LOOKING BUT NOT UPDATING: VISUAL BELIEF INERTIA IN EMBODIED MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

An embodied assistant must judge a scene when its action history describes an outcome that the image contradicts. We study visual belief inertia: state answers that follow an incorrect action-conditioned history in independent single-image queries. VBI-Bench comprises 2,285 robot-state questions and two complementary controls: vary history while fixing the image, or vary the image while fixing all text. These controls reveal that strong performance under both correct and incorrect histories can coexist with poor discrimination between images. On Qwen3-VL-8B, revision-invariant fine-tuning (RI-FT) raises full-conflict accuracy from 14.4% to 98.3% on a 713-item history grid, yet answers both members of only 6.5% of 552 fixed-text image pairs correctly. RI-FT-CF adds image pairs with shared history, question, and options but different targets to the same answer-supervised objective. Across three seeds, both-correct pair accuracy reaches 77.5%, alongside 94.6% full-conflict and 98.3% correct-history accuracy. Component comparisons link the paired gain to the added examples and show how history variation changes the joint performance across conditions. The resulting diagnosis and training recipe make image changes an explicit part of evaluating history-resistant state judgments.

1 INTRODUCTION

A robot's history says that a manipulation failed, but its current image shows the intended outcome. Reliable state judgments connect visual question answering to state estimation (Kaelbling et al., 1998). Vision-language pretraining and visual instruction tuning enable image-conditioned responses (Li et al., 2023a; Liu et al., 2023b; 2024a); we examine their reliability when history and observation disagree.

We call answers that retain an incorrect action-conditioned account *visual belief inertia* (VBI). We study controlled textual histories and single-image questions, without assuming a persistent internal belief state. In Figure 1, the base model repeats that a goblet is empty after ice has entered it. Correcting such errors creates a measurement challenge: training can replace one answer preference with another, raising conflict accuracy without distinguishing different images.

Language-conditioned policies connect visual observations to robot actions (Brohan et al., 2023; Zitkovich et al., 2023; Kim et al., 2025), while robot benchmarks assess execution and failure understanding (Liu et al., 2023a; Luo et al., 2026). We isolate the underlying question: *what supports a state answer when history and vision disagree?*

Prior work shows that reducing visual sycophancy can impair receptiveness to correct user feedback (Li et al., 2025). We ask whether retaining correct-history performance also ensures image dependence. VBI-Bench varies history under fixed images and images under fixed text. Answering both differently labeled images correctly rules out a constant text-only answer.

Conflicting histories reduce accuracy across nine open-weight models. On Qwen3-VL-8B, revision-invariant fine-tuning (RI-FT) supervises correct answers under varied histories and reaches 98.3% full-conflict and 98.7% correct-history accuracy, yet both-correct pair accuracy remains 6.5%. RI-

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: **History interference and the image-dependence gap.** A RoboBench ice-pouring pair shares history, question, and options. Qwen3-VL-8B-Instruct (Base) answers "empty" for both frames; RI-FT answers "contains ice" for both; RI-FT-CF answers both correctly. Archived predictions use independent single-image queries (RI-FT-CF seed 42). Displayed answers expand option letters; belief text is abbreviated.

FT-CF adds fixed-text image pairs with different targets to the same objective. Component comparisons separate the paired gain associated with these examples from the effects of history variation.

We contribute complementary controls for robot-state judgments, an empirical diagnosis of the image-dependence gap after history-resistant training, and a paired recipe tested across conditions, seeds, and groups unseen by the training union. High scores under changing histories leave correct responses to changing images unresolved.

2 VBI-Bench: Two Controls for State Judgments

2.1 Task and complementary tests

Each item contains an image I, a current-state question q, four options $\mathcal{O}$, and history h; the model returns an option letter $f_\theta(I, h, q, \mathcal{O})$. History comprises an action description (Action), an outcome report (Report), and a stated belief. The report's original header is "Expected outcome," but its templates often assert completed success or failure; Section 4 examines sensitivity to its wording.

The *history control* fixes the image, question, and options. NH (No History) omits context; CH (Correct History) supplies a correct belief; WB (Wrong Belief) supplies an incorrect belief (Belief); FC (Full Conflict) combines Action, Report, and Belief (Full history). We measure accuracy and the rate of selecting the false-belief option. CH is a correct-belief condition that restates the target, whereas FC supplies all three history blocks. Their scores measure performance under these specified formats.

The *image control*, visual counterfactual evaluation (VCF), fixes all text across earlier and later frames I^-, I^+ from one episode with different model-verified targets. We query each image independently and score

$$\text{PairAcc} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\left[f_\theta(I_j^-, h_j, q_j, \mathcal{O}_j) = y_j^- \ \wedge \ f_\theta(I_j^+, h_j, q_j, \mathcal{O}_j) = y_j^+\right]. \quad (1)$$

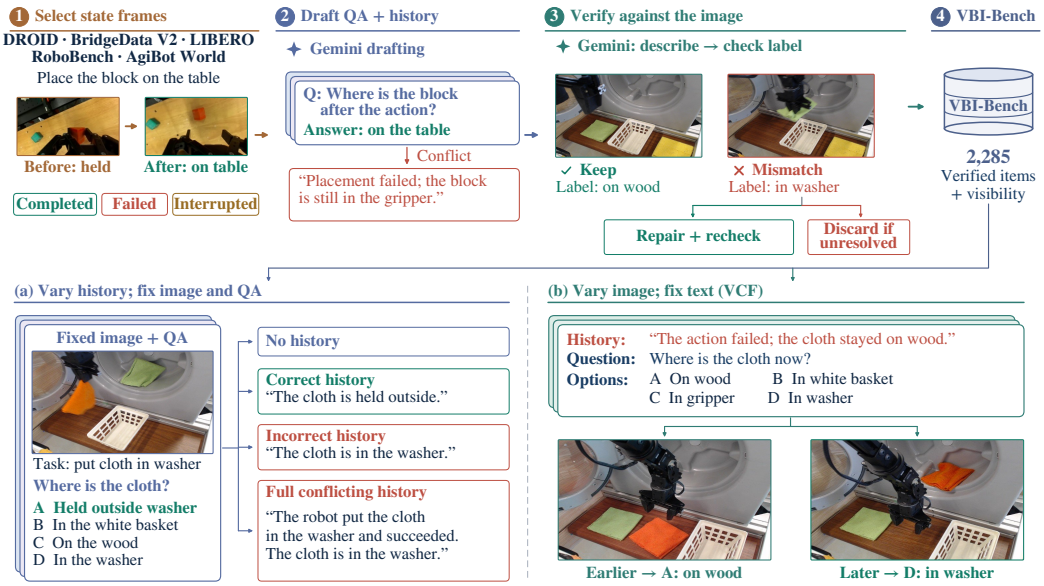

Figure 2: **Construction and complementary controls.** Top: trajectory frames undergo Gemini drafting and image-based verification. Bottom: (a) histories vary with image, question, and options fixed; (b) images vary with all text fixed and are queried independently. Block and cloth examples illustrate construction; verification is model-assisted.

A deterministic image-independent rule cannot satisfy both targets; changing the answer alone is insufficient. Fixed history may agree with one frame and conflict with the other. Labels describe each frame's visible state, but the shared question can retain "after the action" or "final state" wording. The verifier explicitly interprets these as current-state questions; the evaluated models receive a general look-at-image instruction. PairAcc therefore measures agreement with the frame-specific annotations under the stored prompts; earlier-frame errors can also reflect temporal interpretation (Appendix A).

2.2 Construction and Data Examples

VBI-Bench contains 2,285 questions across 12 source identifiers. Sources include DROID, BridgeData V2, LIBERO, RoboBench, and AgiBot World (Khazatsky et al., 2024; Walke et al., 2023; Liu et al., 2023a; Luo et al., 2026; Team AgiBot-World, 2025). There are 1,030 successful and 1,255 failed outcomes; 1,163 items are from LIBERO simulation (Figure 3). Failure records include premature stops and dynamic execution failures, so the mixture is not a prevalence estimate of deployed robot failures.

Source-specific signals localize frames. Gemini drafts questions, labels, histories, and distractors; a separate describe-then-adjudicate pass repairs or rejects inconsistent records (Figure 2). Automated checks validate fields, images, option roles, and outcome polarity. Histories include false success and false failure. VCF selects model-verified state pairs. Appendix A details construction and question forms.

2.3 Splits, Quality, and Scoring

The history grid contains 1,572 training and 713 test items, with the full 31-item LIBERO-10 source held out. This is principally a sample-level split. VCF has 564 training and 552 evaluation pairs, disjoint by recorded scene keys within that component. Auditing both training components finds 359 evaluation pairs whose keys occur in ordinary RI-FT training; the remaining 193 pairs provide the training-union-unseen analysis. The full 552-pair set is the original diagnostic test, rather than a wholly unseen-scene estimate. The paired training sample IDs also exclude the 713 grid-test IDs. Keys denote episodes or task demonstrations, not unique physical environments (Appendix B).

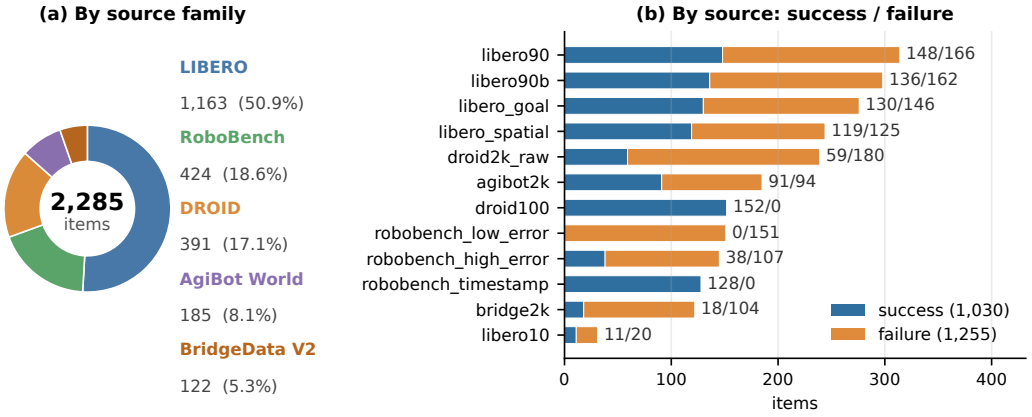

Figure 3: **Source and outcome composition of the 2,285-item benchmark.** (a) Source-family counts and proportions: 50.9% LIBERO simulation, with the rest from real-robot recordings. (b) Success/failure counts by source identifier. Counts describe benchmark items, not paired training or test examples.

Annotations and verification are model-assisted, without independent human annotation agreement. Recorded visibility labels are 2,082 clear, 149 weak, and 54 invisible; Appendix A.3 gives test-set counts and stratified results. No-image NH accuracy is 45.6%, above four-option chance (25%), making image-controlled evaluation essential. Options use a deterministic item-seeded shuffle shared across conditions. Decoding is greedy; invalid outputs count as errors. Unless specified otherwise, grid/pair denominators are 713/552 (Appendix B).

3 Behavioral Diagnosis: What Supports the Answer?

3.1 Incorrect histories systematically redirect answers

Qwen3-VL-8B scores 71.0% under NH and 89.5% under CH. Incorrect belief reduces accuracy to 56.1%; full conflict reduces it to 14.4%, with 82.9% choosing the belief option. All nine general-purpose and robot-oriented models lose accuracy under FC (Table 1). Drops are 19.9–59.9 points from unrounded scores; the table subtracts displayed values.

Table 1: History interference across nine open-weight models (%, 713 items). Drop is NH minus FC; larger values indicate greater interference. Invalid NH/FC outputs count as errors: 16/25 for GLM, zero for the others.

Model	NH ↑	CH ↑	WB ↑	FC ↑	Drop
Qwen3-VL-8B (Bai et al., 2025)	71.0	89.5	56.1	14.4	+56.6
GLM-4.1V-9B-Thinking (GLM-V Team, 2025)	69.3	77.7	56.5	33.5	+35.8
Phi-4-MM (Abouelenin et al., 2025)	55.3	67.5	52.2	35.3	+20.0
Magma-8B (Yang et al., 2025)	51.9	57.8	45.7	27.1	+24.8
Molmo-7B (Deitke et al., 2025)	65.4	56.2	60.2	24.1	+41.3
InternVL3-8B (Zhu et al., 2025)	63.1	82.7	35.3	21.6	+41.5
Phi-3.5-Vision (Abdin et al., 2024)	51.9	80.5	32.1	16.4	+35.5
LLaVA-Video-7B (Zhang et al., 2024)	62.1	78.3	48.4	15.8	+46.3
Cosmos-Reason1-7B (Azzolini et al., 2025)	69.7	95.7	27.1	9.8	+59.9

A 20-participant questionnaire yields 87.9% no-history and 77.3% conflict accuracy. Between-participant assignment makes this an answerability reference rather than a matched human–model causal comparison (Appendix C).

3.2 HIGH CONFLICT ACCURACY CAN CONCEAL WEAK IMAGE DEPENDENCE

Plain SFT repeats no-history QA to match RI-FT's 11,004-row budget; RI-FT uses seven history views per item (Section 5). Table 2 shows the mismatch: RI-FT reaches 98.3% FC accuracy but only 6.5% PairAcc, below plain SFT's 22.5%. FC scores individual items; PairAcc scores two-image trials.

Table 2: History resistance and image dependence (%, Qwen3-VL-8B). FC uses 713 items; other columns use 552 pairs with conflict history. Earlier/later score each image; no-image later matches an image-free answer to the later target.

Method	Grid FC	PairAcc	Earlier	Later	No-image later
Base	14.4	5.4	73.0	8.3	2.0
Plain SFT	62.6	22.5	62.1	52.7	9.6
RI-FT	**98.3**	**6.5**	**7.2**	**99.1**	**94.6**

For a uniformly sampled evaluation pair, let C^-, C^+ denote correct earlier/later answers and $A^\pm = \Pr(C^\pm)$. Then

$$\text{PairAcc} = \Pr(C^- \cap C^+) \leq \min\{A^-, A^+\}. \tag{2}$$

Thus, RI-FT's 7.2% earlier-frame accuracy bounds PairAcc despite 99.1% later-frame accuracy. Without images, it selects the later target on 94.6% of pairs. This identifies an outcome preference available from text alone; it does not separate learned answer priors from interpretation of outcome-oriented questions. Together, the paired marginals and no-image control show why the high FC score cannot establish reliable image-dependent answers (Figure 1).

4 HISTORY COMPONENTS AND ANSWER-SELECTION INTERVENTIONS

4.1 OUTCOME REPORTS AND TEMPLATE SENSITIVITY

Factorial interventions fix images and options while varying Action, Report, and Belief (Figure 4a). Report alone yields 20.2% accuracy, versus 56.5% for Action and 56.1% for Belief, making it the strongest individual interference source among these components.

Report rewrites change the response (Figure 4b). Single-block instruction, goal, prediction, and possibility rewrites reach 57.4–62.6%, versus 51.1% for a success-only report. This report differs in header and content from the original Report-only condition. Within full history, an unverified-hypothesis rewrite reaches 44.7%. These comparisons show sensitivity to factual framing jointly with wording and outcome polarity; Appendix C specifies the prompts.

4.2 ANSWER-SELECTION INTERVENTIONS

On 60 items correct under NH but belief-following under FC, we exchange attention-module outputs $a^h_{\ell,t}$ between history conditions h at layer ℓ and the last prompt token t (Figure 5):

$$\underbrace{a^{\text{NH}}_{\ell,t} \leftarrow a^{\text{FC}}_{\ell,t}}_{\text{corruption}}, \qquad \underbrace{a^{\text{FC}}_{\ell,t} \leftarrow a^{\text{NH}}_{\ell,t}}_{\text{rescue}}. \tag{3}$$

Layer 24 redirects 44/60 NH answers to the belief and rescues 38/60 FC answers. On 60 separately selected non-belief errors paired with correct paraphrases, corruption/rescue reach 48.3/73.3%. Training only layer 24 (93.4% FC) or all layers except it (98.3%) improves FC accuracy. These controls locate an influential, broadly used answer-selection computation.

State probes perform similarly with original, absent, and swapped images, leaving intact perception unestablished (Appendix D). These interventions characterize base-model answer selection. The training design instead follows the behavioral gap: supervise stable targets across history changes and different targets across fixed-text image changes.

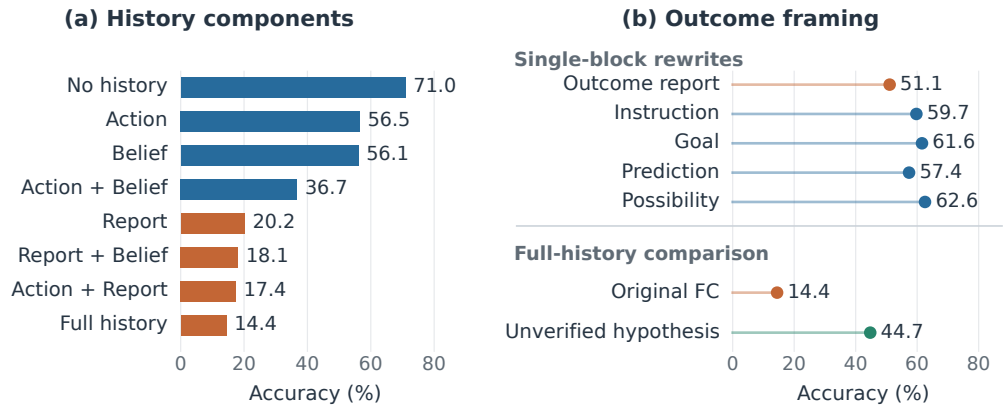

Figure 4: **History components and outcome framing.** Qwen3-VL-8B-Instruct, 713 items. (a) Action, Report, Belief, and combinations; orange bars include Report. (b) Single-block rewrites concern successful completion. The separate unverified-hypothesis rewrite preserves Action and Belief.

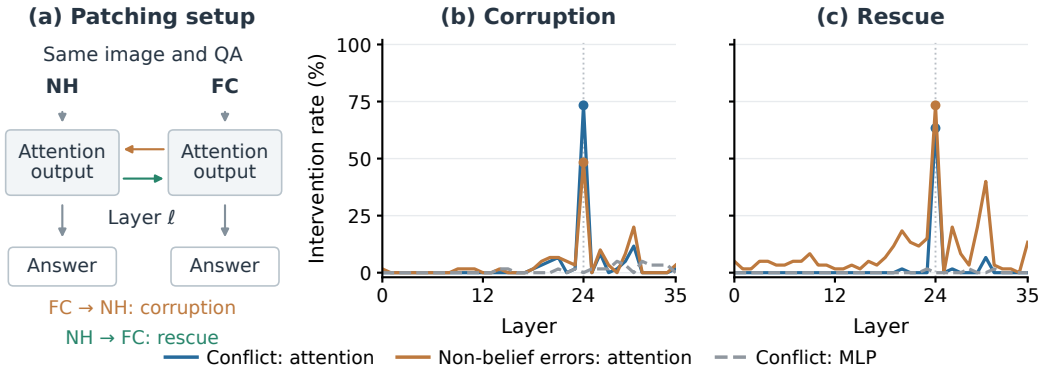

Figure 5: **Activation replacement across layers.** Qwen3-VL-8B-Instruct. (a) Swap attention outputs at the last prompt token. (b) Corruption matches the donor's wrong answer; (c) rescue matches the target. Each group has 60 separately selected items; MLP uses the conflict set. Zero-indexed layer 24 (dotted) was selected on this scan.

5 Counterfactual-Pair Training for State Judgments

RI-FT-CF combines history variation and image variation in one supervised mixture (Figure 6). Its history component, RI-FT, gives each of 1,572 items one NH view, one CH view, three WB phrasings, and two FC phrasings: 11,004 rows. NH anchors ordinary state QA, CH covers correct-belief inputs, and conflict views supervise answers against misleading accounts.

The image component adds 564 verified pairs, or 1,128 rows. Within each pair, history, question, and option ordering are identical, while the image and correct answer differ. The total is 12,132 rows. History augmentation varies text while holding the target fixed; these rows additionally expose different targets under identical text. A constant text-conditioned answer cannot satisfy both pair labels.

Let $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$ concatenate the history and pair-derived rows, retaining repetitions, with $x_i = (I_i, h_i, q_i, \mathcal{O}_i)$ and $N = 12{,}132$. Training minimizes answer-token cross-entropy,

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(y_i \mid I_i, h_i, q_i, \mathcal{O}_i). \tag{4}$$

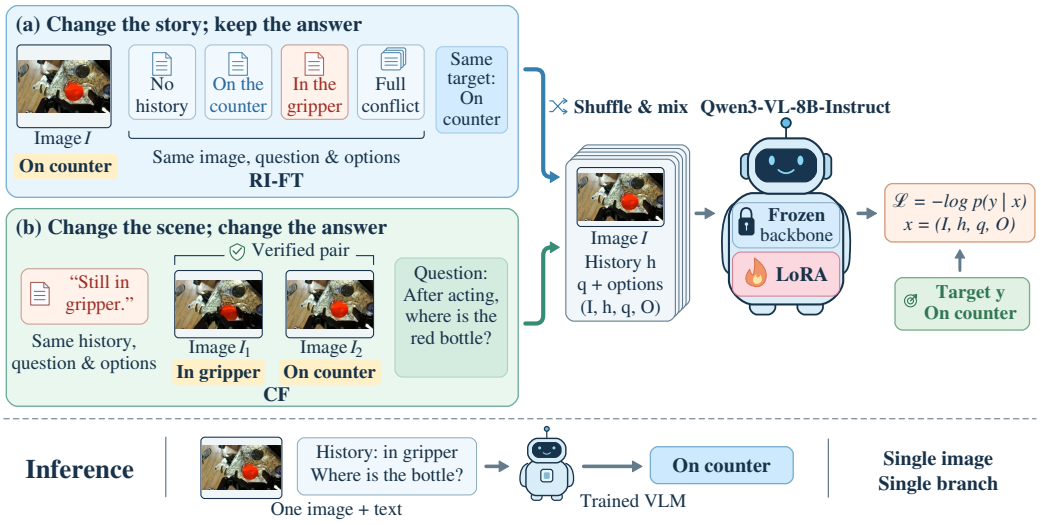

Figure 6: **RI-FT-CF training and inference.** (a) Vary history while preserving the image and target. (b) Vary the image and target while preserving all text. The rows are shuffled into answer-supervised LoRA training with a frozen backbone. Pairing resides in the data; inference uses one image and a single decoding branch. Bottle images illustrate the supervision structure.

Cross-entropy penalizes image-independent answers to distinct pair targets. Rows are shuffled and trained independently; pair identifiers are excluded from the loss.

We initialize Qwen3-VL-8B-Instruct (Bai et al., 2025) and train LoRA (Hu et al., 2022) on language-model attention and MLP projections: rank 16, scaling 32, dropout 0.05, learning rate 10^{-4}, bf16, effective batch size 16, and one epoch (759 updates). The full recipe adds data and steps relative to RI-FT; Appendix E specifies the component budgets. Deployment remains single-image option prediction.

6 Evaluation of Counterfactual-Pair Training

6.1 Joint performance across history and image controls

Table 3 evaluates both axes. Across seeds 42–44, RI-FT-CF obtains mean PairAcc 77.5% (76.6–78.6), FC 94.6% (94.1–95.7), and CH 98.3%. Relative to RI-FT, seed 42 gains 70.8 points in PairAcc and loses 4.2 in FC; paired bootstrap 95% intervals are $[67.0, 74.6]$ and $[-5.9, -2.5]$ points (Appendix B). Earlier/later accuracies change from 7.2/99.1% to 80.6/96.2%, locating the paired gain on earlier frames.

Table 3: Joint evaluation (%). Grid columns use 713 items; CH supplies a correct belief. PairAcc requires both answers correct: Conflict and Clean use the same 552 pairs with and without history; Unseen uses the 193 training-union-unseen pairs with history. Parentheses identify training seeds.

	History grid			PairAcc		
Method	NH	CH	FC	Conflict	Clean	Unseen
Base	71.0	89.5	14.4	5.4	33.7	6.7
Plain SFT	93.3	92.4	62.6	22.5	40.0	18.1
RI-FT	93.1	98.7	98.3	6.5	35.7	5.2
VCF-only	82.9	84.7	86.8	76.1	69.0	82.9
Plain SFT + VCF	93.5	90.7	89.1	77.5	44.7	80.8
RI-FT-CF (42)	92.6	98.0	94.1	77.4	65.2	82.9
RI-FT-CF (43)	93.4	97.9	95.7	76.6	59.8	82.9
RI-FT-CF (44)	93.7	98.9	94.1	78.6	61.6	83.9

6.2 What the two data components contribute

VCF-only and plain SFT plus VCF reach 76.1% and 77.5% conflict PairAcc, recovering most of the improvement over RI-FT. The latter matches the full 12,132-row, one-epoch budget; VCF-only uses 1,128 rows for two epochs. These comparisons associate the paired gain with the added pair-derived examples, which expand state and label coverage as well as providing fixed-text image contrasts. They do not isolate those factors from equally diverse unpaired additions.

The full mixture changes the joint performance across conditions. Compared with budget-matched plain SFT plus VCF, seed 42 gains 7.3 points in CH, 5.0 in FC, and 20.5 in Clean PairAcc, with nearly identical Conflict PairAcc (77.4% versus 77.5%). Relative to VCF-only, its CH/FC scores are 98.0/94.1% versus 84.7/86.8%, but its Clean PairAcc is lower (65.2% versus 69.0%). Across full-mixture seeds, mean Clean PairAcc is 62.2%, compared with 35.7% for RI-FT. The mixture trades off performance across conditions. Component recipes have one checkpoint each, so small differences are descriptive.

Ordinary RI-FT without CH retains 98.3% FC but drops from 98.7% to 26.2% CH. This motivates correct-history coverage; the ablation concerns RI-FT, not RI-FT-CF (Appendix E).

6.3 Unseen groups and domain differences

On the 193 pairs with recorded groups unseen by both training components, RI-FT-CF reaches 82.9–83.9%, versus 5.2% for RI-FT. The subset contains 131 simulated and 62 real-robot pairs, so its higher score does not isolate group novelty. On the 512 pairs marked clear in the existing annotations, the corresponding scores are 79.7–81.6% versus 6.8% (Appendix A.3).

Table 4 separates simulation and real recordings. Pair-derived training improves both domains relative to RI-FT; the full mixture reaches 50.9–54.1% real-robot conflict PairAcc across seeds, compared with 13.2% for RI-FT. It does not improve on the two simpler pair-derived recipes in that domain. Source-level differences, including the difficult DROID-100 group, are detailed in Appendix F.

Table 4: PairAcc (%) by domain on the same 552 pairs. Clean removes history; Conflict retains fixed history. The full mixture's gains over RI-FT extend to real recordings, while the best component recipe depends on domain and history condition.

	Simulation ($n = 332$)		Real robot ($n = 220$)	
Method	Conflict	Clean	Conflict	Clean
Base	0.6	29.5	12.7	40.0
Plain SFT	15.7	40.1	32.7	40.0
RI-FT	2.1	34.6	13.2	37.3
VCF-only	90.4	80.4	54.5	51.8
Plain SFT + VCF	91.0	47.9	57.3	40.0
RI-FT-CF (42)	93.4	74.7	53.2	50.9
RI-FT-CF (43)	93.7	69.0	50.9	45.9
RI-FT-CF (44)	94.9	73.2	54.1	44.1

A separate construction batch yields 81.5% PairAcc on 623 pairs and grid FC 94.8%. A fixed-test dose sweep improves over the tested paired-data range. Appendix F specifies these same-backbone results and their different pools, budgets, and split checks.

6.4 Image dependence and remaining preferences

Without images, RI-FT-CF seed 42 matches the later target on 90.6% of pairs despite a low belief-match rate of 2.4% (Table 10). With images, its 77.4% PairAcc requires both distinct targets to be correct. The image can therefore override a strong textual preference, although this test alone does not identify which visual cues drive the change. Clean PairAcc complements this test by removing the shared history.

Exploratory group-relative policy optimization (GRPO) (Shao et al., 2024) improves paired scores but reduces CH performance under the recorded rewards and unequal update budgets. Appendix F.3 retains the results and diagnostics; prompting and contrastive-decoding comparisons are in Appendix E.

7 Related Work

Embodied belief updating. Estimate-Verify-Update (EVU) compares expected outcomes with observations to revise explicit beliefs in text-only embodied environments (Wang et al., 2026). Theory of Space studies spatial belief construction and revision during exploration (Zhang et al., 2026). Our independent image queries separate history interference from image discrimination in state answers.

Context and visual evidence. Language models are sensitive to irrelevant context (Shi et al., 2023) and evidence position (Liu et al., 2024b). CHAIR and POPE assess visual hallucinations through object labels and presence questions (Rohrbach et al., 2018; Li et al., 2023b); vision–knowledge conflicts concern images contradicting parametric knowledge (Liu et al., 2025). MM-SY tests susceptibility to incorrect user opinions and finds that mitigation can reduce receptiveness to correct feedback (Li et al., 2025). Our action-history controls reveal poor image discrimination even when fine-tuning preserves high correct-history accuracy. Visual-sycophancy and visual-ignorance analyses further distinguish retained information from its use in decoded answers (Hong & Quan, 2026; Zhou et al., 2026); our interventions provide supporting evidence about history-conditioned answer selection.

Controlled evaluation and supervision. VQA v2 uses complementary images with different answers (Goyal et al., 2017); VQA-CP shifts answer priors across splits (Agrawal et al., 2018). Counterfactually augmented text changes target labels through controlled edits (Kaushik et al., 2020). Symmetrical Visual Contrastive Optimization (S-VCO) uses minimal visual contrasts with a symmetric objective rewarding matched image–response pairs and penalizing contradictory pairings, and also evaluates direct SFT on both sides (Wu et al., 2025). Our pairs use recorded robot frames with history, question, and options fixed. RI-FT-CF combines these rows with history variation using ordinary answer cross-entropy. We examine history/image performance tradeoffs rather than compare objectives against S-VCO.

Inference-time correction. Self-Refine iterates feedback and revision (Madaan et al., 2023). Context-aware, instruction-contrastive, and visual-contrastive decoding contrast predictions under changed context, instructions, or images (Shi et al., 2024; Wang et al., 2024; Leng et al., 2024). RI-FT-CF changes training data while retaining single-image, single-branch inference.

8 Discussion, Limitations, and Conclusion

History-resistant training can produce high conflict and correct-history accuracy while leaving fixed-text image discrimination poor. Adding pair-derived examples improves the latter, and the full mixture retains strong history-grid performance. Component comparisons reveal a tradeoff across conditions, making joint evaluation essential.

The evidence concerns four-option state questions with explicit histories, principally on Qwen3-VL-8B for training and interventions. Annotation is model-assisted, and outcome-oriented questions can complicate earlier-frame interpretation. CH and FC use different formats, so their joint accuracy does not establish general history-trust calibration. Recorded-group checks delimit training overlap; real-robot results remain below simulation, and natural frame pairs do not isolate object state from other visual changes. The activation and replay studies do not establish the mechanism of the training gain or improved robot task success. Within this scope, the practical lesson is to evaluate history resistance together with correct responses to image changes, and to supply both forms of variation during supervision.

AI USAGE STATEMENT

Gemini was used to draft and verify benchmark annotations, as described in the construction protocol. Kimi Code (Moonshot AI) assisted with research code, experiment execution, literature checking, and earlier manuscript drafting. OpenAI Codex assisted with restructuring and rewriting this draft, inspecting supplied reviews and experimental artifacts, recomputing reported metrics from archived outputs, producing figures, and checking the manuscript build. No new model-training or evaluation runs were performed as part of this rewrite. The empirical claims refer to the recorded experiments, with their annotation and evaluation limitations stated in the paper. The authors are responsible for the final submission and for checking AI-assisted text, code, analysis, and references.

ETHICS STATEMENT

The benchmark derives from existing robot datasets rather than newly collected robot interactions. Its intended use is analysis of multimodal state judgments, not certification of safe robot behavior. Source datasets have different access and redistribution conditions; derived annotations and source references should be distributed subject to those conditions, without assuming permission to redistribute all original footage. The supplementary human reference uses visual question-answering questionnaires completed by adult colleagues. Participants were informed that their responses would be used for this research and participated voluntarily. We identified no additional ethical concerns in this questionnaire. Only aggregate results are reported here; participant identifiers and raw response files are not included in this manuscript package.

REPRODUCIBILITY STATEMENT

Appendices A–E specify construction, prompts, split units, model settings, ablations, and statistics; Appendix E.4 identifies the archived settings and missing version information. The manuscript folder includes rendered figures, generated tables, and evidence summaries for compilation without model execution. The accompanying repository contains table-reconstruction scripts and the archived per-item and per-pair outputs they read, together with figure-generation code. Those scripts require the repository records rather than the manuscript folder alone. Reproducing model inference or training additionally requires the source images, model checkpoints, and documented runtime dependencies; the manuscript package alone is not a complete redistribution of those assets.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and

Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024. URL https://arxiv.org/abs/2404.14219.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. arXiv preprint arXiv:2503.01743, 2025. URL https://arxiv.org/abs/2503.01743.

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth Gururani, Imad El Hanafi, Zekun Hao, Jacob Huffman, Jingyi Jin, Brendan Johnson, Rizwan Khan, George Kurian, Elena Lantz, Nayeon Lee, Zhaoshuo Li, Xuan Li, Tsung-Yi Lin, Yen-Chen Lin, Ming-Yu Liu, Andrew Mathau, Yun Ni, Lindsey Pavao, Wei Ping, David W. Romero, Misha Smelyanskiy, Shuran Song, Lyne Tchapmi, Andrew Z. Wang, Boxin Wang, Haoxiang Wang, Fangyin Wei, Jiashu Xu, Yao Xu, Xiaodong Yang, Zhuolin Yang, Xiaohui Zeng, and Zhe Zhang. Cosmos-Reason1: From physical common sense to embodied reasoning. arXiv preprint arXiv:2503.15558v1, 2025. URL https://arxiv.org/abs/2503.15558v1.

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

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112v1, 2023. URL https://arxiv.org/abs/2303.08112v1.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe

Yu, and Brianna Zitkovich. RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.025. URL https://roboticsproceedings.org/rss19/p025.html.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006, 2025. URL https://arxiv.org/abs/2507.01006.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.00837.

Rui Hong and Shuxue Quan. To see or to please: Uncovering visual sycophancy and split beliefs in VLMs. arXiv preprint arXiv:2603.18373, 2026. URL https://arxiv.org/abs/2603.18373.

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

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Abraham Lee, Youngwoon Lee, Arhan Jain, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Hatch, Shan Lin, Jingpei Lu, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Walke, Blake Wulfe, Ted Xiao, Jonathan

Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abby O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Andrew Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.120. URL https://www.roboticsproceedings.org/rss20/p120.pdf.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Shuo Li, Tao Ji, Xiaoran Fan, Linsheng Lu, Leyi Yang, Yuming Yang, Zhiheng Xi, Rui Zheng, Yuran Wang, Xiaohui Zhao, Tao Gui, Qi Zhang, and Xuanjing Huang. Have the vision-language models lost confidence? a study of sycophancy in VLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/07be1a0850e58ca29e2b6ce31fc0c791-Paper-Conference.pdf.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023b. doi: 10.52202/075280-1516.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024a.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight: Exploring the vision-knowledge conflicts in multimodal LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 17825–17846, 2025. URL https://aclanthology.org/2025.acl-long.872/.

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 *European Conference on Computer Vision (ECCV)*, 2024c.

Yulin Luo, Chun-Kai Fan, Menghang Dong, Jiayu Shi, Xiangju Mi, Mengdi Zhao, Bo-Wen Zhang, Cheng Chi, Jiaming Liu, Gaole Dai, Rongyu Zhang, Ruichuan An, Kun Wu, Zhengping Che, Shaoxuan Xie, Guocai Yao, Zhongxia Zhao, Pengwei Wang, Guang Liu, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBench: A comprehensive evaluation benchmark for multimodal large language models as embodied brain. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2510.17801v2.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

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

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. URL https://arxiv.org/abs/2402.03300.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 783–791, 2024. URL https://aclanthology.org/2024.naacl-short.69/.

Team AgiBot-World. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3549–3556, 2025. doi: 10.1109/IROS60139.2025.11247088.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736, 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

Hanlin Wang, Chak Tou Leong, Jian Wang, and Wenjie Li. Seeing isn't believing: Mitigating belief inertia via active intervention in embodied agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://aclanthology.org/2026.findings-acl.1884/.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30284–30297, 2025. doi: 10.18653/v1/2025.acl-long.1462. URL https://aclanthology.org/2025.acl-long.1462/.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, and Jianfeng Gao. Magma: A foundation model for multimodal AI agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14203–14214, 2025. URL https://doi.org/10.1109/CVPR52734.2025.01325.

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, 2024.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Fei-Fei Li, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? In *International Conference on Learning Representations*, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713, 2024. URL https://arxiv.org/abs/2410.02713.

Runyu Zhou, Qi Zhang, Qixun Wang, and Yisen Wang. Diagnosing visual ignorance in vision-language models. arXiv preprint arXiv:2606.06890, 2026. URL https://arxiv.org/abs/2606.06890.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025. URL https://arxiv.org/abs/2504.10479.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.

A Dataset Construction and Examples

Relation to visual evaluation and robot data. VQA evaluates image-conditioned questions (Antol et al., 2015), while GQA adds compositional questions and controls answer distributions (Hudson & Manning, 2019). MMBench evaluates a range of multimodal abilities bilingually (Liu et al., 2024c); MMMU emphasizes college-level subject knowledge (Yue et al., 2024); SEED-Bench covers hierarchical multimodal comprehension and generation capabilities (Li et al., 2024). These resources motivate broad capability assessment. VBI-Bench instead holds images or text fixed to test two specific dependencies in robot-state answers. Open X-Embodiment standardizes heterogeneous robot datasets for cross-embodiment learning (Open X-Embodiment Collaboration, 2024). We construct state questions from the individually attributed sources below, retaining source identifiers for domain-level analysis.

A.1 Source composition

Table 5 lists the source groups, their construction signals, and the distinction between successful and failed outcomes. The two LIBERO-90 groups are shown together but retain separate identifiers in the manifests. Counts refer to the frozen 2,285-item history benchmark, not to the paired training or evaluation pools. Figure 3 shows the full source breakdown. Figure 7 summarizes outcomes by source family alongside task and failure counts; Figure 8 adds domain/source breakdowns and question forms.

Table 5: Source composition of VBI-Bench. S/F denote success/failure. Sources are grouped for readability; LIBERO-90 contains two identifiers.

Source	Items	Outcome	Construction signal or notable property
RoboBench low-error	151	All F	Frames from failure-analysis videos
RoboBench high-error	145	38 S / 107 F	More difficult error-video examples
RoboBench timestamp	128	All S	Annotated event intervals
DROID-100	152	All S	Gripper-signal grasp/release transitions
DROID-2K	239	Mixed	Gripper transitions; 102 dynamic failures
LIBERO spatial	244	Mixed	Spatial tasks and two camera views
LIBERO goal	276	Mixed	Goal tasks; environments overlap spatial suite
LIBERO-90 (two groups)	612	Mixed	Task/event boundaries
LIBERO-10	31	Mixed	Task suite held out from history training
BridgeData V2	122	Mixed	Event intervals, including deformable objects
AgiBot World	185	Mixed	Event slicing in longer real-home episodes
Total	2,285	1,030 S / 1,255 F	1,163 simulation items (50.9%)

LIBERO failures are often frames from incomplete execution rather than naturally occurring mechanical failures. DROID-2K includes dynamic failures such as missed grasps and dropped objects. Successful and failed outcomes both receive an incorrect history in the conflict conditions. The resulting task probes state recognition under controlled disagreement; it does not estimate the prevalence of these disagreements in deployment. We retain source identifiers because camera setup, task distribution, state visibility, and option wording all vary by source.

A.2 Annotation workflow and quality boundaries

Transition localization uses gripper events or source-specific event annotations. Gemini drafts the action description, question, target state, false belief, and distractors from the selected frames. A separate describe-then-adjudicate pass first describes the image and then checks the labels against that description. The original construction record reports source-dependent discard rates of 12–39%. This is model-based quality filtering, not an independent human validation set.

Three asserted and two hedged belief phrasings are associated with each item. Integrity checks verify required fields, distinct option roles, image references, and consistency between the outcome label and the false history. Evidence strength is recorded as clear (2,082), weak (149), or invisible (54), using the annotating model's judgment. These categories are not objective guarantees that an

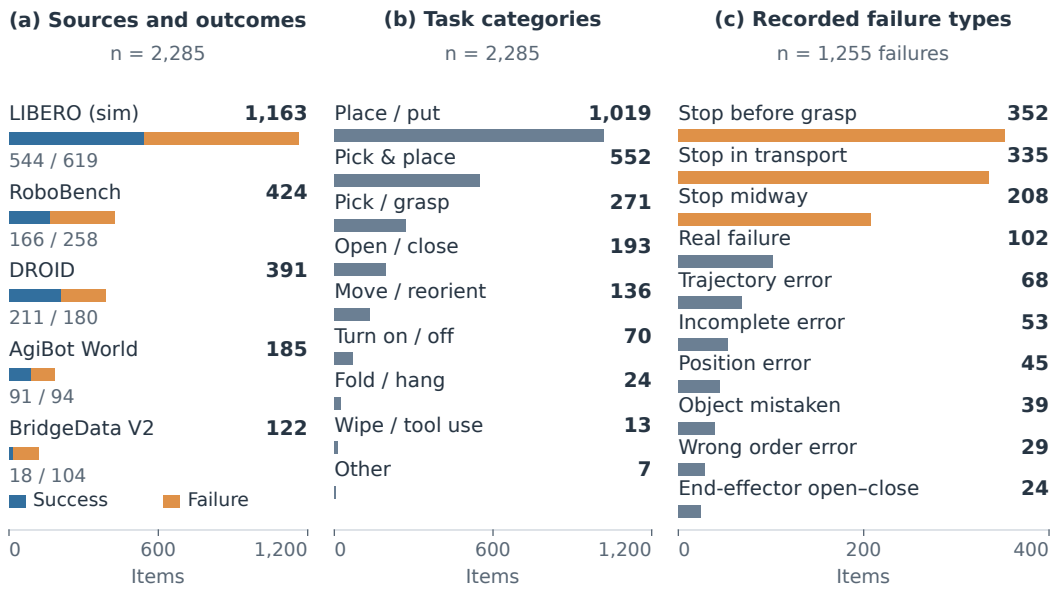

Figure 7: **Aggregated source, task, and failure composition.** (a) Source-family totals, with success/failure counts below each bar. (b) Task categories from action-verb rules. (c) Source-defined failure types; orange marks premature stops. Panels (a,b) cover 2,285 benchmark items, and (c) covers 1,255 failed outcomes.

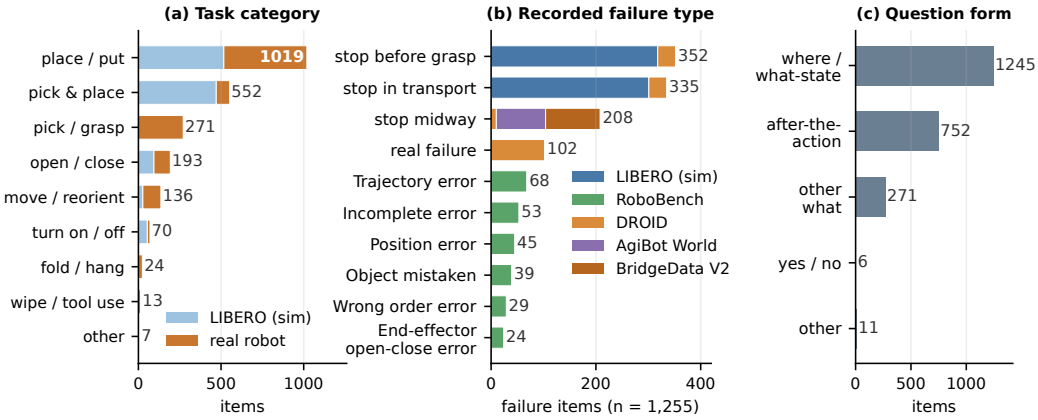

Figure 8: **Detailed task, failure, and question composition.** (a) Task categories by action-verb rules, split into simulation and real recordings. (b) Recorded failure types by source family (1,255 failed outcomes); the labels are source-defined. (c) Question forms by wording rules. Panels (a) and (c) cover all 2,285 items.

evaluated model can infer the answer from a single image. In particular, a robot gripper hovering over a released object may make release difficult to distinguish from continued grasping.

The construction history records two relevant repairs. Some early RoboBench fields contained generation scaffolding and were regenerated before the frozen evaluation. In LIBERO spatial examples, descriptions of an object being held could conflict with the visual appearance of a just-released object under a hovering gripper; option wording was revised to describe the visible relation more precisely. The frozen evaluation uses the repaired records; pre-repair snapshots are separate. Independent human agreement on the frozen labels remains unmeasured.

A.3 Recorded visibility and stratified results

Table 6 separates full-benchmark visibility counts from the evaluated subsets. The pair records inherit the originating item's evidence-strength field; this field does not independently assess both frames. VCF has a separate model-assisted two-image adjudication. Accordingly, the strata describe the existing annotations rather than a human-validated answerability partition.

Table 6: Recorded visibility strata. Counts refer to grid items and pairs; the remaining columns are conflict-history PairAcc (%) within each pair stratum. The RI-FT-CF columns show training seeds.

	Test counts			RI-FT-CF		
Recorded label	Grid	Pairs	RI-FT	42	43	44
Clear	659	512	6.8	80.7	79.7	81.6
Weak	41	27	0.0	33.3	33.3	37.0
Invisible	13	13	7.7	38.5	46.2	46.2

On the 512 pairs marked clear, RI-FT scores 6.8% and RI-FT-CF scores 79.7–81.6%. The paired improvement therefore persists within this existing annotation stratum. Weak and invisible strata are small (27 and 13 pairs), and their scores are sensitive to individual examples. This descriptive breakdown leaves the test set and labels unchanged.

A.4 Record schema

Each history-grid record stores a sample identifier, source identifier, action, earlier and later image references, question, outcome, and options with semantic roles (target, belief, two distractors). It also stores a belief text, asserted and hedged variants, expected-outcome text, and the evidence-strength label. A pair record adds a pair identifier, a scene key, the correct option role for each frame, and verification metadata. A scene key can refer to a source episode or a LIBERO task demonstration. It is an operational grouping key; sharing a room or physical setup across different keys remains possible.

A.5 Two directions of conflicting history

The following example is an unsuccessful placement with a false success history (sample `rb_low_0000`):

> Action: place bread plate.
> History: The robot has already completed the task: place bread plate. The task was finished successfully.
> Question: After the action, where is the bread?
> Target: The bread is on the gray mat next to the plate, not placed on the plate.
> Belief option: The bread is placed on the white plate, centered and stable.
> Distractors: The bread is still in the robot's gripper, held above the plate; the bread has fallen off the table and is on the floor.

The opposite direction is a successful grasp with a false failure history (sample `droid_0000`):

> Action: grasp the marker.
> History: The previous action failed. The marker is still lying on the wooden surface, untouched by the gripper.
> Question: After the action, where is the marker?
> Target: The marker is securely held in the robot's gripper, with its tip pointing downward and the body visible above the gripper.
> Belief option: The marker is still lying on the wooden surface, untouched by the gripper.

The latter excerpt omits distractors for brevity. Showing both outcome polarities prevents an always-success rule from solving every example. It does not prevent a rule that infers which option contradicts the supplied history, nor does it ensure that question and option wording are uninformative about the target.

A.6 PAIRED IMAGE CONSTRUCTION AND THE DISPLAYED EXAMPLE

The VCF construction selects two frames with different verified answer roles under one shared option set. The earlier frame need not depict failure, and its target need not equal the belief option; it can instead be a distractor describing an intermediate state. Both frames retain the original question even when it uses outcome-oriented wording. This is a known source of semantic tension for earlier-frame queries, particularly questions asking for a final state. The annotation and evaluation prompts differ in how explicitly they resolve this tension. The verifier is instructed:

> For EACH image independently, decide which single option best describes the CURRENTLY VISIBLE state of the scene with respect to the question. Treat the question as asking about the state actually visible in that image (even if its wording says "after the action" or "final state", answer based on what the image shows NOW).

It can reject a frame when no option fits. The evaluated model instead receives the following instruction, preceded by history in the conflict condition:

> Look at this image and answer the question.
> *[Original question]*
> *[Four shuffled options]*
> Answer with only the letter.

For example, one pouring pair asks "What is the final state of the objects that were in the yellow cup?" Its earlier target is that the objects remain in the cup; its later target is that they were poured into the bowl. The verifier's current-image interpretation makes those targets distinct, whereas a literal final-outcome reading can favor the later answer in both queries. The words "after" or "final" occur in 308 of 552 pair questions; this lexical count does not determine which individual questions are ambiguous. Thus, earlier-frame errors can mix visual-state discrimination, historical interference, and temporal interpretation. The reported predictions use the original evaluation prompt throughout.

Figure 1 uses pair `vcf_t0485` (sample `rb_ts_0096`, RoboBench timestamp). The task is pouring ice from a scoop into a goblet. The earlier frame shows an empty goblet; the later frame contains ice. The archived outcome report states that the action was attempted but did not succeed, and the belief describes failure. The evaluation uses a deterministically selected paraphrase of that belief; the display abbreviates it to the stored canonical wording. Both frames have the same four options: ice fallen onto the table, an empty goblet with ice in the scoop, ice in the goblet, and a goblet filled with water. The correct option changes from B to C. Archived base predictions are B/B, RI-FT C/C, and RI-FT-CF seed 42 B/C. Figure 1 expands these letters into option meanings. Its photographs are original recorded frames; the accompanying evidence record preserves their mapping and predictions.

B EVALUATION PROTOCOL, SPLITS, AND STATISTICS

B.1 HISTORY CONDITIONS AND PROMPT RENDERING

Table 7: Core conditions under fixed images, questions, and options. Labels follow Figure 4a; Correct belief denotes an accurate statement.

History condition	Measurement purpose
No history (NH)	Reference state QA without supplied history
Correct belief (CH)	Performance with a correct belief statement
Belief (WB)	Interference from a wrong belief alone
Action	Action-description contribution
Report	Outcome-report contribution
Action + Belief	Action plus wrong belief
Report + Belief	Outcome text plus wrong belief
Action + Report	Action plus outcome text
Full history (FC)	Full conflicting history

The core prompt asks the model to look at the image, answer the state question, and emit only the option letter. FC organizes the supplied text into previous-action, expected-outcome, and previous-

belief blocks. The expected-outcome block is often already phrased as a realized success or failure, motivating the framing probes rather than a literal interpretation as an uncertain expectation.

Option roles are shuffled with a deterministic seed based on the item identifier. History conditions use the same permutation for an item. The pair evaluation seeds from the pair identifier and keeps the permutation fixed across the two images. Training pair construction uses the same rendering rule. The earlier-image row stores its image reference in the image field consumed by the ordinary SFT loader; it is not an additional image in the same conversation.

All primary pair results use the conflict-history arms for both images. Clean pair results remove the history from both arms. The image-free arm removes the image and adjusts the image-reference instruction; it yields one response that is compared with both frame targets and the belief option. Because the two frame targets differ, those two target-match rates sum to at most 100%. The belief option may coincide with a frame target, so adding a third belief-match column need not preserve that bound. A high one-sided match rate is not an image-free pair-accuracy score.

B.2 Denominators and split units

The frozen benchmark has 2,285 items. Its history training partition contains 1,572 items, and its test partition 713 items (332 successful and 381 failed outcomes). All items from LIBERO-10 are assigned to the test side. Main intervention comparisons use this 713-item denominator. The full-set base-model evaluation is used only when explicitly labeled as such.

The primary VCF evaluation contains 552 distinct pair identifiers with distinct target letters for the two frames. The paired training component contains 564 pairs. The training-union audit finds no evaluation scene key in the paired component, but 214 evaluation pairs derive from history-training sample identifiers and 359 share scene keys with that component. Consequently, only 193 pairs have keys unseen by the complete mixture. Reporting both columns in Table 3 separates the original paired test from this stricter grouping check. The 193-pair subset is a retrospective restriction of the same evaluation set and changes its source composition. Table 8 summarizes the grouping checks; Table 15 gives source counts. Of the 193 pairs, 131 are simulated (67.9%), compared with 332/552 (60.1%) in the full test set. Its higher accuracy therefore does not isolate an effect of group novelty.

Table 8: Primary split relationships. Pair evaluation is audited against both training components. Sample IDs and recorded scene keys represent different grouping levels.

Relationship	Count
History-grid train / test items	1,572 / 713
Paired train / test pairs	564 / 552
Paired-training sample IDs in grid test	0
Evaluation pairs sharing paired-training scene keys	0 / 552
Evaluation pairs derived from history-training sample IDs	214 / 552
Evaluation pairs sharing history-training scene keys	359 / 552
Evaluation pairs unseen by both training components	193 / 552

The zero sample-ID intersection does not establish scene-level separation between paired training and the history-grid test. Scene keys are operational episode or task-demonstration groups; distinct keys can still share physical environments. The independent 623-pair evaluation is described separately in Appendix F.

A separate suite-exclusion check for ordinary RI-FT removes both LIBERO goal and spatial suites from training; FC accuracy on the 157 test items from those suites is 98.7% (155/157), compared with 99.4% (156/157) for mixed training. This bounds one identified overlap route, not all possible episode or environment overlap. Public source footage could also have appeared during model pretraining. We do not infer pretraining disjointness from newly generated questions alone.

B.3 Parsing and statistical procedures

Recorded option letters are scored against the stored targets. In the primary grid and pair comparisons, a missing, truncated, or invalid answer is an error, not a reason to shrink the denominator.

This matters particularly for models producing lengthy reasoning traces. The caption of Table 1 gives invalid NH/FC counts for the cross-model evaluation.

For the principal seed-42 comparison, we recompute paired percentile bootstrap intervals with 10,000 resamples (random seed 20260922). The sampling unit is an item for the grid and a complete pair for VCF. We also compute exact two-sided McNemar tests from discordant correctness indicators. Table 9 reports changes from RI-FT to RI-FT-CF. The corresponding p-values are 6.04×10^{-7} for the FC decrease and 3.91×10^{-116} for the pair increase. These are descriptive comparisons of fixed checkpoints. The bootstrap does not capture label uncertainty, within-source dependence, or all training variation; three training seeds are reported separately. Seed ranges describe training variation on the same test set, not additional independent test observations. Other component recipes have one archived checkpoint each, so their differences are descriptive and do not establish statistical superiority across training seeds.

Table 9: RI-FT-CF minus RI-FT, seed 42. Discordant counts are improvements/regressions. Intervals are paired bootstrap 95% intervals in percentage points.

Metric	n	Change	95% interval	Discordant
Grid FC	713	-4.2	[-5.9, -2.5]	4 / 34
Pair accuracy	552	+70.8	[+67.0, +74.6]	392 / 1

C Additional Behavioral Evidence

C.1 Cross-model results

Table 1 reports all nine models in the main text. Their absolute levels and conflict-induced drops differ. The Qwen3-VL-8B result is not intended to stand in for the entire model population. General multimodal, video-oriented, and embodied-oriented checkpoints differ in preprocessing and response behavior; their own no-history baseline is therefore an important reference. Main mechanistic and fine-tuning conclusions are restricted to the explicitly tested Qwen configurations.

C.2 Paired marginals and no-image responses

Table 10 extends the main diagnostic table with no-image matching against both frame targets and the belief option. All pair columns use the same 552 pairs under conflict history. A belief target can coincide with a frame target, so those match rates are not disjoint categories.

Table 10: Paired marginals and no-image target matching (%, 552 pairs). No-image columns match one image-free answer to each reference; belief can coincide with a frame target. RI-FT-CF uses seed 42.

	Images present			No-image target match		
Method	Later	Earlier	PairAcc	Later	Earlier	Belief
Base	8.3	73.0	5.4	2.0	69.7	97.8
Plain SFT	52.7	62.1	22.5	9.6	64.5	90.2
RI-FT	99.1	7.2	6.5	94.6	1.6	2.7
RI-FT-CF (42)	96.2	80.6	77.4	90.6	6.9	2.4

C.3 History decomposition and input perturbations

The action-only condition differs sharply by true outcome: it achieves 95.5% on successful examples and 22.6% on failures. This is consistent with action text encouraging a successful-outcome answer. The explicit belief increment on top of action-plus-outcome-report is smaller and outcome-dependent. Such interactions caution against treating the three history ingredients as independent additive causes.

Table 11: History-factor accuracy on Qwen3-VL-8B, 713 items. Labels follow Figure 4a; model parameters are fixed.

History	No history	Action	Belief	Action + Belief	Report	Report + Belief	Action + Report	Full history
Accuracy	71.0	56.5	56.1	36.7	20.2	18.1	17.4	14.4

The original-image FC accuracy of 14.4% falls to 6.9% with the image swapped and 2.9% with a blank image; the corresponding belief-option rates increase from 82.9% to 88.5% and 95.5%. Images therefore affect the answers. This perturbation does not show that the intact image was correctly perceived on every erroneous trial, and an arbitrary image swap can change difficulty as well as state evidence.

No-image, no-history accuracy is 45.6%. Two sources contribute particularly strong text-only signals: DROID-100 reaches 93.5% and LIBERO spatial 85.1%. Excluding both leaves 593 items. On this subset with images present, no-history accuracy is 66.3%, with a 49.4-point drop under full-conflict history. The history effect therefore persists beyond these sources, although the remaining items can still contain textual cues. This is why the main claim does not equate NH with pure perception.

Outcome-report templates. The Report-only factor (20.2%) supplies `Expected outcome:` followed by the item's archived `expectation_text`. This field can describe either successful completion or an attempted action that failed. The separate framing probe (51.1%) instead supplies `Action outcome report:` followed by "The action succeeded; the task '{action}' was completed as intended." Both omit the action and belief blocks and retain the image, question, and options. They differ in both the header and the report content, so their gap does not isolate a header effect.

C.4 Human reference and self-generated history

The single-frame human reference comprises 20 participants, with 10 assigned to no history and 10 to conflicting history. Each participant answers two versions of the same 115-question set; the two measurements are averaged within participant before computing group means. Reported mean accuracies are 87.9% without history and 77.3% with conflicting history. Repeated questions and between-participant assignment prevent treating all responses as independent matched trials. The archive also contains clearly identified model-simulated questionnaire files, which are excluded from this human reference. No population-level claim that humans are immune to the phenomenon follows from these numbers.

Separate two-turn probes allow a model to form its own first-turn prediction before observing the outcome image. Correction is scored only when that prediction selected the subsequently contradicted belief option, making the denominator model-dependent and selected. These probes should not be merged with the injected-history grid. Likewise, the provenance study distinguishes a prediction, an outcome report, an attributed statement, and an impersonal log. Their different effects argue against interpreting every textual-history failure as persistence of a model's own internally formed belief.

D Mechanistic Protocol and Interpretive Limits

D.1 Activation patching and specificity

The primary patching experiment selects 60 Qwen3-VL-8B items answered correctly under NH and with the belief option under FC. It intervenes on attention or MLP block outputs across layers at the answer position, using a matched run as the activation donor. Layer numbers are zero-indexed: layer 24 is the 25th transformer block. The hook replaces the block output at the final prompt token (the position predicting the first answer token), aligned by relative position rather than by absolute sequence length. The corruption direction inserts the conflicted-run activation into the clean run; the rescue direction inserts the clean-run activation into the conflicted run. Layer-24 attention is the

strongest localized intervention in this selected set, with corruption 44/60 and rescue 38/60. It was selected by scanning the same items across all 36 layers; no separate discovery/validation split was used. The archived script filters a 100-ID candidate list to these 60 qualifying items. Figure 5 shows the complete layer curves.

The non-belief control selects a separate 60-item set in which an original question receives an incorrect non-belief answer while a synonymous question is answered correctly. Applying the same bidirectional procedure yields layer-24 attention corruption 48.3% and rescue 73.3%. Thus, transplanted activations can carry general answer-selection information. The two sets are differently selected and do not provide a population-level comparison of equal-difficulty errors. The result supports a local causal contribution of these activations under the tested interventions, not a dedicated belief representation.

D.2 Decodability is not proof of intact perception

An answer-letter linear probe peaks at 67.3% target decoding at hidden-state index 33. Probe indices include the embedding at index 0, followed by 36 block outputs; index 33 therefore follows zero-indexed transformer block 32. This convention differs from the block indices used for patching. Above-chance decoding can arise from text-correlated targets or generic answer information; it does not establish accurate visual-state representation or use.

A separately trained binary completion probe attains peak accuracies across layers of 78.5% with the original image, 76.6% without an image, and 76.2% with a swapped image, against a 68.7% majority-class reference. The respective peaks occur at hidden-state indices 33, 21, and 20, so their differences are not comparisons at a fixed layer. High peak decoding accuracy without the original image, together with the probe-label quality, limits inference of intact perception.

Our logit-lens analyses directly project intermediate states through the output head. Belrose et al. (2023) analyze the brittleness of such projections and introduce learned translators in the tuned lens. We use uncalibrated projections: their layerwise changes can locate candidate interventions but do not directly read out ground-truth perception. For GLM-4.1V-Thinking, the template places a thinking token at the tested position; letter-logit probes there are off-policy relative to normal answer generation. We therefore do not use its lens or contrastive decoding curves to establish cross-model mechanistic equivalence.

D.3 Training-location and framing controls

Training only layer-24 LoRA parameters attains 93.4% FC accuracy, while training all layers except layer 24 attains 98.3%. Sufficiency under one restricted parameterization does not imply necessity under retraining. Head-level effects are distributed rather than reducible to one indispensable head. Together with the ordinary-error control, these results motivate the narrower description of an influential late answer-selection computation.

Rephrasing a belief can change its effect, but the available probes do not support a general law that more detailed or more confident beliefs are always harder to revise. A detail-deletion manipulation produced little change; hedging did not reliably help and increased belief following in one tested setup. A more abstract outcome-level rewrite reduced belief following, but simultaneously changed lexical overlap with options and improved accuracy by much less. These are template-specific findings. We do not use them to establish a general precision-weighting theory or to derive the pairing method uniquely.

E Training Configuration and Intervention Baselines

E.1 Supervised recipes

All principal adapters start from Qwen3-VL-8B-Instruct. LoRA uses rank 16, scaling 32, dropout 0.05, no bias adaptation, and the language-model projections `q,k,v,o,gate,up,down`. The archived trainer uses bf16, per-device batch size 1, gradient accumulation 16, cosine learning-rate scheduling from 10^{-4}, warmup ratio 0.05, and gradient checkpointing. The collator masks the prompt positions in the loss and supervises the assistant answer sequence produced by the model's

chat template. The target content is an option letter; template end tokens may also be included by the tokenizer. The base vision encoder is not the target of the stated language-projection adapters.

Table 12: Training data components. Counts denote training rows, rather than independent scenes. VCF-only and mixed recipes differ in both data composition and optimization exposure.

Recipe	History rows	Pair rows	Composition
Plain SFT	11,004	0	No-history questions repeated to the RI-FT row budget
RI-FT	11,004	0	NH: 1,572; CH: 1,572; WB: 4,716; FC: 3,144
VCF-only	0	1,128	564 pairs, two images per pair
Plain SFT + VCF	11,004	1,128	Repeated no-history QA plus paired rows
RI-FT-CF	11,004	1,128	Full history mixture plus paired rows
Independent mixed pipeline	11,004	1,532	Separate 766-pair construction

The principal full recipe uses one epoch and 759 steps; the independent mixed pipeline uses 784 steps. VCF-only uses two epochs (2,256 row presentations, 142 steps), while plain SFT plus VCF uses one epoch (12,132 row presentations, 759 steps), matching the full recipe's row and step budget. The plain rows for this latter control were reconstructed from the archived RI-FT rows after the original temporary plain-SFT file was lost. The reconstruction matches the documented sample set, targets, and repetition count, but cannot be byte-compared with that lost file. Component comparisons identify what information the supervision supplies rather than establish optimal compute allocation. The principal full-recipe results report seeds 42, 43, and 44.

E.2 Correct-history and phrasing ablations

Table 13: Ablations of ordinary RI-FT, 713-item history grid. These are not direct ablations of RI-FT-CF.

Recipe	NH	CH	WB	FC
RI-FT	93.1	98.7	95.9	98.3
No phrasing augmentation	95.1	95.7	94.7	97.8
No correct-history views	93.8	26.2	96.6	98.3

Removing CH demonstrates the danger of evaluating only false-history questions: a model can retain high FC accuracy while becoming much less accurate with correct context. The much smaller effect of removing phrasing augmentation suggests that polarity coverage matters more than this particular augmentation in the measured grid. Establishing the same necessity within the paired mixture would require its own matched no-CH ablation.

E.3 Prompting and contrastive decoding

We retain representative non-training interventions as context for the supervision study. An EVU-inspired verify/update prompt (Wang et al., 2026) reaches 44.2% FC accuracy; this evaluates the prompting variant, not the original work's training variant. Other prompt rewrites alter the source attribution or remove the action/outcome-report blocks. Since those manipulations change available text, their gains should not be interpreted as equivalent interventions on an unchanged deployment input.

Belief-contrastive decoding (BCD) combines letter-aligned logits from the full-history and no-history prompts:

$$z_{\mathrm{BCD}} = (1 - \lambda) z_{\mathrm{FC}} + \lambda z_{\mathrm{NH}}. \tag{5}$$

It uses two inference branches. At $\lambda = 1$ it is exactly the no-history channel, and at $\lambda = 1.5$ it reaches 78.0% on the grid. Its incremental gain over the 71.0% no-history baseline is 7.0 points, not the entire gain from FC. This linear fusion is closely related to context-aware decoding with the direction of context amplification reversed (Shi et al., 2024); it is not a distinct algorithmic contribution of this paper.

In the matched decoding harness, instruction-contrastive decoding at coefficient 0.5 reaches 78.7% and visual-contrastive decoding at coefficient 0.5 reaches 27.8% (Wang et al., 2024; Leng et al., 2024). The archived VCD baseline is the implemented image-perturbation variant; these results should not be generalized to every filtering or coefficient choice in the original method. At coefficient 1.0, the recorded ICD/VCD runs produce 410 and 441 invalid outputs out of 713, respectively, which must count as errors. Such settings primarily expose a format failure rather than stronger evidence correction. The archived BCD pair experiment does evaluate all 552 primary pairs. It combines full-vocabulary log probabilities at each decoding step as $(1 - \lambda)\ell_{\mathrm{FC}} + \lambda\ell_{\mathrm{NH}}$, using the same pair-specific option order and belief paraphrase as the primary evaluation. Decoding is greedy with a 16-token cap; the first standalone A–D letter is parsed, and missing letters count as errors. Table 14 reports every archived coefficient. At $\lambda = 1$, the output distribution reduces to the no-history channel and PairAcc matches the base clean score of 33.7%. At $\lambda = 1.5$, PairAcc is 23.2%. This generative implementation differs from the grid's letter-aligned readout, so the two scores are not a single protocol-matched joint evaluation. EVU, ICD, and VCD do not have equivalent full-pair results in this comparison.

Table 14: Archived BCD paired decoding on 552 pairs (%). Later/earlier are per-frame accuracies; PairAcc requires both correct. Invalid-output counts are later / earlier.

λ	Later	Earlier	PairAcc	Invalid outputs
0.5	22.3	75.9	16.8	0 / 0
1.0	70.3	53.1	33.7	0 / 0
1.5	79.0	32.6	23.2	0 / 0

E.4 Rendering and reproducibility details

The primary supervised and pair-evaluation scripts load the Qwen3-VL-8B-Instruct processor together with the model. They use its chat template and image preprocessing defaults without explicit pixel-limit or input-length overrides. The training collator pads token sequences to the batch maximum and masks prompt and padding positions in the loss. The primary VCF harness uses greedy generation with at most 16 new tokens and parses the first standalone option letter. "Single-branch inference" means one prompted decoding stream per image; it can contain multiple autoregressive steps. Pair scoring combines two independent image queries.

The main supervised recipes use their recorded final checkpoints at the row/epoch budgets above. Seeds 42–44 are exposed in the full-mixture trainer; the other component results use their single recorded checkpoints. RL checkpoint selection follows the different protocol in Appendix F.3. The annotation client has a default identifier of `gemini-3-flash-preview`, with an environment override; saved verification records generally identify only Gemini. Consequently, the code default does not establish the exact served version for every annotation batch. Pair verification scripts and records retain the two-image prompts, choices, explanations, and a 4,000-token response cap. Library commits, processor revisions, and all remote-model versions are not comprehensively pinned in the archive.

The manuscript source contains the rendered figures and generated tables, so compiling the PDF needs no model execution. Its table scripts read archived item and pair predictions, the 713-item split manifest, pair manifests, and the training-union audit. The experimental archive does not contain the complete original 2,285-item master annotation manifest; full annotation reconstruction and every cross-component scene mapping cannot be independently recovered from this package. Original images, model weights, and runtime dependencies are additional requirements for rerunning inference or training.

F Further Evaluation and Remaining Biases

F.1 Primary pair performance by domain and source

Table 4 compares all principal recipes under both history conditions on the same simulated and real-robot pairs. RoboBench is included in the real-robot group. Pair-derived training improves both

domains relative to RI-FT, while the full mixture's advantage is not uniform across these component comparisons. Its real-source conflict PairAcc is 50.9–54.1% across seeds, compared with 54.5% for VCF-only and 57.3% for plain SFT plus VCF.

Table 15: Source counts and conflict-history PairAcc (%). N is the full-test count and U the training-union-unseen count. Method scores use all N pairs, not just U. Plain denotes Plain SFT; VCF denotes VCF-only; Full denotes RI-FT-CF seed 42.

Source	N	U	Base	Plain	RI-FT	VCF	Plain+VCF	Full
LIBERO spatial	74	51	2.7	2.7	0.0	100.0	100.0	100.0
LIBERO goal	67	17	0.0	19.4	0.0	88.1	91.0	97.0
LIBERO-10	21	21	0.0	9.5	0.0	95.2	81.0	95.2
LIBERO-90	85	21	0.0	22.4	3.5	89.4	89.4	89.4
LIBERO-90b	85	21	0.0	18.8	4.7	83.5	87.1	88.2
BridgeData V2	28	10	35.7	64.3	53.6	85.7	89.3	89.3
DROID-2K	40	6	12.5	32.5	12.5	60.0	62.5	62.5
AgiBot	32	1	15.6	56.2	6.2	59.4	62.5	56.2
DROID-100	65	5	0.0	9.2	0.0	36.9	40.0	23.1
RoboBench high-error	15	11	40.0	53.3	40.0	66.7	73.3	80.0
RoboBench timestamp	40	29	5.0	22.5	2.5	47.5	47.5	55.0

For full-mixture seed 42, the real total is 117/220: 83/165 on non-RoboBench sources and 34/55 on RoboBench. Table 15 shows substantial variation within each group, including 15/65 on DROID-100. Visibility, camera viewpoint, task distribution, and wording vary together across sources; the breakdown establishes a domain boundary without isolating its cause. The unseen subset's higher simulated share also prevents interpreting its higher overall accuracy as an intrinsic benefit of novel groups.

F.2 Independent pipeline and supervision dose

The independent pipeline constructs 766 training pairs and 623 evaluation pairs using a separate candidate and verification batch. Mixing its 1,532 pair-derived rows with the 11,004-row history mixture gives 12,536 rows. The archived mixed adapter scores 81.5% both-correct accuracy with conflict histories, 74.0% with clean histories, and grid NH/CH/FC scores of 93.7/98.0/94.8%. Its pair-component split is scene-key disjoint, but the stricter training-union audit reported for the primary 552-pair set should not be assumed automatically for this different pool.

The VCF-only dose sweep evaluates the same 200 held-out pairs at five recorded training settings. Actual pair-derived row counts are 102, 210, 414, 818, and 1,532; both-correct accuracies are 25.0, 48.5, 64.0, 76.0, and 80.5%, respectively. These finite settings show increasing performance over the tested range. They do not establish absence of saturation or a constant return per doubling. Data quantity, example composition, and training steps are not fully separated by this sweep.

At small doses, low no-image matching to one target can reflect a switch to the other target rather than disappearance of shortcuts. The primary no-image table similarly shows that RI-FT-CF retains a later-frame preference when vision is absent. Pair accuracy is informative because it requires two correct responses under fixed text, not because the text has become incapable of influencing the model.

F.3 Exploratory process-reward training

We explore whether reinforcement learning can encourage correct state answers and image-dependent switching. These runs are separate from the supervised RI-FT-CF recipe. Table 16 gives their joint CH/PairAcc results; the procedures and diagnostics below define the scope of that comparison.

Relation to reward and preference learning. InstructGPT combines supervised demonstrations with reward learning from human preferences (Ouyang et al., 2022), using PPO for policy optimization (Schulman et al., 2017). DPO optimizes preference pairs directly without fitting a separate

Table 16: Exploratory reinforcement learning on the paired state task.

Training configuration	Steps	CH (%) ↑	PairAcc (%) ↑
Base (reference)	—	89.5	5.4
RL without pair-switch reward	800	66.8	12.6
RL with pair-switch reward	150	69.7	31.3
RI-FT-CF (reference, seed 42)	759	98.0	77.4

CH uses 713 items; PairAcc requires both answers correct on 552 pairs. RI-VR names the archived process-reward RL runs: +nosw disables the switch bonus; +VCF includes it. Steps are optimizer updates with unequal RL/SFT budgets. Configuration and checkpoint selection are in Appendix F.3.

reward model (Rafailov et al., 2023). Our exploration uses GRPO with programmatic answer, judgment, and format rewards. It tests a different supervision signal from human preference learning. Limitations of reward proxies and their optimization are discussed by Casper et al. (2023); here, paired and correct-history evaluation check whether the recorded rewards produce the intended behavior.

Training pool and optimization. RI-VR+VCF initializes Qwen3-VL-8B-Instruct and uses group-relative policy optimization (GRPO; Shao et al., 2024). The recorded pool contains 9,600 rows: 3,960 standard NH/CH/FC rows derived from the 1,572-item history-training split, plus five copies of the 1,128 rows from the 564 training pairs. The pool builder assigns roughly 20/30/50% of standard items to NH/CH/FC and varies the history phrasing across copies. This mixture differs from the 12,132-row supervised mixture. Pair frames remain adjacent; the data loader does not reshuffle individual rows. A pair crossing a batch boundary receives no pair bonus in that batch.

The logged run uses LoRA rank 16 and scaling 32 on the language-model attention and MLP projections, learning rate 10^{-5}, KL-loss coefficient 0.001, 64 prompts per batch, and eight sampled rollouts per prompt. Rollout temperature and top-p are both 1.0; prompt/response limits are 4,096/512 tokens. Training uses one GPU with rollout tensor parallelism set to one, 150 updates, and a logged data-loader seed of 42. The actor uses the GRPO advantage estimator and an explicit KL loss against the reference policy. Standard evaluation is greedy and retains the 713-item or 552-pair denominator, with invalid answers scored as errors.

Reward implementation. The response includes a "confirms" or "contradicts" judgment and a final option letter. Each rollout receives

$$R = r_{\text{answer}} + r_{\text{judgment}} + r_{\text{format}} + r_{\text{calibration}} + r_{\text{switch}}. \tag{6}$$

A correct final letter earns $r_{\text{answer}} = 1$. A correct judgment earns $r_{\text{judgment}} = 0.5$: FC targets "contradicts," CH targets "confirms," and paired rows derive the target from whether the belief option is correct for that frame. NH has no judgment reward. A parseable answer and, where required, a parseable judgment earn $r_{\text{format}} = 0.1$. Predicting "contradicts" when the target is "confirms" incurs $r_{\text{calibration}} = -0.3$.

For a pair whose two frames occur in the batch, the switch bonus is 0.5 for each correct rollout if at least one rollout on *each* frame answers correctly. It is zero otherwise. This sampled-rollout event differs from evaluation PairAcc, which requires two independently generated greedy answers to be correct. The reward checks labels and response structure; it does not inspect whether the rationale uses visual evidence. RI-VR+nosw sets the switch-bonus weight to zero in the otherwise corresponding reward manager.

Checkpoints and comparison budget. The paired-reward run was probed every 25 updates using 25 training-pool pairs and 50 standard rows. Pair correctness at steps 25, 50, 75, 100, 125, and 150 was 7, 7, 8, 6, 5, and 8 of 25 (20–32%). Step 150 tied the highest paired score and had the highest FC probe accuracy (88%); it was selected for the held-out evaluation. These are pool probes, not held-out test estimates. The nosw result is the author-verified 800-step checkpoint. Its external-machine summary provides the reported scores; raw training logs and periodic evaluation JSONL are not present in the supplied archive, so its full execution overrides, random seed, and selection history cannot be independently reconstructed here. The scripts' defaults do not establish those executed

settings. The 150/800-step difference and distinct RL/SFT sampling costs prevent interpreting the table as a compute-matched comparison or attributing the entire difference to the switch reward.

Held-out results. The archived RI-VR+VCF records give PairAcc $173/552 = 31.3\%$ and CH $497/713 = 69.7\%$, compared with base values of 5.4% and 89.5%. Its full history grid is NH 73.1%, CH 69.7%, WB 74.9%, and FC 61.0%. The nosw summary reports PairAcc 12.6% and CH 66.8%; no unrecorded grid cells are inferred. The paired-reward configuration improves on the observed base and nosw paired scores, but both RL checkpoints lose substantial correct-history accuracy. RI-FT-CF seed 42 provides a supervised reference at CH 98.0% and PairAcc 77.4% after 759 updates. The table reports individual checkpoint results rather than RL multi-seed averages.

Judgment and no-image diagnostics. The two archived detect-style shards contain 684 unique items, each evaluated under FC and CH. Judgments are parseable for 672/684 FC responses (98.2%) and 671/684 CH responses (98.1%); every parsed judgment is "contradicts" in both conditions. These per-condition counts are recomputed from the archived responses. The separate experiment summary reports a four-shard aggregation and 96.8% "contradicts", which do not match these files. The archived responses support a constant-judgment failure despite the contrasting FC and CH targets. The reward and pool composition are possible contributors, but this probe does not isolate their causal effects. This probe also changes the output protocol, so its accuracy should not be substituted for the ordinary QA grid.

Without images, RI-VR+VCF and nosw match the false-belief option on 80.3% and 81.7% of pairs, versus 97.8% for the base. These rates describe answer preferences rather than visual correctness. Conversely, RI-FT-CF's low belief-match rate does not remove its 90.6% later-target preference (Table 10). In the paired-reward training log, mean score rises from 0.686 at step 1 to 0.788 at step 50 and is 0.719 at step 150; bonus coverage is approximately 25.9% in the first 30 steps and 27.0% in the last 30. The supplied nosw summary reports paired scores flattening near 12.6–12.7% late in its run. These trajectories describe the tested budgets and do not establish a general RL performance ceiling.

Earlier process-reward exploration. An earlier RI-VR configuration without the paired setup reaches FC 81.1% on all 713 items (578/710, or 81.4%, among valid outputs), CH 96.5%, and WB 32.8%; answer-only RL reaches FC 15.7%. Its explicit-judgment probe suffers a different failure: none of 359 archived FC probe outputs is parseable. These heterogeneous runs illustrate reward- and format-dependent behavior; they are not additional rows of a controlled comparison with the paired RL checkpoints. Across the exploration, the evidence supports a remaining challenge in jointly learning image sensitivity and correct-history performance under the recorded rewards and optimization settings.

G Multi-Turn Probes and Decision Replay

The main paper tests a necessary component of an embodied system: image-conditioned state judgments under history conflict. Additional experiments explore multi-turn context and decisions, but neither constitutes a closed-loop robot-policy evaluation.

Multi-turn chains can reveal persistence or recovery after textual context is carried forward. A prediction turn that supplies no new image, however, cannot on its own demonstrate failure to use new visual evidence. The archived short-chain and longer-chain probes therefore need to be interpreted by which turns actually introduce observations, rather than aggregating every history-consistent answer as visual inertia.

The fixed-trajectory replay contains 50 episodes with two branches. One branch presents a recorded progression checkpoint, and the other a checkpoint intended to demand retrying an action. Models choose from a supplied decision set without controlling subsequent frames. An audit of 100 first-checkpoint labels finds 57 consistent labels and 38 contradicted labels, with the remainder ambiguous or unresolved. Only 5 of 50 second-checkpoint images are judged clearly complete. These label limitations prevent interpreting raw replay scores as reliable evidence of downstream utility.

Replay identifiers also require care: five truncated episode identifiers originally collided across different instructions. Bookkeeping must distinguish the instruction or use the corrected identifiers

while retaining the original option seeds. This issue belongs to the replay manifest and does not change the history-grid or VCF counts.

The recorded adapters can change progression and retry preferences, and joint state/decision supervision explores a possible mitigation. Because the replay is off-policy, its labels are imperfect, and its intervals do not establish an overall benefit over the base model, we make no claim that RI-FT-CF improves robot task success. A convincing downstream evaluation would require verified decision labels or interactive rollouts with well-defined costs for missed retries and unnecessary retries. That remains separate from the demonstrated paired state-judgment improvement.