallucination, a minor one that does not affect the case, or a major one in which the pedestrian does something other than what the trace describes. Clips on which neither trace is rated major proceed to the strength-of-reasoning study.

Strength of reasoning is measured in the second study. In Toulmin's terms the grounds carry the bottom-up content of an argument, the evidence taken from the scene, and the warrant carries the top-down content, the rule the reader would otherwise have to supply (Toulmin, 2003; Wickens et al., 2021). The second study separates the two for a reader who cannot see the video (Figure 3). Each trace is shown in two parts. Part 1 shows the observations alone; Part 2 adds the reason. At each part the reader reports the conclusion the text reaches, how well the text supports it on a 1–7 scale (local sufficiency in the sense of Wachsmuth et al. (2017)), and the mental effort the judgement took (NASA-TLX Mental Demand (Hart & Staveland, 1988)). Part 2 adds a clarity rating and a checkbox for a reason that only repeats the observations; a final screen reveals both full traces and asks for trust and preference. If the reason carries top-down content, the reader's conclusion should shift between the parts; if it restates the observations, it should not. This is the reader-side counterpart of the ablation in Section 3.2: there the warrant is removed from the pipeline, here it is withheld from the reader. To build Part 1, every sentence is labelled by GPT-5.4 as observation, general rule, mixed, or outcome, and the verdict-bearing sentences are removed in code. The two methods are shown in random order with names hidden, and Part 1 is locked before Part 2 appears. The primary outcome, fixed before analysis, is the *conclusion shift rate*, the share of traces whose attributed conclusion changed after the reason was revealed; all other measures are secondary.

Table 1: Classification on 200 PSI windows, overall and by ambiguity stratum. $R_{\texttt{yes}}$ is the safety-critical recall. Full per-class precision, recall, and F1 are in Appendix C. Best in **bold**.

	Overall					HIGH (n=100)		LOW (n=100)
Method	Acc	Macro F1	not_sure rate	$R_{\texttt{yes}}$	$R_{\texttt{no}}$	Acc	$R_{\texttt{yes}}$	Acc
SC-CoT (K=6)	0.270	0.213	0.045	0.116	**0.935**	0.090	0.111	0.450
Agentic-Debate	0.415	0.329	0.780	0.159	0.152	**0.680**	0.333	0.150
Toulmin (ours)	**0.480**	**0.466**	0.205	**0.681**	0.435	0.380	**0.778**	**0.580**
dataset prior not_sure rate			0.385					

4 EXPERIMENTS

Dataset. We evaluate on PSI 2.0 (Jing et al., 2025), where each pedestrian track is labelled by multiple annotators at key frames as `cross`, `not_cross`, or `not_sure`. We slice each track into 15-frame windows and give each window the labels of its most recent key frame. For a window with n_c, n_{nc}, n_{ns} votes of each kind and N votes in total, the aggregated intent $a = (n_c + 0.5\, n_{ns})/N$ summarises annotator agreement on crossing; thresholding it at 0.33 and 0.67 gives the three-class ground truth `no` / `not_sure` / `yes`. The ambiguity score $s = \max(n_{ns}/N,\ 1 - 2|a - 0.5|)$ measures disagreement and is used only for sample selection. We keep windows with at least one `not_sure` vote and at least one decisive vote, and draw 100 windows from a HIGH pool ($s \geq 0.60$) and 100 from a LOW pool ($s < 0.30$), one window per track. The 200 windows have ground truth `no` 46, `not_sure` 85, `yes` 69; the `not_sure` share of 38.5% is the prior we compare abstention rates against. Selection details are in Appendix D.

Models and compute. All three methods use Gemini 3 flash-preview as the base VLM, accessed through the API. In the Toulmin pipeline the six vision-grounded agents run on Gemini and the text-only critic and controller run on GPT-5.4; SC-CoT draws K=6 samples from Gemini at temperature 1.0; Toulmin and Agentic-Debate run at temperature 0.3. Sentence labelling for the strength-of-reasoning study and sub-type classification use GPT-5.4 at temperature 0. No model is fine-tuned and no GPU is used for inference; the pipeline's API usage is about 1.3× that of SC-CoT at K=6.

Evaluation. Classification metrics compare all three methods against the ground truth above: accuracy, macro F1, per-class precision, recall, and F1, and the `not_sure` rate, overall and by ambiguity stratum. Recall on `yes` is the safety-critical number, since a missed crossing is the error a planner cannot recover from. The human studies of Section 3.3 compare Toulmin with SC-CoT on the traces these runs produced.

5 RESULTS

5.1 CLASSIFICATION RESULTS

The classification metrics answer one question before the human studies: whether the structure costs accuracy. It does not (Table 1). Toulmin reaches 48.0% accuracy against 41.5% for Agentic-Debate and 27.0% for SC-CoT, and `yes`-recall of 0.681 against 0.159 and 0.116. The two baselines fail in opposite directions. SC-CoT abstains on 4.5% of windows and answers `no` on most of the rest; Agentic-Debate, with a debate but no warrant to check, abstains on 78.0%. Toulmin abstains on 20.5%, nearest the 38.5% prior, which is what a judge with a concrete condition to verify should do. Its errors run in the conservative direction (`no`-recall 0.435 against SC-CoT's 0.935): where SC-CoT commits to `no`, Toulmin abstains or answers `yes`. Agentic-Debate's higher accuracy on the HIGH stratum (0.680 against 0.380) comes from abstaining on a stratum where most ground truth is `not_sure`; its HIGH `yes`-recall is 0.333. Figure 4 shows the three trace structures on one clip.

5.2 HUMAN EVALUATION

A correct label can come from an argument or from a rationalization. The two studies below read the traces to tell which. Section 1 makes two predictions about a rationalizing trace: its stated

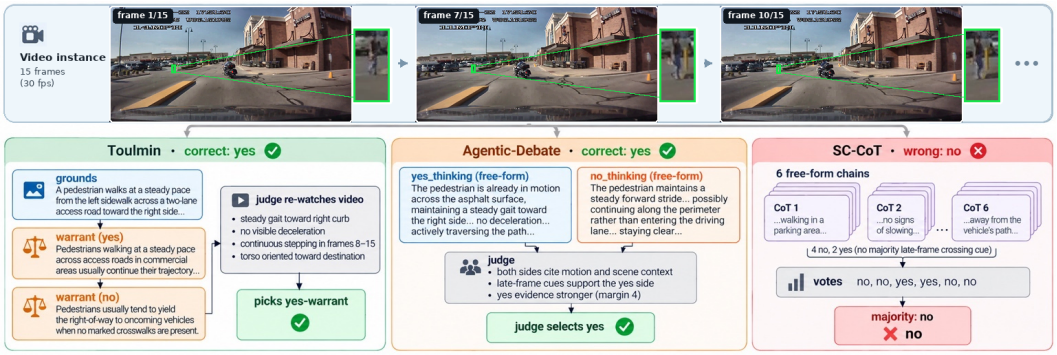

Figure 4: Qualitative example. The reasoning traces of the three methods on one HIGH-ambiguity PSI clip (ground truth `yes`). Toulmin produces three separate components: grounds written before any conclusion, two opposing warrants, and a judge that re-watches the video and selects one warrant. Agentic-Debate produces two free-form arguments and a judge that counts the evidence on each side; no rule is stated or checked. SC-CoT produces six independent chains and takes a majority vote. On this clip the first two answer `yes` and SC-CoT answers `no`. Texts are quoted from the model outputs.

observations should fail a check against the video more often, and revealing its reason should change little, because the observations already carry the conclusion.

5.2.1 PERCEPTION FAITHFULNESS

The first prediction holds. Major hallucinations appear in 51.5% of SC-CoT traces (100 of 194 rating rows) and 31.4% of Toulmin traces (61 of 194), 39% fewer, with the same direction for all three annotators (52.0% against 29.4%, 55.9% against 39.7%, 37.5% against 16.7%). On the 73 clips where only one trace was rated major, it was the SC-CoT trace on 56 (exact McNemar $p = 5.3 \times 10^{-6}$, odds ratio 3.29, 95% CI $[1.89, 6.05]$). Both methods see the same frames through the same base model; what differs is that SC-CoT writes its observations in the same pass as its verdict. Of the 193 clips, 76 had no major rating on either trace and proceeded to the strength-of-reasoning study.

5.2.2 STRENGTH OF REASONING

The second prediction can be tested before any reader is involved. The sentence labelling of Section 3.3 finds that 176 SC-CoT chains (88.0%) contain a sentence that fuses reason and verdict, 24 (12.0%) state no separable reason, and none of the 176 reason clauses states a general rule. Toulmin has none of these by construction. All 76 eligible clips were then rated (45 HIGH, 31 LOW), giving 119 clip–annotator completions and 238 reasoning units; on the 42 clips rated by more than one annotator, agreement on the Part 2 conclusion is 79.5% (Cohen's $\kappa = 0.61$).

The primary outcome separates the methods by an order of magnitude. Revealing the reason changed the attributed conclusion on 37.0% of Toulmin completions and 5.0% of SC-CoT completions (Table 2; odds ratio 11.7, $p = 4.2 \times 10^{-8}$), in both strata and for every annotator. The reason for the gap is visible in Part 1: before any reason is shown, SC-CoT's conclusion is already recoverable from its observations alone in 94.6% of readings, against 63.6% for Toulmin. The readers said the same thing directly, marking the reason as merely repeating the observations on 33.0% of SC-CoT units against 0.8% of Toulmin units. So SC-CoT's reason carries almost nothing its observations did not, and Toulmin's warrant carries enough to change the reader's mind on more than one clip in three.

The rating questions point the other way, and the two findings belong together. SC-CoT is rated better supported, less effortful, and clearer (Table 3), with the largest gap on the observations alone; on the LOW stratum every gap closes once the warrant is revealed, on HIGH it persists and extends to trust and preference (SC-CoT 37 against Toulmin 25). On those same 45 HIGH clips SC-CoT never answered `not_sure` and was correct on 4 (8.9%), Toulmin on 16 (35.6%). Readers recovered SC-CoT's conclusion in 71 of its 78 wrong cases, its clarity and support ratings did not drop when it was wrong, and trust did not track correctness for either method. The trace that reads best on the ambiguous clips is the one that is wrong on nearly all of them.

Table 2: Strength-of-reasoning study on 119 clip–annotator completions (238 reasoning units). The pre-specified primary outcome is the conclusion shift rate: the share of completions in which revealing the reason step changed the conclusion the annotator attributed to the trace. Secondary outcomes are within-method changes from Part 1 (observations only) to Part 2 (observations plus reason). Brackets are 95% confidence intervals.

	Toulmin	SC-CoT
Primary outcome		
Conclusion shift rate (all)	$44/119 = 37.0\%$ [28.8, 45.9]	$6/119 = 5.0\%$ [2.3, 10.6]
HIGH ambiguity	$31/73 = 42.5\%$	$4/73 = 5.5\%$
LOW ambiguity	$13/46 = 28.3\%$	$2/46 = 4.3\%$
Secondary, Part 2 minus Part 1		
Support (1–7)	$+0.87$ [$+0.65, +1.11$]	$+0.21$ [$+0.08, +0.34$]
Mental effort (1–7)	-0.56 [$-0.76, -0.35$]	-0.17 [$-0.30, -0.03$]
Reason "only repeats observations"	$1/119$ (0.8%)	$35/106$ (33.0%)

Table 3: Method effect, Toulmin minus SC-CoT, on the rating questions, from mixed models with annotator as a fixed effect and a random intercept per clip (pooled: additive model with 95% CI; by stratum: interaction model with p). Positive effort means Toulmin was rated more effortful. After the warrant is revealed, the disadvantage is confined to the HIGH stratum.

	Pooled	HIGH (45 clips)		LOW (31 clips)	
Measure	Effect [95% CI]	Effect	p	Effect	p
Part 1 support	-1.07 [$-1.35, -0.78$]	-1.26	< 0.001	-0.76	0.001
Part 1 mental effort	$+0.97$ [$+0.74, +1.21$]	$+1.12$	< 0.001	$+0.74$	< 0.001
Part 2 support	-0.40 [$-0.69, -0.12$]	-0.59	0.001	-0.11	0.64
Part 2 mental effort	$+0.58$ [$+0.35, +0.81$]	$+0.81$	< 0.001	$+0.22$	0.24
Clarity	-0.47 [$-0.72, -0.22$]	-0.71	< 0.001	-0.09	0.67
Trust after reveal	-0.19 [$-0.47, +0.08$]	-0.43	0.017	$+0.17$	0.44

6 DISCUSSION

Rationalization and its cost. The SC-CoT results match the pattern of Section 1: observations that already carry the conclusion, a reason that adds nothing and states no rule, and stated observations that are often not in the video. The reasons SC-CoT does state also cluster: 82% fall in three sub-types that restate the pedestrian's current trajectory, where Toulmin's spread over seven (Figure 5). The three sub-types are separate labels for close to the same statement, that the pedestrian is positioned along the road rather than toward it, so the reasons are more uniform than the count of seven suggests. The first two on their own would be consistent with observations that happen to be sufficient; together with the third, the more plausible reading is that the observations are written to fit the conclusion. This is a claim about the text. We did not intervene on the chains, so we cannot say what the model computed, only that its output has the structure of a rationalization. The Toulmin pipeline removes that structure by changing the order of generation: the grounds are written with no conclusion available, both warrants are stated before either is chosen, and the judge checks two explicit rules against the video. The text shows the difference (grounds that do not reveal the verdict, a warrant that shifts the reader's conclusion), and so do the answers (abstention near the prior, `yes`-recall of 0.681). The cost is in the grounds. Evidence that has not been shaped by a conclusion is harder to read and looks less sufficient on its own. On clear cases the cost disappears once the warrant is revealed; on open cases it persists and extends to trust and preference, which are the cases where SC-CoT's answers are worst.

Fluency and correctness. On the HIGH stratum every presentation measure favours SC-CoT and every outcome measure favours Toulmin. Readers understood 71 of SC-CoT's 78 wrong conclusions, and trust tracked correctness for neither method. Human ratings of reasoning quality collected without ground truth therefore reward presentation, and automated metrics collected without anyone reading the text say nothing about whether the stated evidence is real. Both are needed.

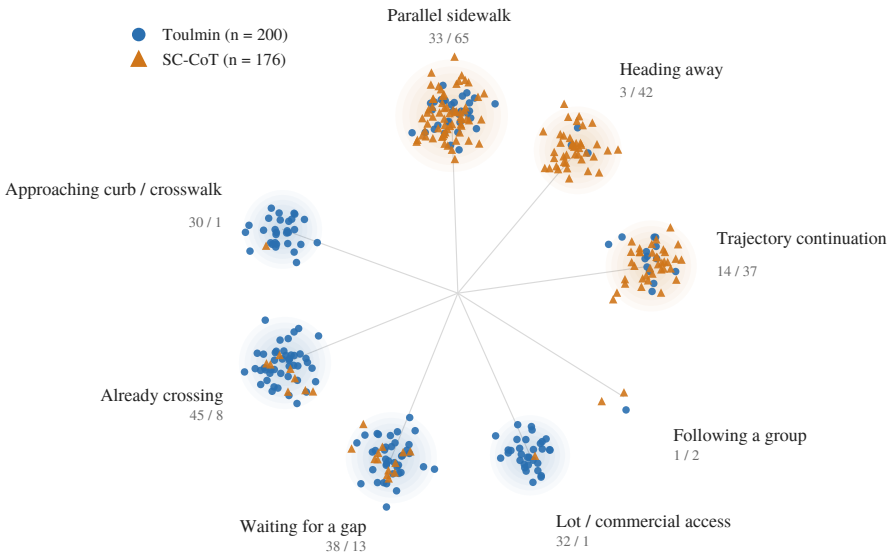

Figure 5: Stated reasons by sub-type, one point per reason (Toulmin circles, n=200; SC-CoT triangles, n=176). Position encodes the assigned sub-type only; placement within a cluster is arbitrary. Counts under each label are Toulmin / SC-CoT. Three sub-types, all descriptions of the pedestrian's current trajectory, hold 82% of SC-CoT's reasons and 58% of Toulmin's (normalised entropy 0.71 against 0.87). Not shown: 11 reasons with no sub-type and 24 SC-CoT chains with no reason.

Limitations. The evaluation uses one dataset and one base VLM, the prompts and warrant taxonomy are hand-written, and the pipeline costs about $1.3\times$ the compute of SC-CoT at latencies too high for real-time use. The strength-of-reasoning study is restricted to clips without a major hallucination, which removes more SC-CoT outputs than Toulmin outputs and so favours SC-CoT. The two studies share three annotators from the research team, and the perception faithfulness interface showed method names. The restatement checkbox cannot be checked for agreement. Clarity, included as a control, favours SC-CoT on ambiguous clips, so a preference for structure as such cannot be ruled out. The sentence labels behind the strength-of-reasoning screens were not validated by a separate human study. We tested whether the Toulmin text has the structure of an argument, not whether the judge is free of post-hoc selection; that requires an intervention on the warrants, left to future work.

7 Conclusion

Free-form chain-of-thought on ambiguous driving video produces text with the structure of a rationalization: its observations already carry the conclusion, its reason adds nothing a reader can use, and half of its traces describe something the video does not show. A Toulmin-structured pipeline that writes the grounds before any conclusion exists and states the warrant as a separate output closes that route by construction without losing accuracy.

We see two uses for this result. The first is in the study of faithfulness. Most work on the topic asks whether a chain reflects the computation behind the answer. We ask a smaller question that can be answered by reading, whether the stated evidence is real and whether the stated reason does any work, and we find that a change in the order of generation moves both. The warrant gives this line of work a unit that can be checked against the input, removed from the pipeline, or withheld from a reader.

The second use is in models that are meant to act in the world. A world model or an embodied agent has to know why things happen, and the why is a rule. Some rules are physical, that a door opens when the latch is released; many are social, that a pedestrian who stops at a curb and looks at the traffic is waiting for a gap. A model that applies such rules without stating them cannot be checked or corrected, and cannot explain its decision to the person who has to trust it. Writing the rule down as a warrant is a small step toward a model whose understanding of the world can be read and tested.

STATEMENT ON THE USE OF AI TOOLS

Large language models were used to assist with writing and editing the manuscript. GPT-5.4 was used to assign sentence-level labels (OBSERVATION, GENERAL_RULE, MIXED, ASSERT_OUTCOME) when constructing the strength-of-reasoning screens; the model returned labels only, deletion was performed in code, and every retained sentence was checked verbatim against its source. GPT-5.4 also assigned each stated reason to a sub-type for Figure 5; the sub-type taxonomy and all analysis decisions were made by the authors. The proposed method itself runs on Gemini 3 flash-preview and GPT-5.4, as described in Section 3. No generative AI tool was used to design the experiments, interpret the results, or write the code.

REFERENCES

Lina Achaji, Julien Barry, Thibault Fouqueray, Julien Moreau, Francois Aioun, and Francois Charpillet. Is attention to bounding boxes all you need for pedestrian action prediction? In *IEEE Intelligent Vehicles Symposium (IV)*, 2022.

Ashutosh Adhikari and Mirella Lapata. Debating for better reasoning in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 15766–15784, 2025.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Maciej Besta, Nils Blach, Ales Kubicek, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proc. AAAI*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

Ege Cakar and Per Ola Kristensson. The argument is the explanation: Structured argumentation for trust in agents. *arXiv preprint arXiv:2510.03442*, 2025.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Charles Corbière, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. DRIVINGVQA: A dataset for interleaved visual chain-of-thought in real-world driving scenarios. In *Findings of the Association for Computational Linguistics: EACL 2026*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proc. ICML*, 2024.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 10764–10799, 2023.

Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, 2017.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1930–1940, 2018.

Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001.

Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati (eds.), *Human Mental Workload*, volume 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4198–4205, 2020.

Taotao Jing, Tina Chen, Renran Tian, Yaobin Chen, Joshua Domeyer, Heishiro Toyoda, Rini Sherony, and Zhengming Ding. PSI: A benchmark for human interpretation and response in traffic interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=yFIpdvQYmx.

Petter Johansson, Lars Hall, Sverker Sikström, and Andreas Olsson. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745):116–119, 2005.

Ziva Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Kamile Lukosiute, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Junxian Li, Xinyue Xu, Sai Ma, Di Zhang, and Sichao Li. Faithful-first reasoning, planning, and acting for multimodal LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6777–6793, 2026.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Qing Lyu, Shreya Havaldar, Adam Stein, et al. Faithful chain-of-thought reasoning. In *Proc. IJCNLP-AACL*, 2023.

Ana-Maria Marcu, Long Chen, Jan Hünermann, et al. LingoQA: Visual question answering for autonomous driving. In *Proc. ECCV*, 2024.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Ming Nie, Renyuan Peng, Chunpeng Wang, et al. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In *Proc. ECCV*, 2024.

Xuefei Ning, Zinan Lin, Zixuan Zhou, et al. Skeleton-of-thought: Prompting LLMs for efficient parallel generation. In *Proc. ICLR*, 2024.

Richard E. Nisbett and Timothy DeCamp Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4658–4664, 2019.

NVIDIA Autonomous Vehicle Research Group, Yan Wang, Wenjie Luo, et al. Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025.

Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

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 (ICCV)*, 2019a.

Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked RNNs. In *British Machine Vision Conference (BMVC)*, 2019b.

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.

Chonghao Sima, Katrin Renz, Kashyap Chitta, et al. DriveLM: Driving with graph visual question answering. In *Proc. ECCV*, 2024.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proc. ICML*, 2024.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017.

Xiaoyu Tian, Junru Gu, Bailin Li, et al. DriveVLM: The convergence of autonomous driving and large vision-language models. In *Proc. CoRL*, 2024.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, Cambridge, UK, updated edition, 2003.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. Journey before destination: On the importance of visual faithfulness in slow thinking. *arXiv preprint arXiv:2512.12218*, 2025.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 176–187, 2017.

Lei Wang, Wanyu Xu, Yihuai Lan, et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proc. ACL*, 2023a.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proc. ACL*, 2024a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22442–22452, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Chain-of-Table: Evolving tables in the reasoning chain for table understanding. In *International Conference on Learning Representations (ICLR)*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Christopher D. Wickens, William S. Helton, Justin G. Hollands, and Simon Banbury. *Engineering Psychology and Human Performance*. Routledge, 5th edition, 2021.

Juncheng Wu, Hardy Chen, Haoqin Tu, Xianfeng Tang, Freda Shi, Hui Liu, Hanqing Lu, Cihang Xie, and Yuyin Zhou. From seeing to thinking: Decoupling perception and reasoning improves post-training of vision-language models. *arXiv preprint arXiv:2605.20177*, 2026.

Zhenhua Xu, Yujia Zhang, Enze Xie, et al. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Proc. NeurIPS*, 2023.

Zhengming Zhang, Renran Tian, and Zhengming Ding. TrEP: Transformer-based evidential prediction for pedestrian intention with uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3534–3542, 2023. doi: 10.1609/aaai.v37i3.25463. URL https://ojs.aaai.org/index.php/AAAI/article/view/25463.

Denny Zhou, Nathanael Schärli, Le Hou, et al. Least-to-most prompting enables complex reasoning in large language models. In *Proc. ICLR*, 2023.

A Baseline Schematics

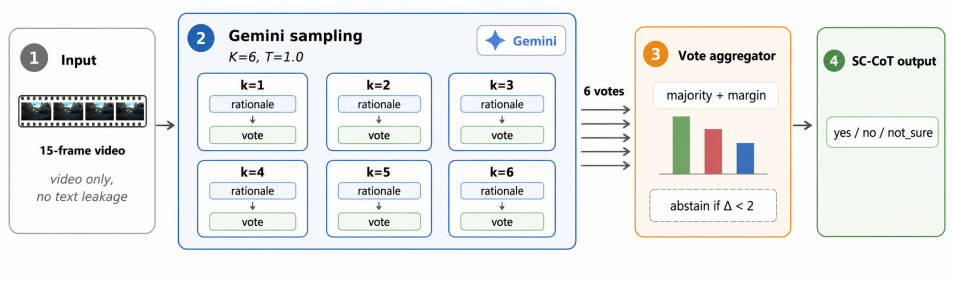

Figure 6: SC-CoT baseline (K=6 free-form chain-of-thought samples, temperature 1.0). Samples are drawn from the same Gemini base model; a vote aggregator returns the majority answer, or `not_sure` when the top class leads the second by fewer than 2 votes.

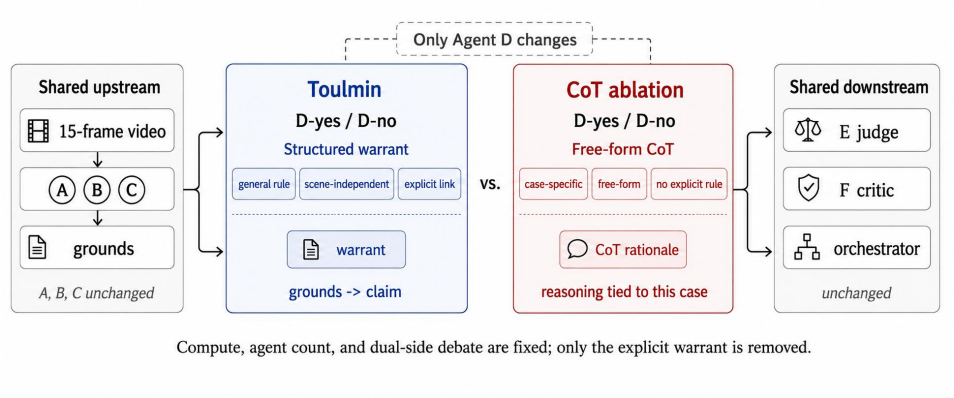

Figure 7: Agentic-Debate ablation. The two reasoning-generator agents (D-yes, D-no) of the full Toulmin pipeline (Figure 2) emit free-form CoT `thinking` and `key_observations` instead of a structured warrant. All other agents are unchanged. Compute, agent count, and dual-side argumentation are held fixed; the only ablated component is the warrant.

B Classification Results in Full

Table 1 gives the classification metrics. Toulmin reaches 48.0% accuracy (96/200) and 0.466 macro F1, against 41.5% / 0.329 for Agentic-Debate and 27.0% / 0.213 for SC-CoT. The gap is largest on the safety-critical `yes` class: of the 69 ground-truth crossings, Toulmin's recall is 0.681 (47), against 0.159 (11) for Agentic-Debate and 0.116 (8) for SC-CoT. The two baselines fail in opposite directions. SC-CoT abstains on 4.5% of windows, far below the 38.5% prior, and answers `no` on 86% of all predictions, including 56 of the 69 true crossings; its recall is 0.935 on `no` and 0.116 on `yes`. Agentic-Debate abstains on 78.0% of windows, including 53 of the 69 true `yes` cases: without a concrete visual target, the two free-form arguments look equally plausible to the judge. Toulmin abstains at 20.5%, nearest the prior; the applicability caveat gives the judge a specific condition to

check, so the abstention triggers fire only when the structured evidence is balanced. Toulmin's errors run in the conservative direction: its `no`-recall (0.435) is well below SC-CoT's, because it abstains or answers `yes` on cases SC-CoT commits to `no`.

By stratum, Toulmin's `yes`-recall on the HIGH pool (0.778) is higher than on the LOW pool (0.667), while Agentic-Debate falls to 0.333 and SC-CoT to 0.111. One number in Table 1 runs the other way: Agentic-Debate's HIGH accuracy of 0.680 exceeds Toulmin's 0.380. It does so by abstaining on 78% of windows on a stratum where most ground truth is `not_sure`; its HIGH `yes`-recall of 0.333 and LOW accuracy of 0.150 show that the number reflects the base rate, not skill. Figure 4 shows the three methods on one sample.

C PER-CLASS CLASSIFICATION METRICS

Table 4: Per-class precision (P), recall (R), and F1. Toulmin is the only method with F1 ≥ 0.37 on all three classes; the recall on `yes` (68.1%) is the safety-critical metric for autonomous driving.

	`no` (n=46)			`not_sure` (n=85)			`yes` (n=69)		
Method	P	R	F1	P	R	F1	P	R	F1
SC-CoT	0.250	**0.935**	0.394	0.333	0.035	0.064	0.421	0.116	0.182
Agentic-Debate	0.269	0.152	0.194	0.417	**0.765**	**0.539**	**0.611**	0.159	0.253
Toulmin (ours)	**0.328**	0.435	0.374	**0.707**	0.341	0.460	0.480	**0.681**	**0.563**

D SAMPLE SELECTION DETAILS

Each PSI 2.0 pedestrian track is sliced into overlapping windows of 15 frames at stride 10 (≈0.5 s at 30 fps). Each window inherits its per-annotator labels from the most recent key frame on or before the window end. Windows are kept only if they have at least one `not_sure` vote and at least one decisive (`cross` or `not_cross`) vote. The filtered windows are bucketed by ambiguity score s into a HIGH pool ($s \geq 0.60$) and a LOW pool ($s < 0.30$); the middle band is discarded so that the two pools are clearly separated. Within each (video, track) pair only the earliest qualifying window is kept, to avoid temporal duplicates. After shuffling with seed 42, the first 100 windows from each pool form the test set. Ground-truth classes are assigned by thresholding the aggregated intent a: `no` ($a \leq 0.33$, n=46), `not_sure` ($0.33 < a < 0.67$, n=85), `yes` ($a \geq 0.67$, n=69).

E HUMAN EVALUATION QUESTIONS

Annotators. Three members of our research team who work on pedestrian crossing behaviour and are familiar with the PSI annotation protocol served as annotators for both studies. They were not told which hypothesis the studies were testing. Seven of the 200 clips are excluded from both studies because the six SC-CoT samples split 3–3 and no single chain could be shown as the method's output.

Perception faithfulness procedure. Figure 8 summarises the perception faithfulness study. The annotator watches the clip with both traces visible and rates each trace on the three-level scale defined in Section 3.3. A clip is eligible for the strength-of-reasoning study when neither of its two traces was rated major.

Strength-of-reasoning procedure. Each clip is shown on five screens. Screens A1 and A2 show one method's trace, first the observations alone (Part 1), then the observations with the reason (Part 2). Screens B1 and B2 do the same for the other method. Screen D reveals both original traces in full, with the sentences hidden on the earlier screens highlighted. Which method is A and which is B is fixed per annotator and clip by hashing the annotator identifier and clip number, and method names are never shown. The video is not shown on screens A and B: the perception faithfulness study has already established whether the observations are true, so this study tests whether the argument works. Part 1 must be submitted before Part 2 appears. For SC-CoT chains that state no separable reason, Part 2 shows a message saying so. Part 1 shows the sentences labelled OBSERVATION; Part 2 adds those labelled GENERAL_RULE and MIXED.

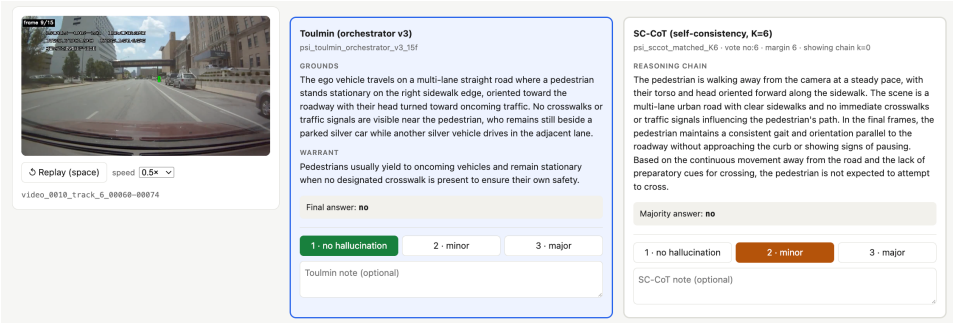

Figure 8: Perception faithfulness study. An expert annotator watches the clip and rates each method's trace as none, minor, or major; clips with no major rating proceed to the strength-of-reasoning study.

Perception faithfulness interface. The screen shows the clip with a replay control on the left and the two traces side by side, labelled "Toulmin" (grounds, warrant, final answer) and "SC-CoT" (reasoning chain, majority answer). Under each trace the annotator selects one of three ratings, "no hallucination", "minor", or "major", with an optional free-text note. *No hallucination*: the description matches the video. *Minor*: the description matches the video overall, with a few slight errors in details that do not matter for the case. *Major*: the description seriously conflicts with the video; the pedestrian does something different from what the trace describes. A final control records whether the clip is usable for the second study (yes / no / unsure), derived from the two ratings.

Strength of reasoning, instruction shown on every screen. "Take the descriptions as true. We already checked these reasonings against the video and removed the ones with major errors. What we want to know is whether the argument works, not whether it got the facts right."

Part 1 (screens A1, B1). The screen shows the heading "Reasoning A — what it observed" followed by the observation sentences.

1. *Based on these observations, what conclusion would you reach?* Options: "The pedestrian will cross in front of the ego vehicle", "The pedestrian will not cross in front of the ego vehicle", "Cannot tell". If "Cannot tell" is selected, a follow-up asks whether the text concludes that the case is uncertain or does not arrive at any conclusion.
2. *Assuming these observations are accurate, do they alone give enough support for concluding [the conclusion selected in Q1]?* Scale 1 (Not at all) to 7 (Completely). The bracketed clause is filled with the annotator's answer to Q1.
3. *How much mental and perceptual activity was required to reach that judgement?* Sub-text: "Include thinking, deciding, and looking things up mentally." Scale 1 (Very low) to 7 (Very high), NASA-TLX Mental Demand.

Footer: "Once submitted this part is locked. You will then see the reason this reasoning gives, and be asked the same questions again."

Part 2 (screens A2, B2). The screen adds the reason sentences under a second heading.

1. *Now, what conclusion does this reasoning reach?* Same options and follow-up as Part 1.
2. *Assuming the observations are accurate, does the reasoning give enough support for concluding [the conclusion selected in Q1]?* Scale 1 (Not at all) to 7 (Completely). Below the scale, a checkbox: "The reason only repeats the observations rather than adding anything new." Sub-text: "Tick this as well as rating above — both are recorded."
3. *How much mental and perceptual activity was required to reach that judgement?* As above.
4. *Is the reasoning easy to follow and is its logical structure clear?* Sub-text: "Only how easy it is to read and follow, not whether you agree with it." Scale 1 (Very hard) to 7 (Very clear).

Footer: "Once submitted this part is locked and cannot be changed." For SC-CoT chains with no separable reason, the panel shows a message that the text gives no reason beyond its observations.

Direct comparison (screen D). Both full traces are shown side by side under the headings "Reasoning A" and "Reasoning B", with the sentences hidden on the earlier screens highlighted. Under each trace the screen shows "You predicted: [Part 2 answer]" and "Model answered: [model answer]".

1. *Based on Reasoning A, how much would you trust the prediction it gives?* Scale 1 (Would not trust) to 7 (Would fully trust).
2. *Based on Reasoning B, how much would you trust the prediction it gives?* Same scale.
3. *If you were the safety driver, which reasoning would you prefer to see for this clip?* Options: "Reasoning A", "Reasoning B", "About the same".