# THE ROLE OF SOCIETY OF THOUGHT IN VISION-LANGUAGE REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reasoning models sometimes produce chain-of-thought traces that look less like a single linear solution and more like a dialogue between internal perspectives. Prior work by Kim et al. (2026) has studied this "Society of Thought" (SoT) behavior in text only models, but it remains unclear whether the same phenomenon appears in vision-language models (VLMs), where reasoning must stay grounded in visual evidence. We analyze how this behavior varies across model families, post-training variants, tasks, difficulty, and answer correctness. Society-like reasoning appears across models, but is most visible in models trained to deliberate and in traces for more difficult benchmark items. In contrast to earlier findings in text only models, higher society scores are associated with lower accuracy, including when questions are separated by difficulty. We further distinguish the presence of multiple voices from whether different perspectives communicate and reach a coherent conclusion. Finally, we use society-based rewards to test whether the observed behavior can be changed through reinforcement learning. The rewards successfully alter the reasoning traces and produce modest performance gains without using answer correctness as a training signal. These results show that SoT is present and steerable in VLM reasoning, while its benefit depends on how the different perspectives are coordinated.

## 1 INTRODUCTION

Recent work by Kim et al. (2026) identifies a phenomenon in reasoning models called the "Society of Thought" (SoT), in which a single model's chain-of-thought trace contains multiple distinct voices or personas that appear to exchange views while solving a task. Instead of prompting separate models to interact, the authors show that dialogue-like exchanges between different perspectives can appear within the reasoning output of one model. They also observe unexpected textual fragments appearing mid-trace, often seemingly unrelated to the task, which may indicate brief shifts in perspective before the model returns to the main line of reasoning. Importantly, they find that the presence and amplification of society-like behavior is correlated with improved model performance. Their analysis, however, is limited to text only reasoning models, leaving open whether similar behavior occurs in vision-language models (VLMs).

The link between society-like behavior and performance is consistent with work on collective human cognition. Woolley et al. (2010) show that group performance depends not only on individual members, but also on how they interact. Bahrami et al. (2010) make a related point in perceptual tasks: two observers can outperform either individual when their abilities are comparable and they communicate effectively, while poor coordination can reduce performance. Applied to VLMs, this suggests that internal multi-perspective reasoning may help only when the different perspectives are coordinated well. The key issue is therefore not just whether diverse viewpoints appear, but whether they interact in a useful way.

Building on these foundations, this work asks whether the society-like reasoning behavior identified by Kim et al. (2026) in text only models also appears in the reasoning traces of VLMs. Human studies on group reasoning suggest that multiple perspectives are useful only when they are coordinated effectively. By analogy, we hypothesize that internal dialogue within reasoning in VLMs may be beneficial when it is structured well, but harmful when it adds noise.



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.

We organize the study around three research questions. **RQ1:** How does society-like reasoning appear across VLMs, model variants, and tasks? **RQ2:** How does this behavior relate to task difficulty and answer correctness, and when does communication between perspectives remain coherent? **RQ3:** Can society-like reasoning be directly changed through judge-based reinforcement learning, and how does this affect reasoning traces and performance? On RQ3, we confirm that Society of Thought can be steered and can contribute to performance gains, though the gains remain modest and do not establish society-based rewards as a generally effective method for improving task accuracy.

This work makes three contributions: a cross-model study of society-like reasoning in VLMs; evidence connecting this behavior to task difficulty and answer correctness, while separating visible personas from communication that reaches a coherent conclusion; and a demonstration, through controlled training, that the measured behavior and the resulting chain-of-thought can be changed.

## 2 Related Work

**Reasoning beyond one chain.** The original chain-of-thought work showed that prompting a model to produce intermediate reasoning steps improves performance on math and logical tasks in LLMs (Wei et al., 2022). Wang et al. (2023) extended this by sampling multiple reasoning paths and taking the majority answer, framing the gain as self-consistency. This can also be seen as an early form of multi-perspective reasoning, where different sampled trajectories provide alternative routes to the answer. Synergy-of-Thoughts (Shang et al., 2024) makes this interaction explicit across models, treating separate thoughts as agents that exchange information. Across these methods, performance improves by generating, selecting, or coordinating multiple reasoning paths rather than relying on a single chain.

**CoT transfers imperfectly to VLMs.** CoT has also been adapted to VLMs, although prior work focuses mainly on what the trace contains rather than how it is organized. Early studies use prompt tuning (Ge et al., 2023) or describe the image before answering (Wu et al., 2023). However, gains from text LLMs do not automatically transfer to visual tasks, possibly because training lacks step-by-step visual reasoning data (Nagar et al., 2024). Later methods link reasoning steps to image regions (Shao et al., 2024a), or combine distilled CoT supervision with preference optimization (Zhang et al., 2025a).

Unlike methods that create several paths or agents, our work studies the interaction that appears within one generated trace. As these traces become longer and more common, the reasoning itself becomes an object of study rather than only a tool for producing the answer.

**Trace signals can mislead.** Hwang et al. (2026) examine trace-level signals such as "wait". Their analysis shows that a well-placed "wait" in DeepSeek-R1-32B can coincide with a jump in the probability of the correct answer, suggesting genuine reconsideration, whereas in other models the same token is overused without producing the same corrective effect. They also find that tokens such as "I", "Therefore", "Let", "Now", and "So" are more associated with correct reasoning, while tokens such as "Wait", "Alternatively", and "But" are more associated with incorrect answers. Their results suggest that surface markers of reflection are not enough; what matters is whether such markers correspond to useful changes in the model's reasoning.

**The evidence on shifts is mixed.** d'Aliberti & Ribeiro (2026) take up this question directly and arrive at a skeptical conclusion, finding that apparent reasoning shifts are rare, do not reliably increase with training, and seldom improve accuracy. Their results challenge the idea that mid-trace turns such as reconsideration or sudden insight should automatically be treated as meaningful reasoning events. Much of their evidence, however, comes from instruction-tuned text only models, leaving open whether the same conclusion holds for reasoning-focused variants and VLMs.

**Society-like thinking.** This token and shift-level skepticism is challenged by Kim et al. (2026), who find that reasoning models show multi-perspective shifts far more often than instruction-tuned models, even after controlling for trace length. They call this broader effect "Society of Thought", where reasoning traces can contain multiple voices, personas, or perspectives that interact in dialogue-like ways within a single model's output. They also find that harder problems trigger more of this behavior and that usefulness depends on whether these perspectives are reconciled into a coherent

conclusion. While they show that society-like traces can be encouraged through internal steering and supervised fine-tuning, they do not use the behavior itself as a training reward.

**Steering reasoning through reinforcement learning.** Reinforcement learning provides a way to test whether trace behavior can be shaped directly. GRPO (Shao et al., 2024b) adapts PPO-style optimization (Schulman et al., 2017) to group-based reasoning updates, while GDPO (Liu et al., 2026) independently normalizes multiple reward channels, allowing task accuracy and trace-level behavior to be combined without one dominating because of its scale. Gandhi et al. (2025) show that models can be steered toward behaviors such as verification, backtracking, and subgoal setting, while other methods use self-certainty, latent structure, or process-level feedback (Zhao et al., 2025; Zhang et al., 2026; 2025b). In VLMs, however, longer reasoning can weaken visual grounding; models remain vulnerable to misleading traces, visual forgetting, and drift toward language priors (Zhao et al., 2026; Tian et al., 2026; Xu et al., 2026).

Together, this literature leaves three questions open. It is unclear whether SoT appears in VLMs, whether its visible structure helps or instead marks uncertainty and drift, and whether this behavior can be changed directly through a judge reward. We use reinforcement learning to answer the last question as an intervention.

# 3 METHOD

## 3.1 MEASURING SOCIETY-LIKE TRACE STRUCTURE

To investigate *RQ1*, whether and how SoT appears, we collect reasoning traces across the selected models and benchmarks and analyze them using a judge model.

For each trace the judge assesses three aspects. **Perspective Multiplicity** asks whether the trace is written by a single narrator or can be interpreted as multiple narrators, e.g., through plural self-reference ("we"), sudden shifts in stance, dialogue-like exchanges, or questions asked and answered within the trace. **Viewpoint Conflict** asks whether there is visible disagreement about choosing an answer within the trace. **Role/Persona Separation** asks whether distinct internal roles are visible and interacting.

Each aspect is scored on a 0–3 scale; the three scores are summed to form the *society score* $\in [0, 9]$, which is then normalized to $[0, 1]$:

$$r_{\mathrm{S}}^{\mathrm{v1}} \;=\; \frac{1}{9}\left(s_{\text{multiplicity}} + s_{\text{conflict}} + s_{\text{role}}\right). \tag{1}$$

To address *RQ2*, the implications of SoT, we bucket responses by society score and check whether higher buckets have higher accuracy across models and benchmarks.

We acknowledge that distinguishing single voice traces from multi-persona traces is inherently ambiguous and has no definitive criteria. During manual inspection, we therefore used a consistent working rule: could this trace reasonably be read as dialogue, or is it better treated as a single voice? We labeled a stratified sample of 100 traces under this rule and used it to calibrate the judge prompt.

The primary judge is `Qwen3-30B-A3B-Instruct-2507-FP8` (Team, 2025), a 30B Mixture-of-Experts model. The main analysis reports the mean and standard deviation across three judge runs with different decoding seeds. The judge receives the question, correct answer, and reasoning trace, following the original evaluation protocol. The correct answer provides context for interpreting disagreement, unsupported alternatives, and reasoning that drifts away from the task. Because access to this answer could bias a judge intended to measure trace structure, we repeat the evaluation without ground truth as an additional sensitivity check. The broad conclusions remain consistent, although some model-level differences are discussed in Appendix A.3. We therefore retain the original scoring for the main analysis.

## 3.2 COMMUNICATION AND CLOSURE

The $v_1$ analysis addresses the first part of *RQ2*, but visible persona structure alone does not show whether multiple perspectives interact productively. A trace may contain conflict and role shifts yet

remain unresolved, while a single narrator may compare alternatives and reach a coherent conclusion.

We therefore introduce the *communication* variant (or $v_2$) as a communication quality rubric. It keeps the idea that reasoning can contain several perspectives, but shifts the target from detecting personas to evaluating how those perspectives interact.

The $v_2$ rubric replaces the three axes of $v_1$ with eight axes organized into three groups: five productive axes, one gate axis, and two failure axes. **Question Answering** tracks whether the trace poses sub-questions and answers them internally; **Perspective Shift**, whether it changes approach or revises an earlier line of reasoning; **Perspective Conflict**, whether it contains a real disagreement between alternatives rather than a list of options; **Ask–Give Balance**, whether the trace both raises questions and responds to them rather than only asking or only asserting; and **Role Reciprocity**, whether corrective and constructive moves both appear and engage with each other. **Reconciliation** gates productive credit by checking whether competing perspectives are resolved into one coherent conclusion. The two failure axes, **Unresolved Drift** (contradictions remain to the end, or the final answer does not follow from the deliberation) and **Ungrounded Persona Split** (declared roles or voices that never meaningfully interact), flag degenerate traces.

The first five axes form the productive term, reconciliation gates it, and the final two axes are subtracted as failure modes. The $v_2$ communication score is:

$$r_{\mathrm{S}}^{\mathrm{v2}} = \underbrace{\frac{s_{\mathrm{rec}}}{3}}_{\text{reconciliation gate}} \cdot \underbrace{\frac{1}{15}\sum_{c\in\mathcal{P}} s_c}_{\text{productive axes}} - \underbrace{\frac{1}{6}\sum_{c\in\mathcal{F}} s_c}_{\text{failure axes}}, \tag{2}$$

where $\mathcal{P}$ contains the five productive axes, $\mathcal{F}$ contains the two failure axes, and $s_{\mathrm{rec}}$ is the reconciliation score. Since productive axes sum to $[0, 15]$, failure axes sum to $[0, 6]$, and the reconciliation gate lies in $[0, 1]$, the final score is bounded by $r_{\mathrm{S}}^{\mathrm{v2}} \in [-1, +1]$.

This formulation has three effects. First, a trace receives no productive credit without reconciliation; if $s_{\mathrm{rec}} = 0$, the productive term is muted. Second, failures are unconditional, so drift and empty persona splits can produce negative scores even when some productive markers are present. Third, the wider $[-1, +1]$ range separates traces that reconcile competing views from traces that only contain multi-perspective phrasing without closure, a distinction $v_1$ cannot make.

For analysis, negative scores indicate that unresolved drift or an ungrounded persona split is larger than the gated productive term. A score of zero often means that reconciliation is absent, while high positive scores require both productive interaction and closure. The judge model is unchanged from $v_1$; only the prompt and aggregation rule differ. This second score therefore tests whether comparing alternatives and closing on an answer is more useful than the visible presence of several personas.

### 3.3 Diagnostic Intervention

To address ***RQ3***, the steerability of SoT, we use reinforcement learning as a diagnostic intervention.

We use the GDPO implementation of Liu et al. (2026), but activate only one reward channel at a time. The *rule-based* accuracy reward $r_{\mathrm{A}}$ extracts the final answer with a regex and grades it against the ground truth. The *judge-based* society reward $r_{\mathrm{S}}$ scores the reasoning trace with a frozen LLM judge under the $v_1$ rubric.

The $v_1$ rubric carries over from the analysis phase with one adjustment. The judge for the society score is no longer given the ground truth answer. Providing the correct answer during training to the judge could leak a direct correctness signal into the society channel, conflating trace style with answer quality. The *interrupted_trace* and *unexpected_fragment* flags are also dropped, as these were offline diagnostics not needed during training.

Due to the judge being queried synchronously during training, inference latency and memory become practical constraints. The offline judge, `Qwen3-30B-A3B-Instruct-2507-FP8`, is too large to host alongside the policy. We therefore switch to `Qwen3.5-4B`, a smaller model from the same family that fits on one GPU.

Accuracy reward is an intentionally advantaged reference because it uses the ground truth. We also include a random reward control, which assigns each rollout a pseudorandom value sampled

uniformly from $[-1, 1]$. It tests whether changes arise merely from applying noisy RL updates rather than from the $v_1$ signal. This intervention can show that the measured SoT is trainable.

# 4 Experimental Setup

## 4.1 Observational Study

We evaluate five models drawn from three independently developed families: Qwen3-VL (Alibaba), GLM-V (Zhipu AI & Tsinghua University), and Pixtral (Mistral). **Qwen3-VL-4B-Instruct** is a dense 4B vision-language model (Bai et al., 2025), fine-tuned for direct instruction following without explicit deliberation, while **Qwen3-VL-4B-Thinking** shares the same backbone but is post-trained with a multi-stage reasoning pipeline. **Qwen3-VL-30B-A3B-Instruct** is a Mixture-of-Experts variant ($\sim$30B total, $\sim$3.3B activated per token) trained under the same Instruct-style objective as the 4B model. **GLM-4.6V-Flash** is Zhipu AI's 9B vision-language model (Team et al., 2025), post-trained with supervised fine-tuning and reinforcement learning. **Pixtral-12B-2409** is Mistral's 12B dense vision-language model (Agrawal et al., 2024), instruction-tuned with the native Mistral chat template.

Each model is evaluated on three benchmarks varying in reasoning difficulty and reliance on visual input, yielding 6760 prompts per model. **MMMU** ($n = 900$) uses the validation split, a college-level multimodal multiple-choice benchmark spanning art, business, health, humanities, science, and technology (Yue et al., 2024), with the reasoning-tagged template requiring free-form reasoning before a final letter answer. **MMBench** ($n = 4329$) uses the English dev split, containing image-based multiple-choice questions on perception, attribute reasoning, logic, and OCR (Liu et al., 2025). **MMLU** ($n = 1531$) uses the FLAN-CoT zero-shot split, covering 57 academic subjects across STEM, humanities, social sciences, and practical domains (Hendrycks et al., 2021); as a text only control, behavior observed here cannot be attributed to visual input.

For difficulty analysis, MMMU supplies native Easy/Medium/Hard labels, and we also use empirical item difficulty $1 - \bar{a}$, where $\bar{a}$ is mean accuracy across the five models. Generation and extraction use lmms-eval library (Zhang et al., 2024) for reproducibility.

## 4.2 Intervention Study

We use three training datasets. **Geometry3k** (Geo3k) contains geometry word problems paired with diagram images (Lu et al., 2021), with open-ended numeric answers extracted via `\boxed{}` ($n = 2101$ train, $n = 601$ test). **ChartQA** pairs questions with chart images and mostly numeric answers (Masry et al., 2022); we subsample to $n = 2101$ train and $n = 2500$ test to match the Geo3k training budget. **GSM8K** consists of text only math word problems with school-level numeric answers (Cobbe et al., 2021); we subsample to $n = 2101$ train and use the full $n = 1319$ test split.

We train the Qwen3-VL-4B Instruct and Thinking policies using verl (Sheng et al., 2025) with GDPO (Liu et al., 2026) as the advantage estimator. Rollout is set to $8$ in all experiments, with sampling temperature $1.0$ during training. After each rollout pass, the policy is updated with PPO using AdamW, a small KL-to-reference loss against the frozen base policy, and no critic. Each job uses one node with $4$ A100 GPUs. Prompts are capped at $1024$ tokens and rollouts at $3072$ tokens, giving a $4096$ token training context.

# 5 Results

We begin with contrasting qualitative examples to make the society score interpretable. Figure 1 shows a low score, concise VLM trace, while Figure 2 shows a high score trace with repeated changes of mind and explicit role shifts. Together they illustrate the visible difference measured by $v_1$.

**Society-Like Traces Are Observable but Uneven.** Across the three judge runs, the highest society $v_1$ score has `Qwen3-VL-4B-Thinking`, $0.299 \pm 0.023$, followed by `GLM-4.6V-Flash` at $0.245 \pm 0.020$. The model ordering is identical across all three runs, showing that the relative comparison is stable. Society-like traces therefore appear across more than one model family. Within



**Image input.**
MMBench id = 1600

**Judge evidence**

`perspective_multiplicity = 1 (others = 0)`

"The trace uses a **third-person analytical tone throughout**, but includes a subtle internal dialogue-like structure where the narrator considers the action, evaluates options, and then **qualifies the final choice with a caveat**: 'even though the person is not actively doing a situp.' This suggests a self-questioning turn that **borders on internal dialogue, though not explicitly plural**."

**Qwen3-VL-30B-A3B-Instruct / MMBench / society $v_1$ = 0.11 / correct = 1**

<analysis>
The image shows a person sitting at the edge of a pool with their legs extended and feet resting on a towel. They are not actively moving through the water, nor are they in the process of jumping or sliding. The person appears to be in a seated position, possibly preparing to enter the water or resting. This posture does not match the actions of water sliding, swimming backstroke, or jumping into the pool. The most fitting description of the action is sitting, which could be related to a situp, but the person is not performing a situp motion. However, among the given options, "**situp**" is the closest to the action depicted, as it involves a seated position, **even though the person is not actively doing a situp.**
</analysis>
<answer>D. situp</answer>

Figure 1: Low-society $v_1$ trace from `Qwen3-VL-30B-A3B-Instruct` scored by the judge. The trace is a single coherent analytical voice that enumerates options sequentially and chooses a correct answer (D), only *perspective_multiplicity* fires at all.

**Judge evidence**

`perspective_multiplicity = 3`

"The trace engages in **dialogue-like self-questioning** ('Hmm, option A claims… But I remember the opposite was true: German loans were actually larger. Wait, let me double-check…'). This indicates **multiple internal voices interacting**, such as a doubter, a verifier, and a synthesizer."

`viewpoint_conflict = 3`

"The trace **explicitly contradicts itself**: it initially considers option A as plausible, then rejects it based on new information, and later reaffirms B as correct. The internal contradiction ('**But I remember the opposite was true…**') shows a clear conflict between competing viewpoints within the same reasoning stream."

`role_separation = 3`

"Distinct roles are evident: a ***proposer*** (initially favouring A), a ***critic*** (questioning that view), a ***verifier*** (checking sources like Wilson Library archives), and a ***synthesizer*** (drawing conclusions from evidence). The flow shows **handoff between these roles**: 'Hmm…' (proposer/critic), 'Pauses to verify sources' (verifier), 'Final verification' (synthesizer)."

**Qwen3-VL-4B-Thinking / MMLU us_foreign_policy / society $v_1$ = 1.00 / correct = 0**

[proposer] Hmm, option A claims US banks funded the Allies more than Germany/Austria-Hungary.
[critic] **But I remember the opposite was true**: German loans were actually larger. **Wait, let me double-check** the timeline… Yes, in 1914–1915, German banks like Kreditverein were aggressively financing German war efforts through US banks.
[proposer] Option B says US banks funded Germany more — this seems correct based on historical records. But wait, the question is about what "impinged" on neutrality. If US banks were pro-German, that would definitely undermine neutrality.
[critic] Option C is ridiculous — Wilson was famously anti-German. Option D is tempting if the facts are messy, but Option B feels solid.
[verifier] *Pauses to verify sources.* **Yep, the Wilson Library archives confirm US banks lent $2.7 billion to Germany by 1917** (mostly through the German government's "German-American Bonds"). Meanwhile, Allied financing was smaller.
[synthesizer] This created huge pressure for the US to side with Germany, but the US never did — so why did this "impinge" on neutrality? Ah! Because if banks were openly pro-German, it would make neutrality look like hypocrisy.
[verifier] *Final verification.* I recall the "War Loan" debates in Congress — bills to support Germany were blocked, but the lending itself existed. **Yes, Option B is historically accurate.**

Figure 2: Trace with high $v_1$ score from `Qwen3-VL-4B-Thinking` on MMLU (text only benchmark). All three $v_1$ axes score the maximum. The trace exhibits a clear multi-persona behavior, yet the model commits to the wrong answer, because its "verifier" fabricates a source citation that confirms the incorrect hypothesis. Bracketed role labels added for readability; not present in the original trace.

Qwen, the contrast is strong, `Qwen3-VL-4B-Instruct` scores less than half of the thinking variant, suggesting that post-training focused on reasoning makes multi-voice behavior more visible, while later instruction tuning suppresses it. The dense `Qwen3-VL-4B` variants also score above `Qwen-MoE`, which is lowest at $0.044 \pm 0.005$ despite being larger. Its traces are short and decisive, consistent with stronger instruction following. This suggests that society-like reasoning is not mainly a scale effect; it may be a behavior that training for reasoning exposes and instruction tuning reduces.

**Difficulty Triggers Society Structure.** To identify when visible society structure occurs, we assign each item an empirical difficulty $1 - \bar{a}$, where $\bar{a}$ is its mean accuracy across the five base models.

Table 1: Accuracy and society $v_1$ across the three benchmarks. Society scores report mean $\pm$ standard deviation over three judge runs. Higher $v_1$ scores occur in models that expose more deliberative reasoning.

| Model | Accuracy | v1 |
|---|---|---|
| Qwen3-VL-4B-Instruct | 0.741 | $0.134 \pm 0.006$ |
| Qwen3-VL-4B-Thinking | 0.714 | $0.299 \pm 0.023$ |
| Qwen3-VL-30B-A3B-Instruct | 0.799 | $0.044 \pm 0.005$ |
| Pixtral-12B-2409 | 0.631 | $0.100 \pm 0.014$ |
| GLM-4.6V-Flash | 0.689 | $0.245 \pm 0.020$ |

Table 2: Mean empirical item difficulty $1 - \bar{a}$ within each $v_1$ bucket, reported as mean $\pm$ standard deviation across three judge runs. Values closer to 0 indicate easier items and values closer to 1 indicate harder items. Higher $v_1$ scores occur on harder items.

| | $v_1 = 0$ | $(0, .33]$ | $(.33, .67]$ | $> .67$ |
|---|---|---|---|---|
| Difficulty | $0.120 \pm .008$ | $0.347 \pm .010$ | $0.452 \pm .012$ | $0.563 \pm .011$ |

A value near 0 means that nearly all models answered the item correctly, indicating an easier item, whereas a value near 1 means that nearly all models answered incorrectly, indicating a harder item. Table 2 shows a monotonic relationship: mean difficulty rises from $0.120$ in the $v_1 = 0$ monologue bucket to $0.563$ in the high-society bucket. Thus, higher $v_1$ scores tend to occur on more difficult items.

**Higher v1 Predicts Errors.** The left panel of Figure 3 shows that accuracy falls as traces move from the $v_1 = 0$ monologue bucket to the $v_1 > 0.67$ high-society bucket. Gray lines represent individual models, while the blue line shows their macro-average. Therefore, $v_1$ identifies traces with more visible voices, but these traces are usually less accurate. As an additional check, we stratify MMMU by its native *topic_difficulty* labels. The decline in accuracy remains within each difficulty level (Appendix Table 7), showing that item difficulty alone does not explain the relationship. In this setting, the broad society signal looks more like uncertainty or drift than productive deliberation.

**Communication Depends on Reconciliation.** The right panel of Figure 3 gives a different pattern. Traces with $v_2 = 0$ can remain accurate when they solve the task procedurally without setting up competing perspectives. Negative and weakly positive traces are less accurate, while scores above $0.5$ recover when alternatives are compared and reconciled. Communication quality therefore depends on reaching closure, not simply on adding more perspectives. The $v_1$ rubric remains closer to the original society idea because it detects visible voices, whereas $v_2$ measures broader decision making that can occur within a single narrator.

**RL Confirms Trainability, Not Competitiveness.** The observational analysis gives a clear intervention direction: because higher $v_1$ is associated with lower accuracy, we use $v_1^-$ to discourage the measured society structure. Figure 4 compares the resulting shifts in evaluation accuracy and $v_1$ judge scores against the base, accuracy, and random controls. The $v_1^-$ reward improves accuracy in five of the six model–dataset cells, with `Qwen3-VL-4B-Instruct` on GSM8K as the only exception, while reducing the society score in every cell. Its mean accuracy increases from $0.766$ to $0.789$, while mean $v_1$ falls from $0.051$ to $0.019$. Interestingly, accuracy training produces a broadly similar reduction in $v_1$, despite not directly rewarding trace style. Together, these results suggest that both rewards can lead to more decisive reasoning traces. The random control shows no consistent direction and remains close to base at $0.769$. Its relatively strong GSM8K results may reflect ceiling effects, since the base models already exceed $0.91$ accuracy on that benchmark. The accuracy reward reaches $0.833$, but it has direct access to the ground truth while $v_1^-$ does not.

Figure 5 illustrates the example of reasoning pattern produced by $v_1^-$. The base policy follows an incorrect geometric relation and then checks it without recovering. Accuracy training and $v_1^-$ both reach the correct answer, but $v_1^-$ produces the shortest trace and removes the visible doubt markers.

**Where Do the Accuracy Gains Come From?** A possible explanation is trace compression. The final column of Figure 4 shows that $v_1^-$ reduces mean normalized response length for both models.

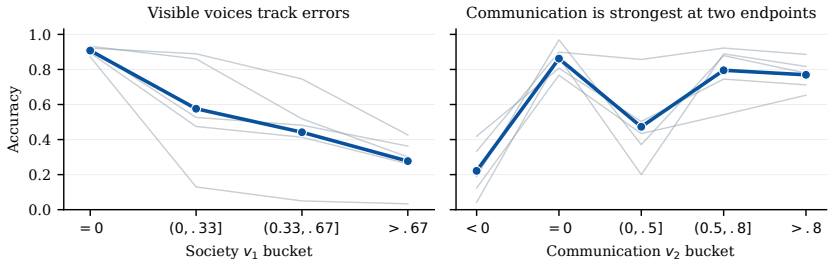



Figure 3: Accuracy by $v_1$ society and $v_2$ communication bucket. Gray lines show individual models and blue lines their macro-average. Accuracy declines as visible society structure increases under $v_1$. Under $v_2$, procedural traces at zero and clearly reconciled positive traces are more accurate than negative or weakly positive traces.

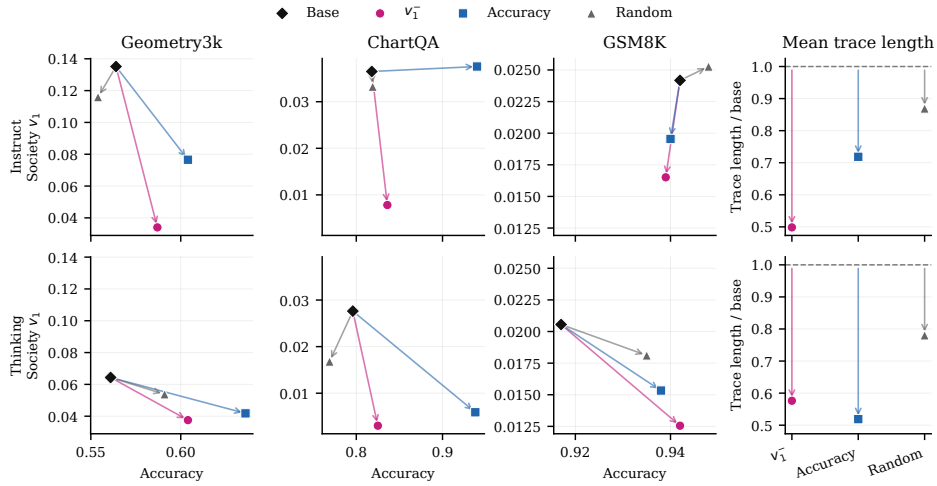



Figure 4: Accuracy, society $v_1$, and response length after training. The first three columns show evaluation accuracy and $v_1$ for each dataset; the final column shows response length normalised by the corresponding base and averaged equally across datasets. Each arrow starts at the base value and ends at a trained checkpoint. The $v_1^-$ reward consistently lowers the measured society score, while the random control tests whether comparable movement follows from RL updates without a task or judge signal.

As illustrated in Figure 5, the lower SoT style also uses fewer markers of doubt, perspective shifts, and repeated checks, which can prevent reasoning loops and may contribute to the accuracy gains. Hallucination results are inconclusive and are reported in Appendix Table 12.

Additional appendix results report per seed judge robustness, high-$v_1$ frequencies, per model difficulty and $v_2$ breakdowns, full intervention results, and further $v_2$ trajectories and reward ablations (Tables 3, 8, 6, 9, and 10; Appendices C.2 and C.3).

## 6 Discussion

**RQ1: How does society-like reasoning appear across VLMs?** Society-like reasoning is observable across the evaluated model families, but its frequency varies substantially. It is most visible in models that externalize longer deliberation and less visible in the instruction-tuned and more concise models. Society-like traces also occur on the text only MMLU subset, so their presence cannot be attributed only to visual inputs. These results instead suggest that post-training and the extent of externalized reasoning are more important than model size or modality.

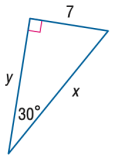



**Image input.**
Geo3k id = 352
(find $x$, GT = 14)

**Base / 1 091 chars / 2 doubt markers / correct = 0**

[...] so $y = 7/2 = 3.5$, $x = 3.5 \cdot \sqrt{3}$.
**But wait, let me double check.** In a 30-60-90 triangle, the side opposite the 30° angle is half the hypotenuse. So yes, $y = 7/2 = 3.5$. The side adjacent is $y\sqrt{3} = (7/2)\sqrt{3}$. [...]
**Final answer:** $\boxed{\frac{7\sqrt{3}}{2}}$ (GT = 14)

**Accuracy / 1 056 chars / 1 doubt markers / correct = 1**

[...] so the hypotenuse is $2 \cdot 7 = 14$, i.e. $x = 14$.
**Let me double check:** the side opposite the 30° angle is 7, so hypotenuse $x = 2 \cdot 7 = 14$. That matches the 30-60-90 ratios. [...]
**Final answer:** $\boxed{14}$

**$v_1^-$ / 567 chars / 0 doubt markers / correct = 1**

[...] In a 30-60-90 triangle, **the side opposite the 30° angle is** $x/2$, where $x$ is the hypotenuse. Given that the side opposite the 30° angle is 7, $7 = x/2 \Rightarrow x = 14$.
**Final answer:** $\boxed{14}$ (no repeated checking, one step to commit)

Figure 5: Same Geo3k question under the base policy, accuracy reward, and $v_1^-$ reward for `Qwen3-VL-4B-Instruct`. The $v_1^-$ trace is shortest, removes repeated checking, and commits directly to the correct relation.

**RQ2: How does this behavior relate to difficulty, correctness, and communication?** Higher v1 scores occur more often on difficult items and are consistently associated with lower accuracy. These two observations are compatible: harder questions may elicit more visible internal conflict, but difficulty does not fully explain the resulting error pattern. The v1 signal therefore appears to identify traces containing uncertainty, repeated checking, or drift, rather than productive multi-persona reasoning by itself.

Communication quality shows a non-monotonic relationship with accuracy. Procedural traces with $v_2 = 0$ can be accurate, performance declines for weak or incomplete deliberation, and recovers when alternatives are clearly reconciled. Thus, successful reasoning depends more on reaching closure than on displaying multiple voices.

**RQ3: Can society-like reasoning be changed through reinforcement learning?** Yes. Penalizing $v_1$ consistently reduces the measured society structure and generally improves accuracy despite having no access to answer correctness. However, accuracy training produces a similar reduction in $v_1$, suggesting that both rewards may lead to shorter, more decisive reasoning. The intervention therefore establishes trainability, but not that reducing SoT specifically causes the accuracy gains or provides a competitive general-purpose objective.

**Judge Reliability and Limitations** This work follows one operational definition of SoT, calibrated against human annotations. The judge therefore measures society structure under this definition rather than establishing a universal boundary for the phenomenon.

A further limitation is that this study identifies when society-like reasoning appears, but not where it originates. Larger post-training experiments could help determine whether it arises from decoding uncertainty, learned styles in post-training data, or rewards for extended deliberation.

**Conclusion.** Society-like reasoning appears across VLM families, especially when models externalize longer deliberation. Its visible multi-persona form is associated with difficult and error-prone traces, whereas useful comparison depends on reconciliation and closure. Reinforcement learning can reduce this structure and often improves accuracy, but the mechanism behind that improvement remains open.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with document restructuring, code assistance, and language editing. The authors verified the analysis, citations and numerical results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

Code and configuration files for reproducing the experiments are available at: https://anonymous.4open.science/r/society-of-thought-in-vlm-85B9.

Additional implementation details are provided in Appendix D.

## REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Singh Chaplot, Jessica Chudnovsky, Saurabh Garg, Théophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Timothée Lacroix, Guillaume Lample, Diego de Las Casas, Thibaut Lavril, Teven Le Scao, Andy Desman Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Reddy Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, and Thomas Wang. Pixtral 12b. *ArXiv*, abs/2410.07073, 2024. URL https://api.semanticscholar.org/CorpusID:273229118.

Bahador Bahrami, Karsten Olsen, Peter E. Latham, Andreas Roepstorff, Geraint Rees, and Chris D. Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010. doi: 10.1126/science.1185718. URL https://www.science.org/doi/abs/10.1126/science.1185718.

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, 2025. URL https://arxiv.org/abs/2511.21631.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Liv G d'Aliberti and Manoel Horta Ribeiro. The illusion of insight in reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 924–966, 2026.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars, 2025. URL https://arxiv.org/abs/2503.01307.

Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of thought prompt tuning in vision language models, 2023. URL https://arxiv.org/abs/2304.07919.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Jaehui Hwang, Byeongho Heo, Sangdoo Yun, and Dongyoon Han. Oops, wait: Discourse tokens matter in reasoning model, 2026. URL https://arxiv.org/abs/2601.17421.

Junsol Kim, Shiyang Lai, Nino Scherrer, Blaise Agüera y Arcas, and James Evans. Reasoning models generate societies of thought, 2026. URL https://arxiv.org/abs/2601.10825.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 216–233, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72658-3.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6774–6786, 2021.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Yu Shang, Yu Li, Fengli Xu, and Yong Li. Synergy-of-thoughts: Eliciting efficient reasoning in hybrid language models, 2024. URL https://arxiv.org/abs/2402.02563.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0275. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0ff38d72a2e0aa6dbe42de83a17b2223-Paper-Datasets_and_Benchmarks_Track.pdf.

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, 2024b. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi,

Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 133928–133965, 2026.

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, 2023. URL https://arxiv.org/abs/2203.11171.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010. doi: 10.1126/science.1193147. URL https://www.science.org/doi/abs/10.1126/science.1193147.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C. Gee, and Yixin Nie. The role of chain-of-thought in complex vision-language reasoning task, 2023. URL https://arxiv.org/abs/2311.09193.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. URL https://arxiv.org/abs/2407.12772.

Nonghai Zhang, Weitao Ma, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, and Jingwen Xu. Silence the judge: Reinforcement learning with self-verifier via latent geometric clustering. *ArXiv*, abs/2601.08427, 2026. URL https://api.semanticscholar.org/CorpusID:284704602.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

*Long Papers)*, pp. 1631–1662, Vienna, Austria, July 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.82. URL https://aclanthology.org/2025.acl-long.82/.

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. Process-based self-rewarding language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18097–18110, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.930. URL https://aclanthology.org/2025.findings-acl.930/.

Rosie Zhao, Anshul Shah, Xiaoyu Zhu, Xinke Deng, Zhongyu Jiang, Yang Yang, Joerg Liebelt, and Arnab Mondal. On robustness and chain-of-thought consistency of rl-finetuned vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2026.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

# A JUDGE CONTROLS

## A.1 HUMAN CALIBRATION

To validate the v1 judge, a human annotator labeled a stratified sample of 100 reasoning traces (five base models $\times$ {5 MMMU, 10 MMBench, 5 MMLU}) as multi-persona or single-voice. Sixteen traces were marked positive. The judge ranks positive traces almost perfectly (AUROC $= 0.998$, AP $= 0.992$), and $v_1$ tracks label probability well (Pearson $r = 0.92$, Brier $= 0.026$). At the conservative cut $v_1 \geq 7/9$, agreement is substantial ($\kappa = 0.68$). Every flagged trace is a true positive, although this cut recalls only $56\%$ of positives.

These measurements support alignment with the study's operational definition; they do not establish that the rubric is the only valid definition of society-like reasoning. This calibration uses the original seed-0 judge annotations; the separate analysis below tests robustness across decoding seeds.

## A.2 JUDGE SEED ROBUSTNESS

We repeat the full annotation with two additional seeds. Table 3 reports benchmark-macro means using the long-trace variants for both `Qwen3-VL-4B` models. All three runs preserve the same model ordering, with rank correlation $\rho = 1.00$ between every pair. Seed 2 scores are systematically lower, especially for Thinking and GLM, which shows that the ordering is more stable than the absolute calibration.

Table 3: Society $v_1$ across judge decoding runs and their mean $\pm$ standard deviation.

| Model | Seed 0 | Seed 1 | Seed 2 | Mean $\pm$ SD |
|---|---|---|---|---|
| Qwen3-VL-4B-Instruct | 0.137 | 0.138 | 0.128 | $0.134 \pm 0.006$ |
| Qwen3-VL-4B-Thinking | 0.313 | 0.311 | 0.273 | $0.299 \pm 0.023$ |
| Qwen3-VL-30B-A3B-Instruct | 0.046 | 0.047 | 0.038 | $0.044 \pm 0.005$ |
| Pixtral-12B-2409 | 0.106 | 0.110 | 0.084 | $0.100 \pm 0.014$ |
| GLM-4.6V-Flash | 0.258 | 0.256 | 0.222 | $0.245 \pm 0.020$ |

## A.3 GROUND TRUTH SENSITIVITY

The main analysis uses the original judge prompt, which includes the correct answer. To test whether the judge is overly influenced by that answer, we repeat the full scoring after removing it from the prompt. This comparison is a sensitivity check.

Table 4 compares the 31,595 traces for which both runs returned a valid score. The score columns are macro-averaged across the three benchmarks on this paired subset, so the ground-truth values differ from Table 1, which uses all valid traces. Removing the answer lowers the score for every model. The change is small for four models, with trace-level rank correlations between $0.67$ and $0.87$, but `Pixtral-12B` changes substantially and has a correlation of only $0.33$. Agreement at the conservative $7/9$ threshold remains above $94\%$ for every model, although Pixtral's near-zero $\kappa$ shows that this percentage is dominated by the large number of low-scoring traces.

Table 4: Sensitivity to removing ground truth from the judge prompt. Scores are benchmark-macro means on traces valid in both runs. Agreement and $\kappa$ use the $v_1 \geq 7/9$ threshold.

| Model | Paired $n$ | GT | No-GT | $\rho$ | Agree / $\kappa$ |
|---|---|---|---|---|---|
| Qwen3-VL-4B-Instruct | 6,305 | .066 | .047 | .743 | .984 / .521 |
| Qwen3-VL-4B-Thinking | 5,415 | .246 | .229 | .870 | .948 / .592 |
| Qwen3-VL-30B-A3B | 6,432 | .047 | .028 | .672 | .989 / .562 |
| Pixtral-12B | 6,755 | .106 | .019 | .328 | .973 / .049 |
| GLM-4.6V-Flash | 6,688 | .260 | .247 | .766 | .941 / .548 |

The two highest-scoring models remain `Qwen3-VL-4B-Thinking` and `GLM-4.6V-Flash`, although their order is reversed, and the rank correlation between the five model means is $0.80$. We therefore find that the main cross-model pattern is not produced only by access to the correct answer,

but the judge is not fully prompt invariant. We retain the ground truth conditioned scores because they match the manually calibrated protocol and provide context for incorrect or hallucinated reasoning.

### A.4 Unexpected Fragments

The $v_1$ judge prompt also includes *unexpected_fragments*, an axis meant to flag off-topic insertions in the trace. We report it separately as a sanity check for the dialogue-like breakdowns described by Kim et al. (2026). Such cases are rare, with the highest rate on MMMU for `Qwen3-VL-4B-Thinking`. Manual inspection found no genuine off-topic dialogues. The flagged cases were artifacts of long generation, including loops, aborted sentences, and word repetitions, as shown in Figures 6 and 7. We therefore find no evidence of off-topic internal dialogue in these models.

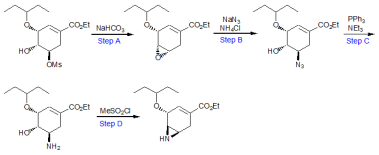



**Image input.** MMMU id = 768

**Judge evidence**

`unexpected_fragments = 3` "The trace contains **24 repeated lines** of 'Wait — perhaps the reaction is intramolecular SN2 — the OH group attacks the carbon bearing the $NH_2$? No.' which is **a clear fragmentary loop** not typical of single-writer reasoning, indicating **a malfunction or internal cognitive loop**."

**Qwen3-VL-4B-Instruct / MMMU**

To determine which reaction in the sequence is regioselective, we must analyze each step [ . . . ]
**Wait — perhaps the reaction is intramolecular SN2 — the OH group attacks the carbon bearing the $NH_2$? No.**
Wait — perhaps the reaction is intramolecular SN2 — the OH group attacks the carbon bearing the $NH_2$? No. ***(24 repetitions follow)***

Figure 6: Example of a degenerate reasoning trace exhibiting a catastrophic generation loop. The model emits the same question 24 times and fails to produce a final answer.

**Judge evidence**

`unexpected_fragments = 1` "The phrase '**salamander's salamander's salamander**' is **an odd, repeated, and clearly erroneous insertion** that breaks the logical flow and appears to be a mental stumble or accidental repetition, **not part of the final reasoning**."

**Qwen3-VL-4B-Instruct / MMLU college biology**

Let's think step by step. Polyteny is a phenomenon observed in some somatic cells (e.g., in the **salamander's salamander's salamander's salamander** — wait, no, [ . . . ] at least not the primary driver. So, the correct answer is: **(C) Chromosome replication without cell division.** Final Answer: C.

Figure 7: Example of a degenerate reasoning trace flagged by the judge. The model produces a repetition artifact mid-trace yet still recovers the correct final answer.

Table 5: Unexpected-fragment rates (%; mean ± standard deviation across judge runs). Flags are rare and concentrate in longer MMMU generations.

| Model | MMLU | MMMU | MMBench |
|---|---|---|---|
| Qwen3-VL-4B-Instruct | $1.83 \pm 0.14$ | $6.72 \pm 0.37$ | $0.47 \pm 0.05$ |
| Qwen3-VL-4B-Thinking | $0.87 \pm 0.31$ | $6.90 \pm 0.40$ | $0.34 \pm 0.17$ |
| Qwen3-VL-30B-A3B-Instruct | $0.00 \pm 0.00$ | $0.05 \pm 0.08$ | $0.01 \pm 0.01$ |
| Pixtral-12B-2409 | $0.02 \pm 0.04$ | $0.22 \pm 0.11$ | $0.00 \pm 0.00$ |
| GLM-4.6V-Flash | $1.25 \pm 0.28$ | $2.38 \pm 0.49$ | $0.19 \pm 0.06$ |

# B Additional Analysis Results

## B.1 Qualitative Examples

The main paper contrasts one high-$v_1$ and one low-$v_1$ trace. Figures 8 and 9 provide two additional examples that separate visible society structure from reconciled comparison.

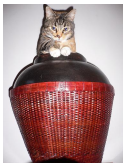


**Image input.**
MMBench id = 1749 (cat)

**v1 judge evidence: $v_1 = 0.78$ (*high*)**

`perspective_multiplicity = 2`

"The trace uses **plural self-reference** ('we must choose', 'if we must pick the least incorrect') and shifts between evaluative stances, suggesting **multiple internal voices debating the options** rather than a single coherent narrator."

`viewpoint_conflict = 3`

"The trace **explicitly rejects option C as incorrect** ('the cat is clearly on top'), then **contradicts itself** by stating 'C is the most plausible if we misinterpret the image, but it's still wrong' — showing internal disagreement about the validity of the chosen answer."

**v2 judge evidence: $v_2 = 0.46$ (*mid*)**

`reconciliation = 2`

"The trace attempts to reconcile by concluding that none are correct, but if forced to choose, it selects C as the 'most plausible' despite knowing it's wrong — a **partial integration acknowledging the error but making a pragmatic choice**."

`unresolved_drift = 1`

"The **final answer (C) contradicts the trace's own conclusion** that the cat is on top, not inside, indicating that the conflict is not fully resolved."

**Qwen3-VL-30B-A3B-Instruct / MMBench / society $v_1 = 0.78, v_2 = 0.46$ / correct = 1**

<analysis> The image shows a cat sitting on top of a large, dark-coloured vase with a woven red base. The cat is positioned *above* the vase, not inside it, and is not to the left or in front of it. The vase is not facing away from the cat. **Therefore, none of the options perfectly describe the relationship**, but the closest correct one is that the cat is on top, which is not listed. **However, if we must choose from the given options:** *(A)* is incorrect (cat is on top, not in front); *(B)* is incorrect (cat is centred on top, not left); *(C) "the cat is inside the vase"* **is incorrect as the cat is clearly on top**; *(D)* is irrelevant (there is no car in the image). None of the options are correct, but if we must pick **the least incorrect**, *C* is the most plausible *if we misinterpret the image*, **but it's still wrong**. </analysis><answer>**C. The cat is inside the vase.** </answer>

Figure 8: Trace with high $v_1$ score but moderate $v_2$ (also present *role_separation*=2 for $v_1$). The trace rejects option C, then selects it under a "least incorrect" rule. The $v_1$ score rewards the visible conflict, while the $v_2$ communication score gives only partial credit because the conflict is not fully reconciled.

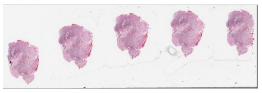


**Image input.**
MMMU id = 339

**Judge evidence**

`perspective_conflict = 3`

"The trace **explicitly contrasts multiple diseases** (Trypanosomiasis, Amebic encephalitis, Cerebral malaria, Aspergillosis) with Tuberculous meningitis, highlighting how each differs in pathology and clinical course, **creating a tension between competing diagnoses**."

`reconciliation = 3`

"The trace concludes by **integrating the clinical, histological, and epidemiological evidence**."

**Qwen3-VL-30B-A3B-Instruct / MMMU / $v_1 = 0, v_2 = 1$ / correct = 0**

**A. Trypanosomiasis** — sleeping sickness, with psychiatric symptoms and ataxia, **not granulomatous**.
**B. Amebic encephalitis** — rapidly progressive necrotizing inflammation, **not granulomatous**.
**C. Cerebral malaria** — sequestration in cerebral capillaries, microhemorrhages, **not granulomas**.
**D. Tuberculous meningitis** — classic chronic meningitis in immigrants; **hallmark is granulomatous inflammation with caseous necrosis, matching the image**.
**E. Aspergillosis** — necrotizing vasculitis, **not granulomatous with caseation**.
<answer>**D. Tuberculous meningitis**</answer> (GT = A; model is wrong)

Figure 9: A trace from a single narrator with high $v_2$ communication score despite $v_1 = 0$. The trace compares answer options and chooses a final choice, so $v_2$ credits reasoning across perspectives rather than persona behavior.

### B.2 Quantitative Results

The following tables provide the full breakdowns behind the aggregate results in the main paper. Table 6 reports empirical item difficulty by model and $v_1$ bucket; Table 7 stratifies the accuracy–$v_1$ relationship by MMMU difficulty; Table 8 reports the frequency of high-$v_1$ traces; and Table 9 reports accuracy and trace share across the communication-score buckets.

Table 6: Mean empirical item difficulty $1 - \bar{a}$ within each $v_1$ bucket for each model, reported as mean $\pm$ standard deviation across three judge runs. Difficulty rises with $v_1$ for every model.

| Model | $v_1 = 0$ | $(0, .33]$ | $(.33, .67]$ | $> .67$ |
|---|---|---|---|---|
| Qwen3-VL-4B-Instruct | $.141 \pm .006$ | $.478 \pm .013$ | $.523 \pm .008$ | $.605 \pm .002$ |
| Qwen3-VL-4B-Thinking | $.100 \pm .009$ | $.135 \pm .007$ | $.270 \pm .005$ | $.504 \pm .013$ |
| Qwen3-VL-30B-A3B-Instruct | $.153 \pm .005$ | $.493 \pm .017$ | $.534 \pm .026$ | $.600 \pm .022$ |
| Pixtral-12B-2409 | $.126 \pm .014$ | $.471 \pm .013$ | $.499 \pm .012$ | $.524 \pm .018$ |
| GLM-4.6V-Flash | $.081 \pm .008$ | $.158 \pm .014$ | $.435 \pm .036$ | $.582 \pm .004$ |

Table 7: MMMU accuracy per $v_1$ bucket, averaged across the five base models and stratified by MMMU's native *topic_difficulty* label (Easy: 295 items; Medium: 424; Hard: 181). Cells show mean $\pm$ standard deviation across three judge runs. Accuracy declines with $v_1$ within every difficulty level.

| Difficulty | $v_1 = 0$ | $(0, .33]$ | $(.33, .67]$ | $> .67$ |
|---|---|---|---|---|
| Easy | $.870 \pm .021$ | $.564 \pm .022$ | $.376 \pm .013$ | $.278 \pm .017$ |
| Medium | $.781 \pm .025$ | $.453 \pm .018$ | $.421 \pm .029$ | $.219 \pm .023$ |
| Hard | $.605 \pm .038$ | $.336 \pm .021$ | $.308 \pm .006$ | $.255 \pm .037$ |

Table 8: Fraction of reasoning traces (%) classified as strongly multi-persona (society $v_1 > 0.67$), reported as mean $\pm$ standard deviation across three judge runs. High-scoring traces appear in every model and remain most frequent for `Qwen3-VL-4B-Thinking`.

| Model | MMMU | MMBench | MMLU | Avg |
|---|---|---|---|---|
| Qwen3-VL-4B-Instruct | $20.5 \pm 1.2$ | $3.1 \pm 0.1$ | $7.7 \pm 0.1$ | $10.4 \pm 0.4$ |
| Qwen3-VL-4B-Thinking | $26.2 \pm 3.8$ | $8.9 \pm 0.8$ | $5.9 \pm 1.8$ | $13.7 \pm 2.1$ |
| Qwen3-VL-30B-A3B-Instruct | $5.3 \pm 1.0$ | $0.9 \pm 0.2$ | $1.5 \pm 0.1$ | $2.6 \pm 0.4$ |
| Pixtral-12B-2409 | $4.1 \pm 1.7$ | $1.5 \pm 0.5$ | $2.7 \pm 0.0$ | $2.8 \pm 0.7$ |
| GLM-4.6V-Flash | $16.1 \pm 2.2$ | $5.8 \pm 0.9$ | $5.8 \pm 1.8$ | $9.2 \pm 1.5$ |

Table 9: Accuracy and trace share within each $v_2$ communication bucket across MMMU, MMBench, and MMLU. The original $v_2$ annotation run is used. Procedural traces at zero and clearly reconciled positive traces are more accurate than negative or weakly positive traces.

| Model | $< 0$ flagged | $= 0$ no reconciliation | $(0, .5]$ weak | $(.5, .8]$ moderate | $> .8$ reconciled |
|---|---|---|---|---|---|
| Qwen3-VL-4B-Instruct | .124 (1.8%) | .767 (2.5%) | .434 (1.4%) | .542 (15.8%) | .653 (78.6%) |
| Qwen3-VL-4B-Thinking | .042 (0.7%) | .867 (0.5%) | .200 (0.9%) | .889 (3.6%) | .817 (94.3%) |
| Qwen3-VL-30B-A3B-Instruct | .333 (0.1%) | .899 (5.4%) | .857 (0.9%) | .922 (20.0%) | .886 (73.6%) |
| Pixtral-12B-2409 | .419 (5.1%) | .807 (13.2%) | .503 (5.6%) | .745 (36.9%) | .712 (39.3%) |
| GLM-4.6V-Flash | .189 (1.1%) | .969 (6.7%) | .371 (1.7%) | .879 (13.8%) | .779 (76.7%) |

# C Additional Intervention Results

## C.1 Intervention details

Table 10 reports the complete accuracy and $v_1$ results for every policy model, training dataset, and main reward condition.

Table 10: Evaluation-set accuracy / common-v1 score for the main intervention conditions. Random reward is a control for RL updates without task or judge information. The accuracy reward has direct access to ground truth, whereas $v_1^-$ uses only the reasoning trace.

| | Geometry3k | | ChartQA | | GSM8K | |
|---|---|---|---|---|---|---|
| Reward | Instruct | Thinking | Instruct | Thinking | Instruct | Thinking |
| Base *(no training)* | .564/.135 | .561/.064 | .818/.036 | .796/.028 | .942/.024 | .917/.021 |
| Random *(control)* | .554/.116 | .591/.054 | .819/.033 | .769/.017 | .948/.025 | .935/.018 |
| Accuracy | **.604**/.077 | **.636**/.042 | **.940**/.038 | **.938**/.006 | .940/.020 | .938/.015 |
| $v_1^-$ society | .587/.034 | .604/.038 | .836/.008 | .825/.003 | .939/.017 | **.942**/.013 |

## C.2 Training Reward Trajectory Reconstruction

Figure 10 contains the training trajectory. Mean $v_1$ falls from the first ten to the last ten steps under $v_1^-$ for both policy models. The $v_2$ curves rise overall under $v_2^+$, but the step-level movement is noisier, so they support only the direction of steering rather than a smooth training trend.

The reconciliation gate makes the $v_2$ reward less smooth than $v_1$ because productive credit changes with both the judged structure and whether the trace reconciles it. A smoother trajectory may therefore require more rollout data or a longer training horizon than used here.

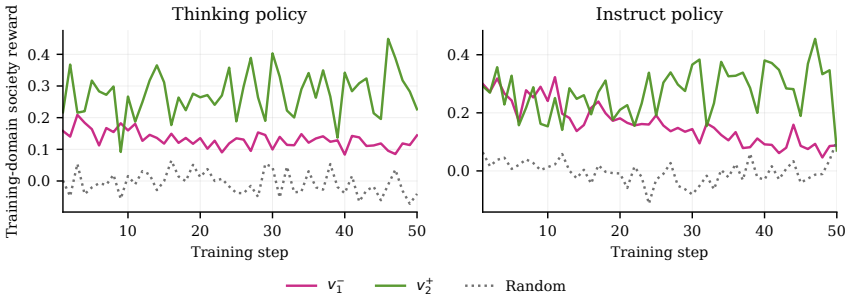



Figure 10: Training-domain society reward trajectories over all 50 steps. Mean society reward decreases under $v_1^-$ and increases under $v_2^+$ for both policy families, while the random control changes only slightly. The opposing directions show that training responds to the definition and sign of the judge reward. The $v_1$ and $v_2$ curves use different rubrics, so their absolute magnitudes should not be compared. These are rollout metrics rather than downstream checkpoint evaluations.

## C.3 Communication reward outcome

Raw $v_2^+$ frequently increases response length and can reward traces that name a candidate answer but continue deliberating without a stable commitment. Accuracy gating repairs part of this behavior on Geometry3k, reinforcing that a communication-style score needs an external correctness or verification constraint.

All variants keep the $v_2$ judge prompt fixed and change only the reward transformation:

- **Anti mixed** negates scores in the weak middle region $0 \leq v_2 \leq 0.5$ and shifts scores above $0.5$ downward.
- **Trap filtered** maps $0 < v_2 \leq 0.5$ to a near-neutral constant under group normalization, while leaving other scores unchanged.

- **Length penalty** subtracts a capped penalty that grows between 512 and 3072 response tokens.
- **Combined with accuracy** adds the standard correctness reward to either $v_1^-$ or $v_2^+$; **accuracy gated** instead gives $v_2$ credit only when the extracted final answer is correct.

Table 11: Geometry3k accuracy for communication reward ablations. Pure $v_2$ reshaping does not recover the Instruct loss, while accuracy gating gives the strongest result for both policy models.

| Reward | Instruct | Thinking |
|---|---|---|
| Base *(no training)* | 0.564 | 0.561 |
| Accuracy reward | 0.604 | 0.636 |
| $v_1^-$ | 0.587 | 0.604 |
| $v_2^+$ | 0.529 | 0.621 |
| $v_2^+$ anti mixed | 0.521 | 0.611 |
| $v_2^+$ trap filtered | 0.522 | 0.602 |
| $v_2^+$ length penalty | 0.569 | 0.592 |
| $v_2^+$ + accuracy | 0.584 | 0.614 |
| $v_2^+$ accuracy gated | **0.616** | **0.647** |

### C.4 Hallucination Scope

Table 12 evaluates checkpoints on POPE, AMBER, and CHAIR, which were not used for training. The $v_1^-$ checkpoints improve several hallucination measures, particularly for Instruct, but the effect is not consistent across model families or stronger than the accuracy control on every metric. We therefore do not claim that SoT introduces hallucination. The more limited interpretation is that long or drifting reasoning creates additional opportunities to lose grounding, and suppressing a correlated trace style can sometimes reduce those failures.

Table 12: Hallucination evaluation for Base, a Geometry3k accuracy checkpoint, and $v_1^-$ checkpoints trained on each task. POPE reports accuracy (↑); AMBER and CHAIR report hallucination or coverage measures with the preferred direction shown in the header. Results are mixed across models and controls.

| | | POPE accuracy | | | AMBER-G | | | | CHAIR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Reward | adv↑ | pop↑ | rand↑ | chair↓ | cover↑ | hal↓ | cog↓ | $i\downarrow$ | $s\downarrow$ | recall↑ |
| Instruct | Base | .879 | .891 | .911 | 6.63 | 75.09 | 55.92 | 5.73 | 12.10 | 57.20 | 75.05 |
| | Accuracy | .886 | .896 | .915 | 6.48 | 73.93 | 52.94 | 5.22 | 11.55 | 55.00 | 75.11 |
| | $v_1^-$ (Geo3k) | .885 | .896 | .914 | 6.08 | 73.86 | 53.83 | 5.06 | 10.38 | 52.80 | 75.74 |
| | $v_1^-$ (ChartQA) | .882 | .894 | .915 | 6.12 | 74.50 | 53.04 | 5.24 | 11.94 | 54.00 | 74.86 |
| | $v_1^-$ (GSM8K) | .885 | .892 | .912 | 6.69 | 74.09 | 55.52 | 6.23 | 11.90 | 58.80 | 74.73 |
| Thinking | Base | .868 | .879 | .896 | 7.02 | 72.55 | 49.95 | 5.04 | 11.42 | 58.00 | 76.43 |
| | Accuracy | .868 | .884 | .898 | 6.21 | 72.29 | 44.88 | 4.22 | 11.96 | 54.40 | 75.43 |
| | $v_1^-$ (Geo3k) | .878 | .881 | .897 | 6.57 | 72.68 | 47.46 | 5.69 | 11.32 | 58.40 | 75.87 |
| | $v_1^-$ (ChartQA) | .872 | .879 | .903 | 6.41 | 72.36 | 44.88 | 4.65 | 11.55 | 56.40 | 75.80 |
| | $v_1^-$ (GSM8K) | .877 | .885 | .902 | 6.42 | 72.27 | 46.37 | 4.89 | 11.22 | 55.80 | 75.55 |

# D REPRODUCIBILITY

## D.1 BASE MODEL TRACE GENERATION

All models are evaluated through lmms-eval using the vLLM backend. The output budget is `max_new_tokens=1536` and is increased to 6144 for the `Qwen3-VL-4B` variants because the shorter cap truncates approximately 30–40% of their traces. Decoding follows each model card: `Qwen3-VL-4B-Instruct` and `Qwen3-VL-30B-A3B-Instruct` use temperature 0.7, top-$p$ 0.8, and top-$k$ 20; `Qwen3-VL-4B-Thinking` uses temperature 1.0, top-$p$ 0.95, and top-$k$ 20; Pixtral uses temperature 0.7 and top-$p$ 1.0; GLM uses temperature 0.8, top-$p$ 0.6, top-$k$ 2, and repetition penalty 1.1.

## D.2 OFFLINE JUDGE CALLS

The primary annotations use `Qwen3-30B-A3B-Instruct-2507-FP8`, served locally with vLLM. Judge decoding uses temperature 0.7, top-$p$ 0.8, top-$k$ 20, min-$p$ 0, and a 512-token output budget for $v_1$ (1024 for $v_2$). The analysis reports the mean and standard deviation across 3 seeds.

## D.3 INTERVENTION RUNS

All runs use verl (Sheng et al., 2025) with GDPO (Liu et al., 2026), rollouts set to 8, and sampling temperature 1.0 during training. PPO uses AdamW, a small KL-to-reference loss against the frozen base policy, and no critic. Each job uses one node with four A100 GPUs: three for training and rollout and one for the local online judge. Prompts are capped at 1024 tokens and rollouts at 3072, giving a 4096 token training context.

Each intervention condition uses one initialization seed (42). Geometry3k runs for 50 outer steps and takes roughly 4–6 hours; ChartQA and GSM8K use 25 steps and take approximately 2 and 1.5 hours, respectively. Because each condition requires three A100 GPUs for training and rollout plus one for the online judge, this computational cost limited the intervention runs to one initialization seed.