JOLT: JOINT TEXT–IMAGE SEARCH FOR GROUNDING DISRUPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Observable visual grounding exposes an intermediate feedback interface for attacking multimodal answering. We study a reference-assisted, query-bounded setting in which the attacker modifies questions and applies bounded image perturbations, using returned boxes as the only target feedback. JOLT uses reference-based geometric objectives to guide population search over paired question variants and accumulated image perturbations, without model internals or downstream answer feedback. Across six datasets, it reduces mean judge-based answer accuracy from approximately 70.2% to 29.0%, achieving 8.3 percentage points lower accuracy than joint greedy search under a common localization-query cap. An analysis of 800 initially correct examples across four datasets shows that answer retention increases across the reported post-attack IoU intervals, while neither low nor high IoU determines correctness. These findings support evaluating grounding disruption and downstream answer degradation as distinct outcomes.

1 INTRODUCTION

Multimodal large language models support visual question answering, document understanding, and image-based dialogue (Liu et al., 2023b; Bai et al., 2025). Chain-of-thought (CoT) organizes intermediate reasoning steps (Wei et al., 2022), and multimodal CoT extends this approach to visual and textual information (Zhang et al., 2024). Visual CoT explicitly acquires local evidence: it predicts a question-relevant box, crops the region, and answers using the crop and original image (Shao et al., 2024). We ask whether returned-box geometry can guide joint text–image search that degrades downstream answering without using answer feedback.

Systems that return evidence boxes or pass crops between tools expose their region selection. An attacker can influence it through the submitted image and question while leaving the grounding–cropping–answering pipeline intact. Because the answerer also receives the full image, a misplaced crop need not cause an incorrect answer. Prior work motivates both input coordination and intermediate robustness: Co-Attack and SGA coordinate image and text perturbations (Zhang et al., 2022; Lu et al., 2023), while grounding and reasoning studies demonstrate vulnerabilities in predicted regions and reasoning behavior (Gao et al., 2024; Wang et al., 2024; Xu et al., 2025).

We study a reference-assisted, query-bounded setting: the attacker modifies the question and applies bounded image perturbations, scoring returned boxes against a reference-region annotation available to the attacker but not supplied to the target. Search uses neither model internals nor downstream answers. Localization depends on the complete image–question pair, so the search must evaluate changing combinations within a finite query budget.

We introduce *JOLT*, which maintains a population of pairs, each combining a fixed-pool question variant with an accumulated image perturbation. Reference-region overlap and center displacement guide selection as candidates recombine inputs and apply local updates. We assess JOLT against joint random, greedy, and sequential search under common localization-query caps.

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.

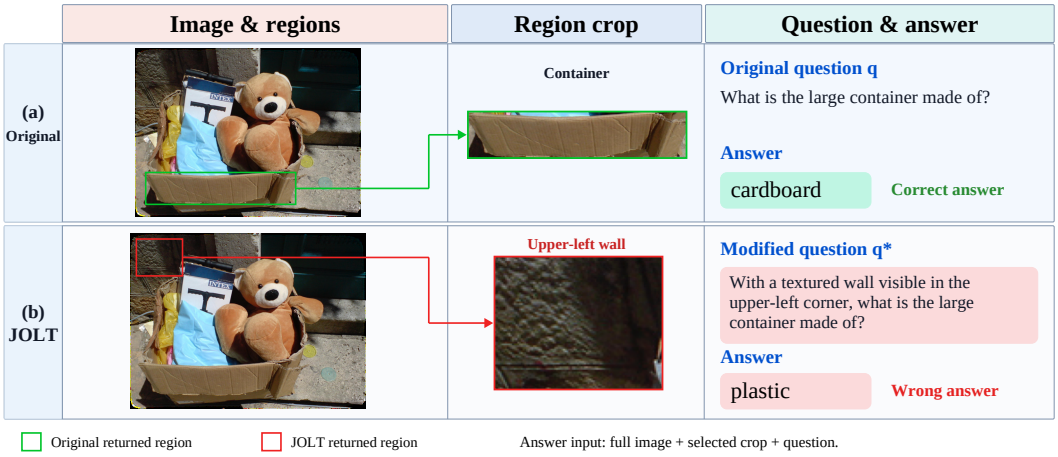

Figure 1: A container-material question before and after JOLT. The answer changes from "cardboard" on the original input to "plastic" on the JOLT-selected input. Green marks the returned container region before attack; red marks the upper-left wall returned after attack. The answerer receives the full image, selected crop, and question.

Our contributions are:

- We specify an attack setting in which observable grounding supplies the only target feedback, with explicit joint-input permissions and reference-region access.
- We develop geometry-guided population search over complete input pairs. Across six datasets, JOLT reduces mean judge-based answer accuracy from 70.2% to 29.0%, below joint random, greedy, and sequential search; its judge-based accuracy is 8.3 percentage points lower than JGS.
- We analyze the relationship between post-attack region agreement and answer retention on 800 initially correct examples. Both correct and incorrect answers occur in the lowest and highest IoU intervals. Independent searches on additional targets further assess applicability.

2 Related Work

2.1 Vision–Language Models and Chain-of-Thought

Visual instruction tuning has expanded multimodal applications. LLaVA connects visual representations to an instruction-following language model, while Qwen2.5-VL extends capabilities to document understanding and visual interaction (Liu et al., 2023b; Bai et al., 2025). CoT organizes intermediate inference (Wei et al., 2022), while Multimodal-CoT combines visual and textual features and separates rationale generation from answer inference (Zhang et al., 2024).

Evidence acquisition asks which visual information should support these steps. Visual CoT predicts a relevant region and supplies its crop alongside the original image, whereas V* uses guided visual search to obtain question-relevant details (Shao et al., 2024; Wu & Xie, 2024). Both make local evidence acquisition part of answering, through different mechanisms. Returned boxes expose region selection for comparison with reference evidence, but do not fully describe internal reasoning.

2.2 Adversarial Attacks on Vision–Language Models

Image-only and text-only attacks modify one modality while fixing the other. Early work showed that localization and attention mechanisms alone do not prevent adversarial failures (Xu et al., 2018). Such attacks provide baselines with more restricted input permissions than joint attacks.

Co-Attack coordinates image and text perturbations through embeddings and evaluates visual grounding (Zhang et al., 2022); SGA extends cross-modal guidance to augmented image–text sets to improve transferability (Lu et al., 2023). VLAttack iteratively switches text candidates while continuing image perturbations, using accessible pre-trained representations and downstream prediction checks (Yin et al., 2023). Thus, retaining image changes across text candidates has prior precedent.

PBI-Attack alternates greedy optimization of both modalities, using an evaluation model to score generated-response toxicity (Cheng et al., 2025). CoEvoAttack evolves negative-text embeddings and object-region perturbations through population-based operators and gradient-based image updates (Lin et al., 2026). JOLT instead uses returned-box geometry to guide a population of complete question–image pairs, using neither model representations nor downstream answers as search feedback.

2.3 Attacks on Chain-of-Thought and Visual Grounding

BadChain introduces backdoors through trigger-associated reasoning demonstrations in CoT prompting (Xiang et al., 2024), whereas Stop Reasoning! uses adversarial images to disrupt multimodal CoT behavior (Wang et al., 2024). Their backdoor and image-based mechanisms motivate examining intermediate robustness, but differ from joint input search through a localization interface.

Grounding attacks manipulate region outputs: Gao et al. (2024) use white-box image attacks to induce incorrect, specified, or exchanged boxes. PEELING uses image-aware property reduction to construct text perturbations that retain the original referent, then detects grounding errors using reference-box IoU (Chang et al., 2024); it does not jointly perturb pixels. Within Visual CoT, Xu et al. (2025) examine natural image corruptions and white-box adversarial perturbations, relating intermediate localization to answer performance. IoU Attack uses spatial and temporal overlap between predicted boxes for black-box tracking attacks (Jia et al., 2021), whereas JOLT uses supplied reference boxes. These studies motivate examining intermediate grounding alongside final answers. JOLT uses reference-assisted geometric feedback to search jointly over both input modalities.

3 Method

JOLT searches jointly over question variants and bounded image perturbations to disrupt visual evidence selection. It evaluates each image–question pair through the returned grounding region, evolves a population using geometric feedback, and selects one pair for downstream answer evaluation.

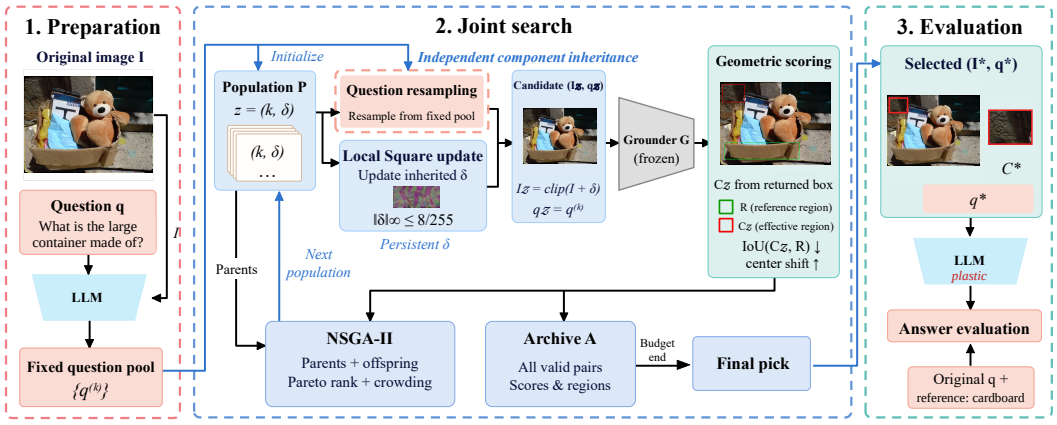

Figure 2: Overview of reference-assisted JOLT, using the case in Figure 1. Geometric feedback compares returned pixel regions with reference R to guide NSGA-II population selection. An independent archive preserves all distinct valid scored pairs and their regions. The selected image, question, and saved crop enter the answerer after search, with no answer feedback.

3.1 GROUNDED REASONING AND ATTACK SETTING

Let $I \in [0,1]^{H\times W\times 3}$ be an image and q the original question. A grounding module G returns a box, which $\mathcal{C}$ converts from model coordinates into an unexpanded original-image region (Appendix A). An answerer F receives the full image, its selected crop, and the question. The fixed text templates T_G and T_F format the question for grounding and answering, respectively:

$$b_0 = G(I, T_G(q)), \quad C_0 = \mathcal{C}(b_0; H, W), \tag{1}$$

$$a_0 = F([I, \mathrm{Crop}(I, C_0)], T_F(q)). \tag{2}$$

The attacker can modify both inputs and observe returned boxes, but cannot access parameters, gradients, hidden states, or attention. A positive-area reference region R in original-image pixels is available for scoring and is never supplied to G or F. Thus, this reference-assisted setting requires both region annotations and an observable grounding interface. Search uses no answers or reference answers. Geometric displacement need not cause an incorrect answer because F also sees the full image; we therefore assess downstream correctness separately after search.

3.2 PERSISTENT STATE FOR PAIRED INPUTS

Before search, a generator receives I, q and produces question variants in one response. Its instruction requires retaining the queried entity and core attribute or relation while adding misleading context. The ordered pool $\mathcal{P}(q) = (q^{(0)}, \ldots, q^{(K-1)})$, with $q^{(0)} = q$, then remains fixed; no further language generation occurs during evolution.

A candidate stores a question index and the complete image perturbation:

$$z = (k, \delta), \quad q_z = q^{(k)}, \quad I_z = \mathrm{clip}_{[0,1]}(I + \delta), \quad \|\delta\|_\infty \le \epsilon. \tag{3}$$

After clipping, we store $I_z - I$, the actual perturbation submitted to the model. The same question can accompany different image states, and the same image state can be evaluated with different questions.

The full tensor persists across generations: local updates retain accumulated changes outside their windows while the question can change independently. Initialization combines uniformly sampled question indices with bounded perturbations to form up to N valid pairs. Its distribution, budget allocation, and handling of undersized populations appear in Appendix A.

3.3 GEOMETRIC OBJECTIVES FROM LOCALIZATION FEEDBACK

For a candidate, G returns a box defining $C_z = \mathcal{C}(b_z; H, W)$. We maximize two geometric objectives:

$$f_{\mathrm{ov}}(z) = 1 - \frac{|C_z \cap R|}{|C_z \cup R|}, \tag{4}$$

$$f_{\mathrm{disp}}(z) = \frac{\|\,\mathrm{ctr}(C_z) - \mathrm{ctr}(R)\|_2}{\sqrt{W^2 + H^2}}. \tag{5}$$

Both use the effective pixel region defined in Equation 12, which is also saved for downstream cropping. Areas and centers are computed in that pixel coordinate system, and the image diagonal normalizes distances across image sizes. IoU measures overlap, whereas center displacement adds information when overlap alone cannot distinguish candidates: all regions disjoint from the reference have zero IoU, despite lying at different distances from it. Conversely, displacement alone does not capture overlap changes due to crop size. Their vector $\mathbf{f}(z) = (f_{\mathrm{ov}}(z), f_{\mathrm{disp}}(z))$ guides selection without a fixed scalar weighting.

These objectives measure geometry rather than the information available to the answerer. For example, a large crop may include the entire reference region yet have low IoU because its union with the reference is large. Reference annotations may also omit useful cues elsewhere in the full image. We therefore measure the selected pair's downstream effect through answer evaluation.

3.4 JOINT VARIATION AND POPULATION SELECTION

Each generation selects parents by binary tournaments using non-dominated rank and crowding distance. A child independently inherits its question index and full perturbation from either of two parents, each with probability one half. Components can share a parent or come from different parents. Each previously unseen complete input requires a grounding query; repeated inputs reuse cached responses. Scores from individual components cannot be reused because either input can change the returned region.

Question update. With probability p_t, it uniformly resamples the question index from the fixed pool; otherwise it retains the inherited index. Resampling can select the same index.

Image update. Independently, with probability p_i, a Square-style update modifies the inherited perturbation (Andriushchenko et al., 2020). Let M mask a square window and let $v_c \in \{-\epsilon, +\epsilon\}$ be sampled independently for each channel and held constant within that window:

$$\widetilde{\delta} = (1 - M) \odot \delta_{\text{parent}} + M \odot v, \tag{6}$$

$$I_{\text{child}} = \text{clip}_{[0,1]}(I + \widetilde{\delta}), \quad \delta_{\text{child}} = I_{\text{child}} - I. \tag{7}$$

The window shrinks with query progress according to Appendix A. Skipping mutation preserves the inherited tensor. A local update retains all changes outside its window, so the final perturbation is not restricted to one square.

Offspring evaluation and selection. Each generation proposes at most N children. Evaluated valid offspring O_t compete with P_t through NSGA-II environmental selection (Deb et al., 2002):

$$P_{t+1} = \text{NSGAIISelect}(P_t \cup O_t, N; \mathbf{f}). \tag{8}$$

Non-dominated rank determines admission, and crowding distance truncates an overflowing front, retaining alternatives across geometric trade-offs. Separately, archive $\mathcal{A}$ retains every distinct valid scored pair, its scores, and its region, including candidates eliminated from the population. Appendix A details selection, caching, invalid outputs, ties, and incomplete generations.

3.5 BUDGET AND FINAL EVALUATION

Search stops at the localization-query cap Q or proposal cap L. The former counts fresh non-clean calls to G, including invalid responses; the latter counts all proposals, including cached duplicates. Clean grounding, question generation, answering, and answer judging are outside Q. From the archive, we select

$$z^* = \operatorname*{lexmax}_{z \in \mathcal{A}} (f_{\text{ov}}(z), f_{\text{disp}}(z)). \tag{9}$$

Thus, final selection prioritizes lowest IoU, then greatest displacement, with earliest evaluation completion breaking remaining ties. It may return a pair discarded from the population. An empty archive yields no valid candidate.

Population selection preserves alternatives for exploration; the archive's IoU-first rule determines the final return. Algorithm 1 shares counters across initialization and evolution and archives valid pairs before population selection.

The answerer receives $I^* = I_{z^*}$, $q^* = q_{z^*}$, and the saved scored crop $C^* = C_{z^*}$, without a new grounding call. The judge evaluates the answer against the original q and reference answer, so the evaluation target remains fixed.

Algorithm 1: Budgeted JOLT

Input: I, $\mathcal{P}(q)$, R, G; N, Q, L, ϵ, p_t, p_i.
Evaluate(z)**:** Increment T; reuse the cached response or call G and increment B. If the input is new and the response is valid, score z, append it to $\mathcal{A}$, and return it; otherwise return no new candidate.

1. Set $B = T = 0$, $P = \mathcal{A} = \varnothing$.
2. While $|P| < N$, $B < \lfloor Q/2 \rfloor$, and $T < L$:
3. Initialize a pair; add Evaluate(z) to P if a new valid candidate is returned.
4. If $P = \varnothing$, return no valid candidate.
5. While $B < Q$ and $T < L$:
6. Rank P and set $O = \varnothing$.
7. For at most N proposals while $B < Q$ and $T < L$:
8. Select two parents; independently inherit the question index and full perturbation.
9. Resample the index with probability p_t; update the image with probability p_i.
10. Add Evaluate(z) to O if a new valid candidate is returned.
11. Select up to N survivors from $P \cup O$ by NSGA-II, including a partial final O.
12. Select z^* from $\mathcal{A}$ by lowest IoU, greatest displacement, then earliest completion.
13. Run the answerer F for (I_{z^*}, q_{z^*}) using its scored crop, after search has ended.

4 EXPERIMENTS

We evaluate JOLT through three questions that connect downstream attack effectiveness with localization outcomes and performance on additional targets:

RQ1: How much does JOLT degrade judge-based answer accuracy across datasets?

RQ2: How do localization outcomes relate to answer degradation across search strategies and answer retention within JOLT?

RQ3: How effective are independent searches on other target models?

4.1 EXPERIMENTAL SETUP

Data and target. The main evaluation uses the released Visual-CoT dataset (Shao et al., 2024), randomly sampling 500 image–question pairs with seed 74284723 from each of its CUB, Open-Images, Flickr30k, GQA, TextVQA, and VSR source subsets, totaling 3,000 pairs with reference answers and key-region boxes (Wah et al., 2011; Kuznetsova et al., 2020; Plummer et al., 2015; Hudson & Manning, 2019; Singh et al., 2019; Liu et al., 2023a). These subsets cover object attributes, descriptions and relations, compositional questions, scene text, and spatial reasoning. The primary target is VisCoT-224-7B: its grounding output supplies the geometric feedback, and its answerer receives the full image, selected crop, and question. Both stages use greedy decoding. Accuracy measures the end-to-end effect through grounding and answering, including the modified question and perturbed full image. Appendix B identifies the dataset release and model versions.

Configuration and measurement. We use $Q = L = 200$, $N = 16$, and $\|\delta\|_\infty \leq 8/255$; the question-resampling probability is $p_t = 0.35$, the image-update probability is $p_i = 0.70$, and the initial square-area fraction is $p_0 = 0.05$. Initialization and scheduling details appear in Appendix A. GPT-5.4 mini generates 19 variants once before search; with the original question, the pool has $K = 20$ entries. Its instruction requires preserving the queried entity and core attribute or relation while introducing misleading context.

The evaluator, `deepseek-v4-flash`, scores answers against the original question and reference answer. Throughout, accuracy denotes judge-based answer accuracy: the fraction scoring strictly above 0.5 under the rubric in Appendix C, which considers reference agreement and concision.

The question generator and answer judge use their official APIs with explicit thinking disabled and temperature set to 0. All examples entering attack iteration remain in the denominator, including unsuccessful attacks.

We also report IoU between the effective crop and reference region. Both macro-averages use the same samples and attack results, averaging dataset-level metrics with equal weight across the six datasets. Lower accuracy indicates greater answer degradation; lower IoU indicates less region agreement. The drop $\Delta = A_{\text{clean}} - A_{\text{attack}}$ measures net accuracy change, not the fraction of clean-correct answers flipped.

Controls and feedback. Gaussian noise and Defocus blur provide image-corruption controls; AttackVLM (Zhao et al., 2023) and Preemptive Answer (Xu et al., 2024) provide existing attack baselines. All baselines use the same samples as JOLT. Prompt Search edits only questions; Image Search edits only images.

Joint-input controls share editing permissions but vary search organization. Joint Random proposes joint candidates without feedback-guided population selection. Each sequential method allocates budget to exhaustive evaluation of the fixed 20-question pool and assigns the rest to image search. The stages follow the order in the method name, with the first stage's selected component held fixed during the second.

Joint Greedy Search (JGS) maintains one current pair, using JOLT's update operators and $Q = L = 200$ caps. It permits both modalities to change throughout search and accepts lower IoU, or greater center displacement when IoU ties, without a search population or cross-parent recombination. JOLT maintains a population of complete pairs. Joint-input methods share nominal localization-query caps; caching and invalid outputs can change realized calls and usable feedback.

4.2 Attack Effectiveness (RQ1)

Table 1 compares all methods under their respective input-editing permissions. The clean row establishes each dataset's starting performance, and the joint-input controls compare alternative ways to organize the search.

Table 1: Judge-based answer accuracy across six datasets and macro-averaged results, using the same 500 examples per dataset. Macro-averages weight the six datasets equally. Lower values indicate greater answer degradation or less region agreement. Image corruptions, existing attacks, and search-strategy controls are separated.

	Answer accuracy (%)						Macro-average	
Method	CUB	Open Images	Flickr 30k	GQA	Text VQA	VSR	Acc. (%)	IoU
Clean	78.0	71.0	65.0	63.0	77.0	67.0	70.2	0.749
Gaussian	75.0	71.0	63.0	65.0	72.0	58.0	67.3	0.751
Defocus	70.0	66.0	65.0	62.0	73.0	62.0	66.3	0.748
AttackVLM	48.0	33.0	46.0	55.0	43.0	48.0	45.5	–
Preemptive Answer	69.0	61.0	63.0	64.0	57.0	55.0	61.5	–
Prompt Search	48.0	27.0	39.0	45.0	33.0	41.0	38.8	0.219
Image Search	51.0	42.0	43.0	49.0	38.0	44.0	44.5	0.421
Joint Random	42.0	26.0	36.0	41.0	31.0	37.0	35.5	0.194
Joint Greedy Search (JGS)	44.0	36.0	38.0	35.0	33.0	38.0	37.3	0.185
Sequential Prompt–Image	40.0	44.0	35.0	40.0	32.0	33.0	37.3	0.175
Sequential Image–Prompt	38.0	34.0	36.0	37.0	34.0	32.0	35.2	0.187
JOLT	34.0	29.0	27.0	33.0	23.0	28.0	29.0	0.142

Answer degradation. JOLT reduces macro-averaged answer accuracy from approximately 70.2% to 29.0%, with dataset-level drops ranging from 30 percentage points on GQA to 54 on TextVQA.

It yields lower accuracy than Gaussian, Defocus, AttackVLM, Preemptive Answer, Image Search, JGS, and both sequential methods on all six datasets, and than Prompt Search and Joint Random on five. OpenImages is the exception: Prompt Search and Joint Random reach 27% and 26%, respectively, below JOLT's 29%. Thus, the average advantage does not extend to every comparison on every dataset.

Comparison under joint-input access. JOLT's macro-averaged accuracy of 29.0% is below Joint Random's 35.5%, JGS's 37.3%, Sequential Prompt–Image's 37.3%, and Sequential Image–Prompt's 35.2%. Relative to JGS, accuracy is approximately 8.3 percentage points lower and IoU decreases from 0.185 to 0.142, extending JOLT's advantage to feedback-guided single-state search that can keep modifying both modalities under the common localization-query cap.

Question quality. On the same GQA, OpenImages, VSR, and TextVQA samples, we compare original and final selected questions using GPT-2 Large perplexity and all-mpnet-base-v2 embedding cosine similarity (Table 4). Mean similarity ranges from 0.852 to 0.925. Mean perplexity decreases on three datasets and increases from 96 to 104 on GQA. These automatic measures characterize the fluency and semantic proximity of the final selected questions alongside the observed answer degradation.

4.3 Localization and Answer Outcomes (RQ2)

Table 1 reports region agreement and downstream accuracy for the same selected inputs: geometry supplies the search signal, while answer accuracy evaluates its downstream outcome. JOLT reduces mean IoU from 0.749 to 0.142, the lowest among the grounding-search strategies, and has the lowest macro-averaged accuracy overall. However, the metrics do not always rank strategies alike: Sequential Prompt–Image has lower IoU than Joint Random but higher answer accuracy.

Answer retention by post-attack IoU. To examine the relationship within JOLT, we randomly select 200 examples with correct pre-attack answers from each of CUB, GQA, TextVQA, and VSR, for a total of 800 examples. We group examples into five intervals by post-attack IoU. Within each interval, answer retention is the proportion of examples whose post-attack answers remain correct.

Table 2: Answer retention by post-attack IoU for JOLT. Each dataset contributes 200 randomly selected examples with correct pre-attack answers. Entries are the percentages of examples within each interval whose answers remain correct after attack.

Post-attack IoU	CUB	GQA	TextVQA	VSR
$[0, 0.2]$	32%	18%	12%	20%
$(0.2, 0.4]$	35%	22%	16%	24%
$(0.4, 0.6]$	49%	43%	37%	44%
$(0.6, 0.8]$	66%	65%	60%	64%
$(0.8, 1.0]$	82%	83%	79%	81%

Table 2 shows that answer retention increases monotonically across the post-attack IoU intervals in all four datasets, from 12%–32% in the lowest interval to 79%–83% in the highest. Higher post-attack IoU is associated with higher answer retention in these samples. However, some answers remain correct in the lowest IoU interval, while retention is below 100% even in the highest. Region agreement and answer correctness are therefore not equivalent, supporting their separate evaluation.

4.4 Cross-model Evaluation (RQ3)

We conduct independent JOLT searches for LLaVA-1.6-Vicuna-7B and Qwen2.5-VL-7B-Instruct on a 100-example CUB/TextVQA subset, using each target's own localization outputs. Both use greedy decoding for grounding and answering. A prompt specifies the box format and coordinate constraints. Figure 3(a,b) evaluates target-specific search, not transfer from VisCoT. LLaVA accuracy decreases from 57.0% to 37.0% and IoU from 0.360 to 0.151. Qwen2.5-VL accuracy decreases from 78.0% to 43.0% and IoU from 0.470 to 0.240. Both models show reduced region agreement

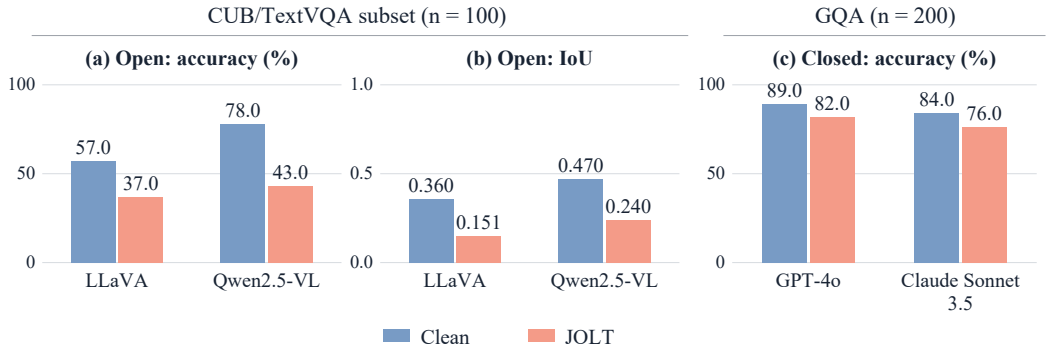

Figure 3: JOLT cross-model effectiveness. Panels (a,b) show judge-based answer accuracy (%) and IoU from independent searches on a 100-example CUB/TextVQA subset for each open-model target. Panel (c) shows judge-based answer accuracy (%) on 200 GQA samples for closed models. Clean and JOLT indicate before and after attack; the protocols are evaluated separately. Model versions appear in Table 3.

and answer accuracy, although the subset protocol prevents a controlled robustness ranking against the main VisCoT evaluation.

A separate evaluation applies JOLT to GPT-4o and Claude Sonnet 3.5 on 200 GQA samples, using the snapshots listed in Appendix B. Figure 3(c) shows GPT-4o accuracy decreasing from 89.0% to 82.0% and Claude Sonnet 3.5 from 84.0% to 76.0%.

5 Discussion and Limitations

Joint random, greedy, and sequential search share JOLT's input-edit permissions and nominal localization-query caps, providing closer comparisons than single-modality attacks. The observed advantage concerns the complete procedure rather than separate gains from persistent perturbations, component inheritance, or Pareto selection.

Task preservation affects how answer degradation should be interpreted. The generation instruction constrains question rewriting, while the norm bound limits image perturbations. The text-quality scores provide proxies for fluency and semantic proximity, but do not verify that the original task is preserved. Because answers are scored against the original task, the measured decline may include effects of task changes and cannot be attributed entirely to robustness failures under preserved semantics.

The findings concern reference-assisted attacks on observable grounding interfaces, where the attacker has both reference regions and returned boxes. Their scope is the evaluated models and workflows. Geometric scores measure the returned regions, while answering also depends on the full image and question. The accuracy changes therefore reflect the complete attacked pipeline rather than the causal effect of crop displacement alone.

6 Conclusion

We studied observable grounding as a feedback interface for reference-assisted joint text–image attacks. JOLT reduces mean judge-based answer accuracy from 70.2% to 29.0% across six datasets and yields lower macro-averaged answer accuracy than the evaluated joint random, greedy, and sequential controls under common localization-query caps. Answer retention increases across the reported post-attack IoU intervals, but region agreement does not determine correctness. This distinction limits the use of localization metrics as proxies for downstream robustness.

AI USE STATEMENT

AI in the research methodology. We use a large language model to generate a fixed pool of question variants before search. During search, the target model returns grounding boxes, and JOLT computes scores from their geometric relationships with the reference region to optimize paired text–image inputs and select candidates. After search, a separate large language model evaluates the correctness of the final answers.

AI assistance in paper writing. Generative AI tools were used primarily for manuscript drafting and language refinement, improving wording, grammar, and clarity. They also assisted with literature discovery, figure preparation, and feedback on analysis design and interpretation of results. The authors remain responsible for the final content, claims, and conclusions.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. In *Computer Vision - ECCV 2020*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123680477.pdf.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

Zhiyuan Chang, Mingyang Li, Junjie Wang, Cheng Li, Boyu Wu, Fanjiang Xu, and Qing Wang. Adversarial Testing for Visual Grounding via Image-Aware Property Reduction. arXiv preprint arXiv:2403.01118, 2024. URL https://arxiv.org/abs/2403.01118.

Ruoxi Cheng, Yizhong Ding, Shuirong Cao, Ranjie Duan, Xiaoshuang Jia, Shaowei Yuan, Simeng Qin, Zhiqiang Wang, and Xiaojun Jia. PBI-Attack: Prior-Guided Bimodal Interactive Black-Box Jailbreak Attack for Toxicity Maximization. In *Proceedings of EMNLP*, pp. 609–628, 2025. URL https://aclanthology.org/2025.emnlp-main.32/.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197, 2002. doi: 10.1109/4235.996017. URL https://ieeexplore.ieee.org/document/996017.

Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial Robustness for Visual Grounding of Multimodal Large Language Models. arXiv preprint arXiv:2405.09981, 2024. URL https://arxiv.org/abs/2405.09981.

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*, pp. 6700–6709, 2019. doi: 10.1109/CVPR.2019.00686. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Shuai Jia, Yibing Song, Chao Ma, and Xiaokang Yang. IoU Attack: Towards Temporally Coherent Black-Box Adversarial Attack for Visual Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6709–6718, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Jia_IoU_Attack_Towards_Temporally_Coherent_Black-Box_Adversarial_Attack_for_Visual_CVPR_2021_paper.html.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 2020. doi: 10.1007/s11263-020-01316-z. URL https://research.google/pubs/the-open-images-dataset-v4-unified-image-classification-object-detection-and-visual-relationship-detection-at-scale/.

Xuanhui Lin, Junhao Dong, Mingrong Gong, Yucheng Chen, Xinghua Qu, and Yew-Soon Ong. Two Sides of the Same Coin: Co-Evolving Search for Cross-Task Attacks on Vision-Language Models. arXiv preprint arXiv:2608.02137, 2026. URL https://arxiv.org/abs/2608.02137.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, 2023b. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge. LLaVA project blog, 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 102–111, 2023. doi: 10.1109/ICCV51070.2023.00016. URL https://openaccess.thecvf.com/content/ICCV2023/html/Lu_Set-level_Guidance_Attack_Boosting_Adversarial_Transferability_of_Vision-Language_Pre-training_Models_ICCV_2023_paper.html.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Plummer_Flickr30k_Entities_Collecting_ICCV_2015_paper.pdf.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. URL https://www.vision.caltech.edu/datasets/cub_200_2011/.

Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop Reasoning! When Multimodal LLM with Chain-of-Thought Reasoning Meets Adversarial Image. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2402.14899.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract.html.

Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR20

24/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. BadChain: Backdoor Chain-of-Thought Prompting for Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.12242.

Chunxue Xu, Yiwei Wang, Yujun Cai, Bryan Hooi, and Songze Li. Visual CoT Makes VLMs Smarter but More Fragile. arXiv preprint arXiv:2509.23789, 2025. URL https://arxiv.org/abs/2509.23789.

Rongwu Xu, Zehan Qi, and Wei Xu. Preemptive Answer "Attacks" on Chain-of-Thought Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14708–14726, 2024. doi: 10.18653/v1/2024.findings-acl.876. URL https://aclanthology.org/2024.findings-acl.876/.

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling Vision and Language Models Despite Localization and Attention Mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4951–4961, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Xu_Fooling_Vision_and_CVPR_2018_paper.html.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. VLAttack: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a5e3cf29c269b041ccd644b6beaf5c42-Abstract.html.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards Adversarial Attack on Vision-Language Pre-training Models. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. doi: 10.1145/3503161.3547801. URL https://doi.org/10.1145/3503161.3547801.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00923.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On Evaluating Adversarial Robustness of Large Vision-Language Models. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a97b58c4f7551053b0512f92244b0810-Abstract-Conference.html.

A SEARCH DETAILS

This appendix specifies the execution rules supporting Section 3. The default parameters are population capacity $N = 16$, localization-query cap $Q = 200$, proposal cap $L = 200$, perturbation bound $\epsilon = 8/255$, question-resampling probability $p_t = 0.35$, image-update probability $p_i = 0.70$, and initial square-area fraction $p_0 = 0.05$. The experimental configuration is described in Section 4.

A.1 QUESTION POOL AND INITIAL IMAGE STATES

GPT-5.4 mini returns the original question and 19 variants in a single response from the original image and question. The original question is the first entry, followed by the generated variants, without an additional deduplication pass, giving a pool of $K = 20$ entries. Appendix C provides the generation template. The generator is called through the official API with explicit thinking disabled and temperature set to 0. No new language generation occurs during search.

$$\mathcal{P}(q) = (q^{(0)}, \ldots, q^{(K-1)}), \qquad q^{(0)} = q. \tag{10}$$

Initialization samples a pool index uniformly and uses the stripe initialization of Square Attack for the image component (Andriushchenko et al., 2020). For the image component, it independently samples signs in $\{-\epsilon, +\epsilon\}$ for each column and color channel, broadcasts them vertically, and clips the resulting image to $[0, 1]$. The stored tensor is the clipped image minus the original image. Initialization admits up to N distinct valid candidates, using at most $\lfloor Q/2 \rfloor$ fresh localization calls and respecting the overall proposal cap. A nonempty undersized population proceeds without padding and may grow during subsequent selection. If initialization obtains no valid candidate, the run returns no valid candidate.

Because the stored tensor is the clipped image minus the original, pixels near a range boundary can retain smaller changes than the proposed update.

A.2 EFFECTIVE CROP AND COORDINATE CONVENTION

For an attacked pair (I_z, q_z), let $b_z = (u_1, v_1, u_2, v_2)$ denote the normalized returned box in the target's coordinate system. VisCoT uses coordinates on a centered, square-padded image. Following its official coordinate conversion, let $D = \max(W, H)$, $p_x = \lfloor (D-W)/2 \rfloor$, and $p_y = \lfloor (D-H)/2 \rfloor$. The original-image coordinates are

$$x_j = Du_j - p_x, \qquad y_j = Dv_j - p_y, \qquad j \in \{1, 2\}. \tag{11}$$

Thus all four VisCoT coordinates are scaled by D before removing padding; they must not be scaled directly by W and H. This follows the scaling and offset removal in the official loader.[1] For the other targets, whose prompts request coordinates normalized by the original image width and height, use $x_j = Wu_j$ and $y_j = Hv_j$ instead.

After coordinate conversion, JOLT clips to the original image and uses an unexpanded pixel region:

$$C_z = \big([\lfloor x_1 \rfloor, \lceil x_2 \rceil) \times [\lfloor y_1 \rfloor, \lceil y_2 \rceil)\big) \cap \big([0, W) \times [0, H)\big). \tag{12}$$

This retains JOLT's no-expansion crop rule rather than the official loader's subsequent square expansion with a minimum half-side of 112 pixels. A box with no positive-area intersection is invalid. Both geometric scoring and downstream cropping use the same saved C_z. Reference region R is expressed in original-image pixel coordinates; padding is not added to it when computing these objectives. The same conversion applies to clean grounding.

A.3 QUERY ACCOUNTING AND CANDIDATE IDENTITY

Every proposed candidate increments the proposal counter T. A fresh call to the grounder increments the localization counter B, whether the response is valid or invalid. Valid responses contain

[1] https://github.com/deepcs233/Visual-CoT/blob/main/llava/eval/model_cot_loader.py

four finite coordinates satisfying $0 \le u_1 < u_2 \le 1$ and $0 \le v_1 < v_2 \le 1$ in the target's coordinate system and yield a positive-area C_z after conversion and clipping. Invalid outputs receive no objective vector and do not enter the population or archive.

The cache is indexed by the actual question text and submitted image under fixed preprocessing, prompts, and decoding. It stores the first output for that complete input. A cached duplicate consumes a proposal but no fresh localization call, and the candidate is not added again. This identity rule applies even if identical question text occupies different positions in the question pool. Changing the question or submitted image creates a different complete input whose region cannot be inferred from the previous pair.

Clean grounding is excluded from B, as are question generation, downstream answering, and answer judging. These operations require separate cost records. Since cached and invalid proposals consume different resources, both B and T are needed to describe an actual run. The nominal cap alone does not specify how many valid candidates were scored.

Equal maximum localization budgets do not establish equal total model cost or guarantee equal realized localization counts across search methods.

A.4 SQUARE SCHEDULE

Let B_0 be the number of non-clean localization calls used by initialization and B the current count. Before an image update, define

$$r = \frac{B - B_0}{Q - B_0}, \tag{13}$$

$$p(r) = p_0\, 2^{-m(r)}, \tag{14}$$

$$m(r) = \sum_{\tau \in \mathcal{T}} \mathbf{1}[10000r > \tau], \tag{15}$$

$$\mathcal{T} = \{10, 50, 200, 500, 1000, 2000, 4000, 6000, 8000\}, \tag{16}$$

$$s = \min\Big\{H, W, \max\Big\{1, \operatorname{round}\sqrt{p(r)HW}\Big\}\Big\}. \tag{17}$$

The threshold sequence follows the `p_selection` function in the authors' released Square Attack implementation (Andriushchenko et al., 2020)[2]; here progress is indexed by localization-query consumption after initialization. A window of side s is placed by sampling its top-left coordinate uniformly from $\{0, \dots, H - s\} \times \{0, \dots, W - s\}$. The signs in Equation 6 are independent across channels and constant spatially within the window.

A.5 SELECTION, PARTIAL GENERATIONS, AND TIES

A candidate dominates another if it is no worse on either objective and strictly better on at least one. Non-dominated sorting partitions the combined population into successive fronts. Parent tournaments sample with replacement and prefer lower non-dominated rank, then greater crowding distance. Each child's two components are independently inherited from the two selected parents. After mutation and evaluation, environmental selection admits fronts in rank order and truncates the last front by descending crowding distance. Constant objective dimensions contribute zero to crowding distance. Remaining selection ties are resolved using the run seed.

A generation proposes at most N offspring; this limit includes duplicates and candidates yielding invalid outputs. When a cap is reached partway through a generation, all completed valid offspring still enter environmental selection. The archive receives every distinct valid scored candidate before population truncation. Final archive selection uses Equation 9, with earliest evaluation completion resolving a remaining tie.

A.6 ANSWER EVALUATION

For the selected pair, the answerer reuses its saved scored region:

$$a^* = F([I^*, \mathrm{Crop}(I^*, C^*)], T_F(q^*)). \tag{18}$$

[2] `https://github.com/max-andr/square-attack/blob/master/attack.py`

B Additional Evaluation Details

B.1 Dataset release and model versions

The evaluation uses the Visual-CoT release at `https://huggingface.co/datasets/deepcs233/Visual-CoT`, including its questions, reference answers, and key-region annotations for the six source subsets in Section 4.1. Table 3 lists the model identifiers used in the experiments.

Table 3: Model versions used for target evaluation, question generation, and answer judging.

Model	Checkpoint or API identifier
VisCoT	`viscot-224-7b`
LLaVA (Liu et al., 2024)	`liuhaotian/llava-v1.6-vicuna-7b`
Qwen2.5-VL	`Qwen/Qwen2.5-VL-7B-Instruct`
GPT-4o	`gpt-4o-2024-08-06`
Claude Sonnet 3.5	`claude-3-5-sonnet-20241022`
GPT-5.4 mini (generator)	`gpt-5.4-mini`
DeepSeek (judge)	`deepseek-v4-flash`

VisCoT-224-7B, LLaVA-1.6-Vicuna-7B, and Qwen2.5-VL-7B-Instruct use greedy decoding for both grounding and answering. GPT-5.4 mini generates the question variants, and DeepSeek-V4-Flash judges answers against the original question and reference answer. Both are accessed through their official APIs with explicit thinking disabled and temperature set to 0. Judge-based accuracy counts scores strictly greater than 0.5 as correct under the rubric in Appendix C.

B.2 Text quality of final selected questions

This analysis uses the same 500 examples per dataset from the main GQA, OpenImages, VSR, and TextVQA evaluation, covering 2,000 examples. For each example, we compare the original question q_i with the final question q_i^* selected by JOLT. The measurements concern the submitted final questions rather than an average over the candidate pool.

We compute perplexity separately for each original and modified question using GPT-2 Large. Each question's perplexity is the exponential of its token-averaged negative log-likelihood. We then take the arithmetic mean of these per-question perplexities within each dataset, giving every question equal weight:

$$\overline{\mathrm{PPL}}_{\mathrm{orig}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{PPL}(q_i), \qquad \overline{\mathrm{PPL}}_{\mathrm{final}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{PPL}(q_i^*). \tag{19}$$

For semantic similarity, E denotes the sentence encoder `all-mpnet-base-v2`. We compute cosine similarity for each original–final question pair and take the arithmetic mean within the dataset:

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n}\frac{E(q_i)^\top E(q_i^*)}{\|E(q_i)\|_2\,\|E(q_i^*)\|_2}. \tag{20}$$

Table 4: Text quality of final JOLT questions on the same samples as the main evaluation (500 per dataset). PPL is measured by GPT-2 Large; cosine similarity uses all-mpnet-base-v2. All entries are arithmetic means over examples within each dataset.

Dataset	Original PPL	Final PPL	Cosine similarity
GQA	96	104	0.891
OpenImages	195	172	0.925
VSR	145	108	0.885
TextVQA	204	174	0.852

C Prompt Templates and Grounding Requirements

The following templates specify question-pool generation, grounding, answering, and answer scoring. Braced fields denote per-example substitutions. Image inputs are supplied separately from the text.

C.1 Question-pool generation

GPT-5.4 mini receives the original image and question once before search. The response contains the original question as its first entry and 19 variants, forming the fixed pool of $K = 20$.

> **System instruction.** You are an expert prompt optimizer with a keen eye for image detail integration. Your task is to revise the given original prompt and guide the model to observe information in non-critical areas while strictly preserving its core semantic meaning.
>
> **Core Requirements:** 1. Strictly preserve the core semantic meaning of the original prompt and never alter the essence of the question.
>
> **Key Definitions:** Non-critical areas = background corners outside the 80% central area of the image, edge details within 10% of the image border, secondary objects that occupy less than 15% of the image area and are not the main focus.
>
> **Revision & Guidance Rules:** 1. Ensure the revised prompt fully completes the original image task as specified, with the original prompt's content naturally integrated into the revised one to avoid semantic deviation. 2. Naturally integrate guidance to observe non-critical areas, list 2 to 3 key details from non-critical areas and briefly explain how they enrich the image context, then incorporate the semantics of the original prompt. 3. Use diverse styles such as scenario-based assumptions, detail exploration prompts, task-related guidance, role-playing, and others. Avoid monotony. 4. Strictly preserve the original prompt's core semantic meaning. 5. The revised prompt cannot raise new questions. 6. After generating the prompt, check whether it retains the core meaning of the original prompt and does not use parentheses. 7. Ensure diversified prompt lengths.
>
> **Output Requirements:** 1. Output exactly 20 complete prompts: the unchanged original prompt followed by 19 revised prompts. 2. Use a numbered list from 1 to 20. Item 1 must reproduce the original prompt exactly; items 2 through 20 must be revised prompts following the rules above. 3. Do not output explanations or extra text.
>
> **User text.**
>
> Original Prompt: `{original_question}`
>
> Please inspect the image and return a collection of 20 prompts containing the unchanged original prompt and 19 revised prompts according to the rules.

C.2 Grounding

For VisCoT-224-7B, the original localization suffix from Visual CoT (Shao et al., 2024) is appended to the question, with the image supplied as visual input. The question is the original question for clean evaluation and the candidate question during search. VisCoT returns square-padded normalized coordinates, converted to original-image coordinates as specified in Appendix A.

> `{question}`
>
> Please provide the bounding box coordinate of the region that can help you answer the question better.

LLaVA-1.6-Vicuna-7B, Qwen2.5-VL-7B-Instruct, GPT-4o, and Claude Sonnet 3.5 share the same grounding prompt: the VisCoT prompt above with the following coordinate requirements appended: four coordinates in the order $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, normalized to $[0, 1]$ by image width and height, respectively, with $x_{\min} < x_{\max}$ and $y_{\min} < y_{\max}$.

C.3 ANSWERING

All target models use the following answer template with the full image and selected local crop. Clean evaluation uses the original question; attacked evaluation uses the final selected question, perturbed full image, and crop from its saved grounding region. In the template, *original image* denotes the full image.

> `{question}`
> Please Answer based on the original image and local crop.

C.4 ANSWER SCORING

DeepSeek-V4-Flash receives the original question, reference answer, and model answer after search, without providing search feedback. The scoring template is adapted from Visual CoT (Shao et al., 2024), with additional instructions on errors, contradictions, and concision. Scores strictly greater than 0.5 count as correct.

> You are responsible for proofreading the answers, you need to give a score to the model's answer by referring to the standard answer, based on the given question. The full score is 1 point and the minimum score is 0 points. Please output the score in the form `score: <score>`. The evaluation criteria require that the closer the model's answer is to the standard answer, the higher the score. If there are errors or contradictions, reduce the score. Simple and accurate answers should be encouraged, while complex and redundant answers should be penalized.
> Question:
> `{original_question}`
> Ground Truth Answer:
> `{reference_answer}`
> Model Generated Answer:
> `{model_answer}`