GUIDING SELECTIVE SELF-CORRECTION IN VISION-LANGUAGE MODELS THROUGH EVIDENCE AUDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can produce coherent reasoning, but misread visual information may persist through later reasoning and self-checking. Revision can also turn correct answers into wrong ones. We propose PACT, a training-free framework with context-isolated evidence auditing and selective revision. Perception, option-level auditing, and task-adapted consistency testing run without the initial answer or its reasoning. Fusion hides earlier answer identities and combines the image with diagnostic reports. A deterministic gate accepts or rejects its candidate based on consistency and contradiction. All stages share one frozen model, without a separately trained verifier or reference answers at inference time. Across 19 open-source model configurations on MMStar and RealWorldQA, PACT scores higher than six ablations in all 38 model–dataset pairs. Its mean differences from the listed Baseline reference scores are 3.09 and 3.12 percentage points, respectively. On both benchmarks, perception has the highest single-module mean score, while removing auditing causes the largest mean loss. These results show that a module's effect depends on the full configuration.

Keywords: vision-language models; visual reasoning; context isolation; evidence auditing; selective revision.

1 INTRODUCTION

Visual reasoning connects image objects, attributes, and relations to question conditions. Reading chart changes or relative positions requires detailed perception and reasoning across facts, as highlighted by MathVista (Lu et al., 2024). LLaVA-CoT organizes problem solving through staged reasoning and search (Xu et al., 2025), while VReST explores candidate paths through tree search and self-reward (Zhang et al., 2025a). As reasoning grows longer, a key question arises: with model parameters fixed, how can extra computation obtain visual evidence that checks earlier judgments and supports useful revisions?

LLaVA encodes images into visual features and generates answers with the question (Liu et al., 2023). Qwen3-VL further improves visual understanding through multi-level feature fusion and cross-modal position encoding (Bai et al., 2025). Generation still depends on both visual input and prior text, without checking every intermediate claim against the image. A visible detail may be mapped to the wrong value, unit, or reference object. Fluent text and correct calculations therefore do not guarantee correct visual premises.

Figure 1 illustrates two such errors. In panel A, the vertical axis starts at 80, while the actual values are 90 and 100. The incorrect path uses the visible bar heights, 10 and 20, and reports 100% growth. The correct path reads the axis values and computes $(100 - 90)/90 \approx 11.1\%$. The error lies in the inputs, not the growth formula. Panel B asks whether the mug is to the right of the fruit bowl. The answer is Yes. The incorrect path correctly observes that the bowl is to the left of the mug but maps this observation to No. Recognizing a relation is therefore insufficient: its target, reference object, and direction must match the question. VCD also shows that statistical biases and unimodal priors can cause generated content to depart from image evidence (Leng et al., 2024).

Incorrect premises may also affect later checks. TVC finds that long multimodal reasoning can increase reliance on prior text and reduce visual information use (Sun et al., 2025). Reusing a misread value or relation may repeat the error. Same-model feedback can also be wrong, and self-correction

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.

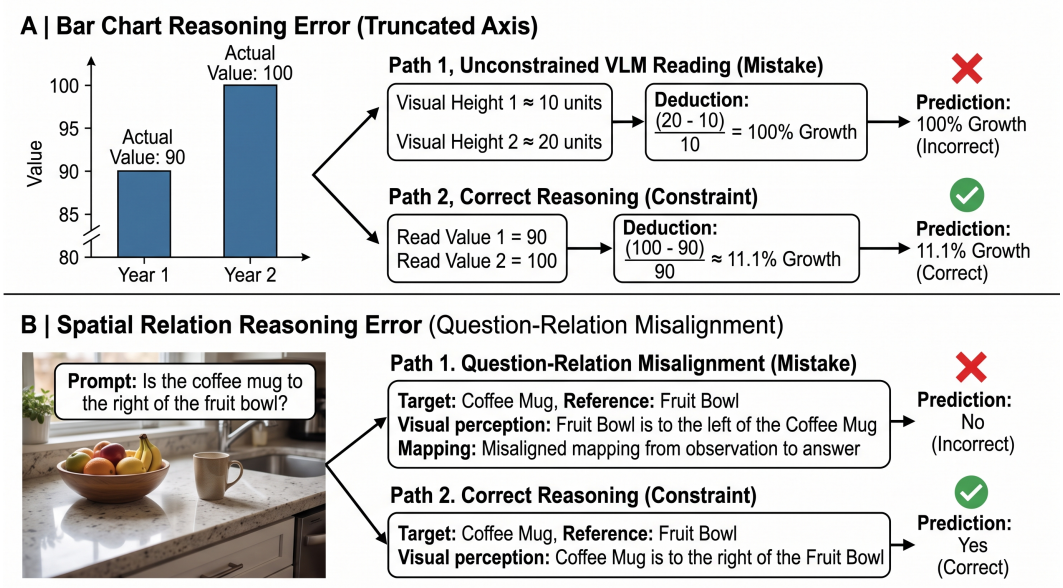

Figure 1: Misalignment between visual premises and question conditions. (A) Truncated bar heights differ from actual values; the correct growth is about 11.1%. (B) The mug is right of the bowl, but a valid relation is mapped to the wrong answer.

without external feedback can reduce reasoning performance (Huang et al., 2024). Reliable revision thus requires checks that avoid directly inheriting the initial judgment, as well as clear conditions for changing the answer.

We propose PACT (Perception, Audit and Counterbalanced Testing). PACT defines which information each diagnostic branch can access. Perception sees only the image and question stem and separates observations from uncertainty. Auditing checks all options without the initial answer or reasoning. When option meanings permit reordering, a permutation test checks agreement after restoring the original option identities. These input boundaries limit direct influence from earlier answers.

PACT also separates proposing a revision from accepting it. Fusion receives the image and reports while earlier answer identities remain hidden. A deterministic gate first protects an initial answer that passes the paired consistency test. It also rejects a revision when only the proposed answer is explicitly contradicted. Context isolation follows the idea of independent verification (Dhuliawala et al., 2024), and permutation testing uses known sensitivity to option order (Zong et al., 2024). We study how these constraints shape evidence and balance correction gains against harmful changes.

Our contributions are threefold:

- We link question-guided perception, auditing without the initial answer, and paired consistency diagnosis through explicit information boundaries in one frozen model.
- We separate evidence fusion from revision acceptance and express accuracy changes as correction gains minus losses from changing correct answers.
- PACT exceeds six ablations across 19 configurations on two benchmarks (38 pairs), with mean differences of 3.09 and 3.12 points from the references. We distinguish single-module performance from effects in larger configurations.

2 Related Work

Visual CoT uses key-region and reasoning-step annotations to guide answers with local visual information (Shao et al., 2024). LLaVA-CoT divides solving into a task summary, visual interpretation, logical reasoning, and a conclusion, with stage-level search (Xu et al., 2025). ViperGPT combines vision and language modules into executable programs for perception, calculation, and reasoning

(Surís et al., 2023). PACT structures information flow across calls to a frozen model, without training new reasoning behavior.

Sampling and search can improve a single generation. Self-Consistency aggregates answers from different reasoning paths (Wang et al., 2023). VReST combines Monte Carlo tree search with multimodal self-reward to explore and assess candidate paths (Zhang et al., 2025a). PACT assigns calls to different diagnostics, then applies fusion and acceptance rules. Search and diagnostics could be combined, but benefits must be separated from repeated solving through controls with matched call counts and generation budgets.

VCD contrasts outputs for original and perturbed images to reduce statistical bias and unimodal priors (Leng et al., 2024). TVC reintroduces visual content and compresses visual tokens to reduce visual forgetting during long reasoning (Sun et al., 2025). Without new training, Woodpecker extracts concepts, forms verification questions, and uses pretrained visual experts to correct hallucinations (Yin et al., 2024). VISCO evaluates fine-grained critique and correction; its LookBack strategy extracts atomic information from prior reasoning and checks it against the image (Wu et al., 2025).

PACT checks the task and options directly, so its diagnostics are not limited to claims in the initial reasoning. The audit separates support, contradiction, and uncertainty, and fusion revisits the image to assess reports. Diagnostic quality still depends on the base model's visual ability.

Self-Refine uses one model to generate drafts, feedback, and revisions (Madaan et al., 2023). CoVe isolates verification contexts to reduce the influence of earlier answers (Dhuliawala et al., 2024), while Critic-V trains a multimodal critic to provide revision feedback (Zhang et al., 2025b). Effects depend on the task and feedback protocol. Huang et al. (2024) show limits of self-correction in some reasoning settings, whereas Liao et al. (2025) demonstrate its potential for specific semantic grounding tasks with binary verification and visual feedback. PACT models feedback generation and revision acceptance separately, since a new judgment need not be a better answer.

Option order provides a controlled diagnostic change when meaning is preserved. Predictions may change with option positions or labels despite unchanged content (Zheng et al., 2024; Zong et al., 2024). PACT performs one nontrivial cyclic permutation on applicable samples and uses agreement after inverse mapping to control revision. This measures stability under the tested change, not correctness. Our focus is the joint design of checking contexts, visual diagnostics, and acceptance rules, rather than treating self-feedback or context isolation as new basic operations.

3 METHOD

PACT first obtains y_0 and runs perception (P), auditing (A), and consistency diagnosis (T) in separate contexts. Fusion then rechecks the image and reports without earlier answer identities, followed by deterministic gating (Figure 2). All calls share one frozen model. The complete procedure is in Algorithm 1, Appendix A.

3.1 TASK AND SHARED INITIAL PREDICTION

For a visual multiple-choice task $x = (I, q, \mathcal{O})$, I is the image, q is the stem without options, and $\mathcal{O} = (o_1, \ldots, o_K)$ is the ordered option set. We predict $\widehat{y} \in \mathcal{Y} = \{1, \ldots, K\}$. The implementation uses $K = 4$ (A–D); notation retains general K. Let $\mathcal{M}_\theta(\cdot; \tau, \xi)$ denote generation with fixed parameters θ, prompt template τ, and seed ξ. The initial prediction is

$$y_0 = \mathcal{M}_\theta(I, q, \mathcal{O}; \tau_S, \xi_0). \tag{1}$$

PACT reuses the Baseline output with the same input, prompt, decoding settings, and seed. Thus, paired comparisons share y_0 and avoid initial sampling differences. The notation denotes decoded outputs; answer probabilities or log probabilities are not required.

3.2 QUESTION-GUIDED PERCEPTION WITHOUT OPTIONS

Option descriptions may influence how a model reads an image. To separate observation from option selection, the perception branch receives only the image and question stem. It returns

$$P = (\mathcal{E}, \mathcal{U}) = \mathcal{M}_\theta(I, q; \tau_P, \xi_P), \tag{2}$$

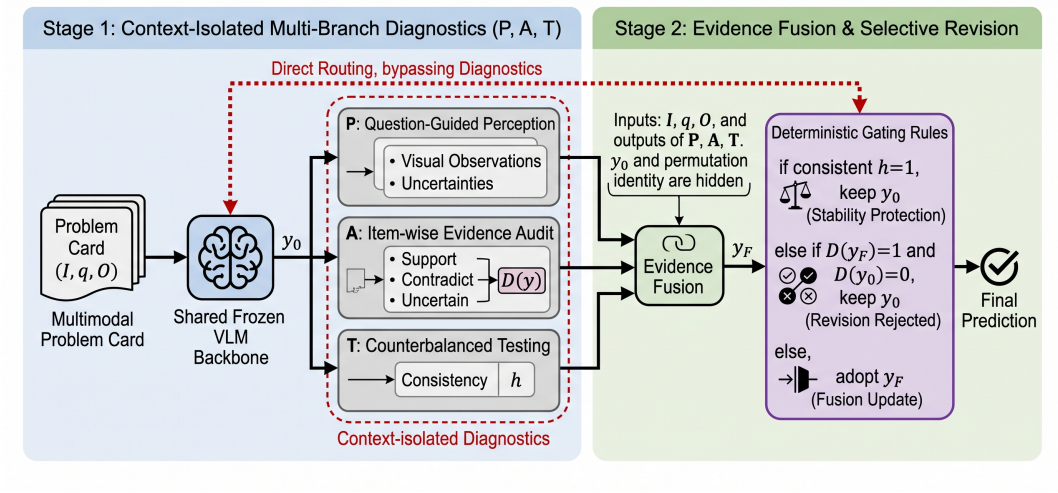

Figure 2: PACT uses separate diagnostic contexts, followed by fusion and gating. Earlier answer identities are hidden from fusion. The gate protects paired stability and rejects revisions when only the proposed answer is contradicted.

where $\mathcal{E} = \{e_j\}_{j=1}^{m}$ contains brief, relevant observations that can be checked against the image. The set $\mathcal{U} = \{u_j\}_{j=1}^{n}$ records attributes or relations that cannot be read reliably. This separates evidence from guesses and makes scales, units, and reference objects explicit. Options cannot directly affect this branch through its input. Perceptual errors remain possible, so fusion also receives the image.

3.3 Evidence auditing without the initial answer

The audit branch checks each option against the image and question. It does not receive y_0, the initial reasoning, or outputs from other diagnostic branches. One call produces all option-level results:

$$A = \{(s_i, a_i)\}_{i=1}^{K} = \mathcal{M}_\theta(I, q, \mathcal{O}; \tau_A, \xi_A), \\ s_i \in \{S, C, U\}, \tag{3}$$

where a_i explains the evidence. States S, C, and U mean supported, contradicted, and uncertain: evidence favors the option, opposes it, or supports neither judgment. Define

$$D(i) = \mathbf{1}[s_i = C]. \tag{4}$$

The three states distinguish missing support from contradiction. $D(i) = 0$ does not verify option i. Uncertainty alone does not reject a revision, whereas contradiction can. Fusion must still check the model-generated audit against the image.

3.4 Consistency under option permutation

When option meanings do not depend on labels or positions, reordering the options should change only the answer label. Sensitivity to option order therefore provides a diagnostic signal (Zheng et al., 2024; Zong et al., 2024). Using a fixed sample seed, PACT selects a nonzero cyclic shift $r \in \{1, \ldots, K-1\}$. The mapping from original to new positions is

$$\pi_r(i) = 1 + ((i - 1 + r) \bmod K), \qquad \mathcal{O}^{\pi_r}_{\pi_r(i)} = o_i. \tag{5}$$

The permutation branch solves the reordered question and maps the prediction back to the original option index:

$$\widetilde{y}_T = \mathcal{M}_\theta(I, q, \mathcal{O}^{\pi_r}; \tau_S, \xi_T), \qquad y_T = \pi_r^{-1}(\widetilde{y}_T). \tag{6}$$

An ideal solver would satisfy label equivariance:

$$f^\star(I, q, \mathcal{O}^{\pi_r}) = \pi_r\big(f^\star(I, q, \mathcal{O})\big). \tag{7}$$

That is, the label follows the new position while the selected content stays the same. For successfully completed solves, let v_T indicate whether the test is enabled and applicable. Define

$$c_T = \begin{cases} \text{consistent}, & v_T = 1 \ \wedge \ y_T = y_0, \\ \text{inconsistent}, & v_T = 1 \ \wedge \ y_T \neq y_0, \\ \text{unavailable}, & v_T = 0. \end{cases} \tag{8}$$

Heuristic rules conservatively skip options with label or position references, such as "all of the above," "option A," or "A and B," and record the reason. Disabled or skipped tests are unavailable, not inconsistent. One permutation measures only local paired stability. Under stochastic decoding, disagreement can reflect both order and sampling changes, so it cannot isolate position bias.

3.5 Evidence fusion without earlier answers

Fusion combines the original input, the perception and audit reports, and the consistency state:

$$y_F = \mathcal{M}_\theta(I, q, \mathcal{O}, P, A, c_T; \tau_F, \xi_F). \tag{9}$$

Fusion does not receive y_0, $\widetilde{y}_T$, y_T, or π_r. It knows whether earlier solves agree but cannot copy their answers directly. It rechecks the image and reports before selecting a candidate; audit labels are not votes.

Input boundaries do not imply statistical independence. The shared model may repeat perceptual errors, and fusion reads the audit report. Agreement is therefore not independent confirmation. The acceptance rule controls the effect of these fallible reports.

3.6 Selective revision with stability priority

Revision can correct a wrong answer or turn a correct answer into a wrong one (Huang et al., 2024). PACT separates proposing a new answer from accepting it. Let

$$h = \mathbf{1}[c_T = \text{consistent}]. \tag{10}$$

The final prediction is

$$\widehat{y} = \begin{cases} y_0, & h = 1, \\ y_0, & h = 0 \ \wedge \ D(y_F) = 1 \ \wedge \ D(y_0) = 0, \\ y_F, & \text{otherwise}. \end{cases} \tag{11}$$

The first rule keeps an initial answer that is consistent across the tested orders, even if fusion disagrees. Audit contradictions are recorded as conflicts. The second rule rejects a revision when only the fusion candidate is contradicted. Other cases use y_F.

Uncertainty alone does not prevent revision. If both answers are contradicted, the rule uses y_F and records a conflict, including when fusion repeats the contradicted initial answer. Stability protection can prevent harmful changes but also preserve stable errors. Both effects require evaluation. Fusion always runs before gating, so protection does not provide early stopping.

3.7 Revision behavior and accuracy change

Equation (11) changes the initial answer if and only if

$$\widehat{y} \neq y_0 \iff (y_F \neq y_0) \wedge (h = 0) \wedge \big(D(y_F) = 0 \vee D(y_0) = 1\big). \tag{12}$$

Thus, fusion must propose a different answer, consistency protection must be absent, and contradiction must not apply only to the fusion candidate. The candidate need not be marked supported. The model may still revise its answer based on the image when the audit is uncertain.

Let $R = \{\widehat{y} \neq y_0\}$ be the revision event, and let $y^\star$ be the ground-truth label, used only for evaluation. Under the same sample distribution and the same criterion for successful predictions, the accuracy change is

$$\begin{aligned} \Delta\mathrm{Acc} &= \Pr(\widehat{y} = y^\star) - \Pr(y_0 = y^\star) \\ &= \underbrace{\Pr(R, \ y_0 \neq y^\star, \ y_F = y^\star)}_{\text{gain from correcting initial errors}} \\ &\quad - \underbrace{\Pr(R, \ y_0 = y^\star, \ y_F \neq y^\star)}_{\text{loss from changing correct answers}}. \end{aligned} \tag{13}$$

Unchanged answers retain their initial correctness. After revision, the output is y_F, so only wrong-to-correct and correct-to-wrong changes affect accuracy. Equation (13) separates these gains and losses. Samples with $h = 1$ produce neither corrections nor harmful changes. Net improvement depends on the difference between the two event rates.

3.8 Inference budget and module control

Without retries, full PACT uses five calls when permutation applies: one initial solve, one call to each of P, A, and T, and one fusion call. Let $b_P, b_A, b_T \in \{0, 1\}$ indicate enabled modules and $e(x) \in \{0, 1\}$ indicate permutation applicability. Then

$$N_{\text{calls}}(x) = 2 + b_P + b_A + b_T e(x). \tag{14}$$

The fixed calls are initial solving and fusion. Disabling P or A leaves its report empty; disabling A sets $D(i) = 0$. Disabling or skipping T sets $c_T =$ unavailable. All variants retain fusion and the gate. Core, with no diagnostics, still uses two calls and can separate fusion from diagnostic effects. A required stage that fails after bounded recovery produces an error record, not an empty report.

4 Experiments

We evaluate full PACT, module removal, and single-module settings across 19 frozen open-source configurations on MMStar and RealWorldQA (Table 1). Public Baseline references appear in Table 2, Appendix B. Differences from these references are distinct from paired gains on the same questions.

4.1 Datasets

MMStar (Chen et al., 2024) contains 1,500 selected visual questions across six core abilities: coarse perception, fine-grained perception, instance reasoning, logical reasoning, science and technology, and mathematics. These are divided into 18 subcategories. Sample selection and human review reduce questions that can be answered from text alone. We use MMStar to assess general visual reasoning, including the extraction and use of question-relevant visual information.

RealWorldQA (xAI, 2024) contains 765 questions about real scenes. Its original formats include multiple-choice questions and short open-ended answers. Questions cover object attributes, spatial relations, relative positions, and scene states. Models must locate relevant objects and combine local details with the wider scene. We use this benchmark to assess real-world scene understanding and convert all questions into multiple-choice format for our experiments.

The datasets provide complementary views of general visual reasoning and real-world problem solving. Nominal dataset sizes are distinguished from the actual evaluation subsets, which are defined by the sample list for each run. The converted RealWorldQA setting must also be distinguished from its original answer formats when interpreting public reference scores.

4.2 Base models and Baseline

Open-source models. We evaluate 19 configurations from three families. Qwen3-VL-Thinking includes 2B, 4B, 8B, 30B-A3B, and 32B. Qwen3.5 includes 0.8B, 2B, 4B, 9B, 27B, and 35B-A3B. InternVL3.5 includes 1B, 2B, 4B, 8B, 14B, 20B-A4B, 30B-A3B, and 38B. These comparisons examine whether the process depends on a particular base model.

Closed-source references. Table 2 includes Claude, GPT, and Gemini, with the supplied version names and high settings for GPT-5-Mini and GPT-5. Scores may differ in publisher, model snapshot, and evaluation protocol. They provide context, are excluded from averages, and do not establish rankings under a shared protocol.

Initial solving and reference scores. In the method, Baseline means one solve of the original image and question, retaining any reasoning produced by a Thinking model. The open-source reference scores in Table 2 come from public reports; the Qwen3.5 entries use Thinking settings. These public scores are distinct from the shared initial predictions used for paired analysis in Section 3.

4.3 EXPERIMENTAL SETTINGS AND EVALUATION

Parameters stay fixed, and reference labels are used only for scoring after prediction. Between the initial solve and fusion, PACT records relevant observations, audits candidate evidence, and checks consistency. T reorders options and restores original option identities. Branch execution and skips must match the run records. For RealWorldQA, this applies to the converted multiple-choice setting.

The six ablations either remove one branch (w/o P, w/o A, w/o T) or retain one (P only, A only, T only). All keep the common solve-and-fusion process and use the same base model and evaluation scope. Single-module differences therefore include the common revision process.

Scores are reported as percentages. MMStar uses option-label accuracy. For the converted RealWorldQA setting, the option construction, answer mapping, normalization, equivalence rules, and aggregation protocol need to be specified alongside the run records. For model m and dataset d, let the score be $S_{m,d}$. Define

$$\Delta_{m,d} = S_{m,d}^{\mathrm{PACT}} - S_{m,d}^{\mathrm{Base}}, \qquad \Delta_{d}^{\mathrm{mean}} = \frac{1}{M}\sum_{m=1}^{M} \Delta_{m,d}, \tag{15}$$

where $M = 19$. We calculate differences and equally weighted model means before rounding. These are not pooled sample accuracies; related configurations are not independent repeats. Base is a public reference, so Eq. (15) differs from the paired change in Eq. (13). Without question-level paired tests or variance across runs, analysis is limited to reported magnitudes, directions, and rankings.

4.4 MAIN RESULTS

Full PACT has the highest listed score in every model configuration. Across all 38 model–dataset pairs in Table 1, it exceeds all six ablations and the corresponding public Baseline reference. On MMStar, the reference and PACT means are 71.55 and 74.64, a difference of 3.09 percentage points. On RealWorldQA, they are 72.67 and 75.79, a difference of 3.12 points. Paired improvements still require initial and final predictions on shared samples.

On MMStar, positive differences from the references occur in all three model families. Individual differences range from 2.50 to 3.90 points. Qwen3.5-0.8B has the largest difference, from 58.3 to 62.2. Qwen3.5-2B and InternVL3.5-20B-A4B differ by 3.6 and 3.5 points, respectively. Qwen3.5-35B-A3B reaches the highest PACT score, 84.9, compared with its reference of 81.9. Mean differences are 2.82, 3.18, and 3.19 points for Qwen3-VL-Thinking, Qwen3.5, and InternVL3.5, respectively.

RealWorldQA shows the same positive direction, with individual differences of 2.70–3.70 points. Qwen3.5-0.8B and InternVL3.5-1B differ from their references by 3.5 and 3.7 points. InternVL3.5-8B scores 70.6 with PACT versus 67.5 in the reference, a 3.1-point difference. Qwen3.5-35B-A3B again has the highest PACT score, 87.0, compared with 84.1. Mean differences for the three families are 3.10, 2.98, and 3.23 points.

GPT-5.2 has an MMStar reference of 77.1, while Qwen3-VL-8B-Thinking with PACT scores 78.1. On RealWorldQA, GPT-5.2 and Gemini-3-Pro both have references of 83.3; Qwen3.5-27B with PACT scores 86.5. These are score comparisons across protocols. Any lead already present in the base model cannot be attributed to PACT.

4.5 ABLATION STUDY

Removal losses compare entries within Table 1. Single-module differences use public references and include the shared fusion process.

MMStar. P only has the highest single-module mean. The means for w/o P, w/o A, and w/o T are 71.59, 71.54, and 71.56, respectively, which are 3.05, 3.09, and 3.07 points below full PACT. P only, A only, and T only score 72.17, 72.12, and 72.04, with differences of 0.63, 0.57, and 0.49 points from the references. Thus, P has the highest mean when used alone, but removing A produces a slightly larger mean loss. Module rankings also vary across models.

This fits MMStar's emphasis on visual information: explicit observations may help with attributes, relations, and multi-step reasoning. Removing P also changes fusion inputs, so the difference includes downstream effects. Total scores cannot identify which perception errors were corrected.

Table 1: PACT and ablation scores (%). Base repeats the public reference score for context. It is not a paired rerun of the initial solve. Both panels contain all 19 open-source model configurations.

(a) MMStar

Model	Baseline	w/o P	w/o A	w/o T	P only	A only	T only	PACT
Q3-VL-2B	68.1	68.2	68.2	68.0	68.8	68.9	68.7	**71.2**
Q3-VL-4B	73.2	73.3	73.4	73.0	73.8	73.8	73.8	**75.8**
Q3-VL-8B	75.3	75.6	75.2	75.3	75.9	75.7	75.9	**78.1**
Q3-VL-30B-A3B	75.5	75.5	75.6	75.4	76.3	76.3	76.0	**78.5**
Q3-VL-32B	79.4	79.3	79.5	79.4	79.9	80.1	80.1	**82.0**
Q3.5-0.8B	58.3	58.3	58.5	58.4	59.2	58.8	58.6	**62.2**
Q3.5-2B	71.7	71.8	71.8	71.7	72.2	72.1	72.1	**75.3**
Q3.5-4B	78.3	78.3	78.2	78.3	79.0	79.0	78.9	**81.2**
Q3.5-9B	79.7	80.0	79.7	79.9	80.6	80.4	80.4	**82.2**
Q3.5-27B	81.0	81.2	81.1	81.1	81.8	81.7	81.6	**84.2**
Q3.5-35B-A3B	81.9	82.1	81.7	81.8	82.6	82.6	82.2	**84.9**
IVL3.5-1B	51.9	51.8	51.9	52.0	52.6	52.3	52.2	**55.3**
IVL3.5-2B	62.7	62.7	62.7	62.7	63.1	63.3	63.4	**65.7**
IVL3.5-4B	65.0	65.1	64.9	65.0	65.4	65.6	65.5	**68.0**
IVL3.5-8B	69.3	69.2	69.2	69.2	69.8	70.1	69.9	**72.7**
IVL3.5-14B	70.4	70.2	70.3	70.6	71.0	70.7	70.7	**73.4**
IVL3.5-20B-A4B	70.4	70.5	70.3	70.6	71.2	70.7	70.8	**73.9**
IVL3.5-30B-A3B	72.0	72.0	72.0	72.0	72.5	72.3	72.3	**75.2**
IVL3.5-38B	75.3	75.1	75.1	75.3	75.6	75.9	75.6	**78.3**

(b) RealWorldQA

Model	Baseline	w/o P	w/o A	w/o T	P only	A only	T only	PACT
Q3-VL-2B	69.5	69.6	69.6	69.4	70.4	70.2	69.8	**72.8**
Q3-VL-4B	73.2	73.3	73.2	73.2	74.0	73.8	73.8	**76.4**
Q3-VL-8B	73.5	73.5	73.3	73.7	74.1	73.7	73.8	**76.2**
Q3-VL-30B-A3B	77.4	77.3	77.3	77.2	77.7	78.0	77.9	**80.5**
Q3-VL-32B	78.4	78.4	78.6	78.5	79.2	79.1	79.0	**81.6**
Q3.5-0.8B	63.4	63.6	63.3	63.6	64.0	63.8	63.9	**66.9**
Q3.5-2B	74.5	74.7	74.5	74.5	75.2	75.0	75.1	**77.3**
Q3.5-4B	79.5	79.4	79.5	79.3	79.9	79.9	80.2	**82.5**
Q3.5-9B	80.3	80.3	80.3	80.3	81.0	80.9	80.9	**83.2**
Q3.5-27B	83.7	83.8	83.7	83.9	84.2	84.5	83.9	**86.5**
Q3.5-35B-A3B	84.1	84.1	83.9	84.2	84.9	84.6	84.5	**87.0**
IVL3.5-1B	57.6	57.7	57.6	57.6	58.3	58.1	58.0	**61.3**
IVL3.5-2B	62.0	62.2	62.0	62.2	62.8	62.2	62.6	**65.3**
IVL3.5-4B	66.3	66.4	66.2	66.3	67.1	67.0	66.7	**69.5**
IVL3.5-8B	67.5	67.6	67.5	67.6	67.9	68.2	68.0	**70.6**
IVL3.5-14B	70.5	70.6	70.4	70.4	71.0	71.2	71.0	**73.9**
IVL3.5-20B-A4B	71.2	71.0	71.0	71.3	71.9	72.0	71.7	**74.5**
IVL3.5-30B-A3B	72.3	72.3	72.2	72.1	72.9	73.1	72.7	**75.0**
IVL3.5-38B	75.9	75.8	75.9	76.0	76.5	76.1	76.5	**79.0**

Q3-VL: Qwen3-VL-Thinking; Q3.5: Qwen3.5; IVL3.5: InternVL3.5. "w/o X" removes X; "X only" retains only X. All ablations keep the common solve-and-fusion process. Bold marks the highest score for each model and dataset. Means give equal weight to all 19 configurations. For RealWorldQA, T follows the multiple-choice conversion described in the text; comparisons with public references may use different protocols.

RealWorldQA. Single-module and removal-loss rankings also differ. The means for w/o P, w/o A, and w/o T are 72.72, 72.63, and 72.70, with losses of 3.07, 3.16, and 3.09 points from full PACT. P only, A only, and T only score 73.32, 73.23, and 73.16, differing from the references by 0.64, 0.56, and 0.48 points. As on MMStar, P has the highest single-module mean, while removing A gives the largest numerical loss. These close losses, without variance estimates or significance tests, do not establish a larger contribution from one module.

Removing T lowers the RealWorldQA mean from 75.79 to 72.70, while T only scores 73.16. Its effect therefore needs to be assessed together with other diagnostics and fusion. Results should also be separated by original question format, including originally multiple-choice questions and converted open-ended questions. The overall difference cannot be attributed entirely to reduced option-position bias.

Joint configuration. PACT exceeds each model's best ablation by averages of 2.39 points on MMStar and 2.38 on RealWorldQA. All single-module settings exceed the references, while some removal settings match or fall below them. Joint evaluation is useful, but these results do not establish independent errors or causal complementarity.

Dependence on other modules. On MMStar, adding A to P only (w/o T versus P only) lowers the mean by 0.61 points. Adding A to P+T (PACT versus w/o A) raises it by 3.09 points. RealWorldQA changes are -0.62 and $+3.16$ points. Module effects thus depend on the task and other branches. The full configuration's highest score does not imply that every combination exceeds the sum of its separate gains.

4.6 Further analysis and discussion

Model size does not predict differences from the references. On MMStar, InternVL3.5-1B and Qwen3.5-0.8B differ by 3.4 and 3.9 points, exceeding several larger models. On RealWorldQA, Qwen3-VL-32B-Thinking and InternVL3.5-14B have larger differences than their 8B counterparts. Smaller models therefore do not always benefit more. Parameter counts across architectures also do not directly measure inference cost.

With PACT, InternVL3.5-2B scores 65.7 on MMStar, above the 4B reference of 65.0. InternVL3.5-4B scores 69.5 on RealWorldQA, above the 8B reference of 67.5. Without unified protocols and budgets, these comparisons do not establish greater accuracy or efficiency. Shared-sample evaluation and measured computation are needed.

5 Conclusion

PACT combines question-guided perception, evidence auditing, and task-adapted consistency diagnosis with deterministic revision acceptance, without training. Across 19 open-source configurations, it exceeds six ablations in all 38 model–dataset pairs. Mean scores are 74.64 and 75.79, differing from the listed references by 3.09 and 3.12 points. Module effects depend on the configuration; adding modules does not always help. These findings support jointly designing evidence extraction, checking, and revision acceptance. Future work will use shared-sample paired evaluation, question-level correction analysis, and matched-budget controls to test practical gains and scope.

AI USE STATEMENT

Generative AI tools assisted with English translation, language editing, and LaTeX formatting of this manuscript.

REFERENCES

S. Bai, Y. Cai, R. Chen, et al. Qwen3-VL Technical Report. arXiv:2511.21631, 2025.

L. Chen, J. Li, X. Dong, et al. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024. doi: 10.52202/079017-0850.

S. Dhuliawala, M. Komeili, J. Xu, et al. Chain-of-Verification Reduces Hallucination in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, 2024. doi: 10.18653/v1/2024.findings-acl.212.

J. Huang, X. Chen, S. Mishra, et al. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations (ICLR)*, 2024.

S. Leng, H. Zhang, G. Chen, et al. 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)*, pp. 13872–13882, 2024.

Y.-H. Liao, R. Mahmood, S. Fidler, and D. Acuna. Can Large Vision-Language Models Correct Semantic Grounding Errors By Themselves? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14667–14678, 2025.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516.

P. Lu, H. Bansal, T. Xia, et al. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

A. Madaan, N. Tandon, P. Gupta, et al. 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.

H. Shao, S. Qian, H. Xiao, et al. Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024. doi: 10.52202/079017-0275.

H.-L. Sun, Z. Sun, H. Peng, and H.-J. Ye. Mitigating Visual Forgetting via Take-along Visual Conditioning for Multi-modal Long CoT Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5158–5171, 2025. doi: 10.18653/v1/2025.acl-long.257.

D. Surís, S. Menon, and C. Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11888–11898, 2023.

X. Wang, J. Wei, D. Schuurmans, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023.

X. Wu, Y. Ding, B. Li, et al. VISCO: Benchmarking Fine-Grained Critique and Correction Towards Self-Improvement in Visual Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9527–9537, 2025.

xAI. RealWorldQA. Dataset, 2024. URL `https://huggingface.co/datasets/xai-org/RealworldQA`.

G. Xu, P. Jin, Z. Wu, et al. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098, 2025.

S. Yin, C. Fu, S. Zhao, et al. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *Science China Information Sciences*, 67:220105, 2024. doi: 10.1007/s11432-024-4251-x.

C. Zhang, J. Peng, Z. Wang, et al. VReST: Enhancing Reasoning in Large Vision-Language Models through Tree Search and Self-Reward Mechanism. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3922–3941, 2025a. doi: 10.18653/v1/2025.acl-long.199.

D. Zhang, J. Lei, J. Li, et al. Critic-V: VLM Critics Help Catch VLM Errors in Multimodal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9050–9061, 2025b.

C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large Language Models Are Not Robust Multiple Choice Selectors. In *International Conference on Learning Representations (ICLR)*, 2024.

Y. Zong, T. Yu, R. Chavhan, B. Zhao, and T. Hospedales. Fool Your (Vision and) Language Model with Embarrassingly Simple Permutations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62892–62913, 2024.