RETHINKING POST-PRUNED EVIDENCE: COUNTERFACTUAL DECISION REASONING FOR FIXED-BUDGET VISUAL TOKEN REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual token pruning improves the efficiency of vision-language large models (VLLMs) but often discards query-relevant evidence under aggressive compression, leading to degraded model decisions. Existing post-pruning recovery methods restore discarded information according to predefined relevance or representation criteria, without explicitly evaluating how the resulting token composition affects the current decision. Consequently, restoration introduces a recovery–risk trade-off: useful evidence is recovered, while unfavorable interventions disrupt otherwise reliable predictions. To address this issue, we propose Counterfactual-Aware Restoration through Token Exchange (CARE), a training-free and fixed-budget post-pruning recovery framework based on query-adaptive token exchange and counterfactual decision-guided selection. Ours first identifies replaceable retained tokens and retrieves complementary discarded evidence to construct diverse one-to-one exchange proposals without increasing the visual-token budget. Ours then treats alternative token compositions as counterfactual interventions and evaluates their decision-level effects against the original pruned representation using the frozen VLLM. An alternative composition is adopted only when it provides stronger decision support than the no-change reference. Experiments across multiple benchmarks, pruning methods, and token budgets demonstrate consistent performance improvements, substantially fewer recovery-induced errors, and preserved pruning efficiency without additional training or token-count increase.

1 INTRODUCTION

Vision-language large models (VLLMs) rely on hundreds of visual tokens to represent high-resolution inputs, resulting in substantial inference cost. Visual token pruning methods He & Xiao (2023); Mao et al. (2025); Bolya et al. (2022) reduces this cost by removing redundant visual information, but aggressive compression can also discard decision-critical evidence and disrupt query-relevant information available before pruning. As a result, the model shift from a correct prediction to an incorrect one after pruning. This limitation has motivated post-pruning recovery methods Chen et al. (2025); Cho et al. (2026); Wang et al. that recover discarded visual information or compensate for representation degradation after token reduction. Existing methods determine which visual evidence to restore Chen et al. (2025); Cho et al. (2026) or rectify Wang et al. based on predefined criteria such as text–visual relevance, safety, or representation quality, enabling increasingly effective recovery from pruning-induced information loss. These developments have substantially improved the ability to recover information lost during aggressive visual-token compression. However, current post-pruning recovery methods alter the model's original decision because restored tokens are introduced according to relevance or representation criteria without evaluating their decision-level consequences. The resulting token composition changes the evidence supporting prediction and may introduce redundant or competing visual information. Consequently, restoration that appears informative at the token level does not necessarily strengthen the current decision and can even overturn an originally correct prediction. The key challenge is therefore not only what evidence to restore, but also whether the resulting composition provides stronger decision support than the original pruned representation.

Fig. 1 illustrates how post-pruning recovery both recovers useful evidence and introduces new decision errors. As shown in Fig. 1(a), a base pruner (e.g., OC-VTP Li et al. (2026)) retains a compact

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.

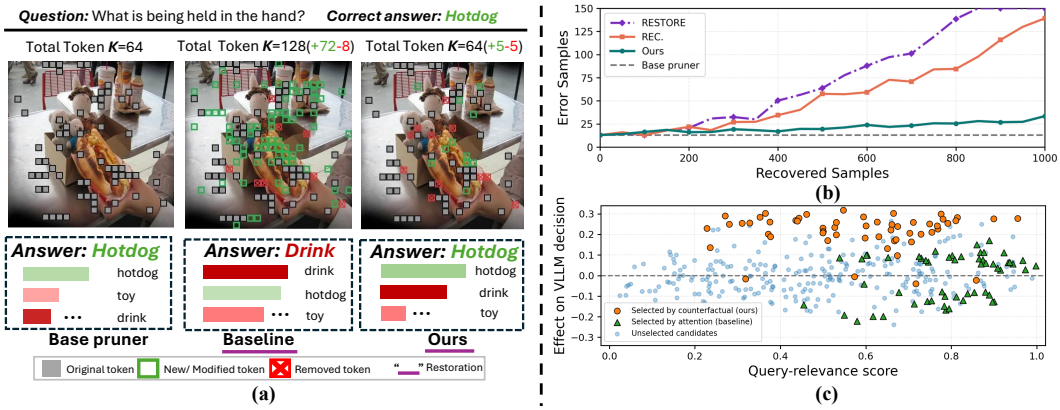


Figure 1: Recovery–risk comparison. (a) Interference with the original decision during post-pruning recovery. (b) The increasing occurrence of recovery-induced errors as post-pruning recovery is applied to more samples. (c) Analysis of the impact of query-relevant tokens on the final decision.

token representation that already produces the correct prediction *Hotdog*, whereas a conventional restoration baseline (e.g., REC. Chen et al. (2025)) reintroduces a large number of discarded visual tokens, increasing the token count from 64 to 128, but shifts the prediction to the incorrect answer *Drink*. Consequently, as recovery is applied to an increasing number of samples, the newly introduced visual information increasingly interferes with previously correct decisions, causing some of them to shift toward incorrect predictions, as illustrated in Fig. 1(b). To further identify the source of this behavior, Fig. 1(c) compares query-relevance scores with the actual effects of token interventions on VLLM decisions. Although highly relevant tokens often provide positive decision support, a substantial portion produces only weak or negative effects after recovery. This mismatch reveals a key limitation of relevance-based recovery: token relevance does not reliably indicate whether introducing the corresponding visual evidence strengthens or weakens the current decision.

To address this recovery–risk trade off, we propose novel **C**ounterfactual **A**ware **R**estoration through Token **E**xchange (**CARE**), a training free post pruning recovery framework that restores potentially useful discarded evidence through fixed budget token exchange and explicitly evaluates the decision level effect of each alternative composition before adoption. CARE consists of two main components: Diverse High Recall Exchange Proposal and Counterfactual Decision Guided Selection. To improve the coverage of useful discarded evidence without increasing the token budget, the first component identifies replaceable retained tokens, retrieves promising discarded evidence, and constructs diverse one to one token exchanges. As shown in Fig. 1(a), this fixed budget exchange restores visual evidence while preserving the original token count and prediction. To prevent recovery from introducing unfavorable decision changes, the second component evaluates alternative exchange compositions with the frozen VLLM relative to the original pruned representation. We treat each alternative composition as a counterfactual intervention on the same image and query while keeping the model fixed. As shown in Fig. 1(c), counterfactual evaluation better distinguishes interventions with positive and negative effects on VLLM decisions. Consequently, as shown in Fig. 1(b), CARE substantially suppresses the growth of recovery induced decision errors without additional training or any increase in visual token count. Our main contributions are summarized as follows:

- We identify a recovery–risk trade off in post-pruning correction, where restoring relevant visual evidence improves pruning-induced failures but unnecessary intervention introduces new prediction errors.
- We propose CARE, a training-free and fixed-budget framework that combines query-adaptive token exchange with counterfactual decision verification to jointly account for token-level relevance and decision-level effects.
- Extensive experiments across multiple visual-token pruners, benchmarks, and pruning budgets demonstrate consistent task-performance improvements, fewer recovery-induced errors, and preservation of the efficiency benefits of visual-token pruning.

2 Related Work

2.1 Counterfactual Reasoning in Vision–Language Models

Counterfactual reasoning has been widely explored in vision–language learning to analyze how controlled changes in multimodal evidence affect model predictions. Early studies construct counterfactual visual or linguistic samples to reduce spurious correlations Abbasnejad et al. (2020), improve VQA robustness Chen et al. (2020), and mitigate language bias Niu et al. (2021). More recent work extends counterfactual intervention to broader settings. CF-VLM Cai et al. (2026) introduces counterfactual examples into vision–language fine-tuning, while representation-level counterfactual calibration suppresses context-dependent shortcuts during recognition Peng et al. (2026). Counterfactual strategies have also been applied to multimodal bias mitigation Howard et al. (2024), hallucination reduction in large vision–language models Wu et al. (2025b); Liang et al. (2026), and visual in-context learning Xiong et al. (2026). These studies demonstrate that comparing model behavior under alternative evidence provides an effective way to assess the decision impact of multimodal interventions. However, counterfactual evaluation of alternative visual-token compositions after pruning remains largely underexplored, particularly for determining whether a restored composition should replace the original pruned representation.

2.2 Post-Pruning recovery

Post-pruning recovery aims to recover visual information or compensate for representation degradation after token reduction. Recoverable Compression restores discarded tokens according to text–visual relevance Chen et al. (2025), while SparseVLM Zhang et al. (2024), LLaVA-PruMerge Shang et al. (2025), and VFlowOpt Yang et al. (2025) preserve pruned information through token recycling or merging. Hi-Lo Prune further transfers useful information before token removal Sun et al. (2026). SAP Wang et al. restores benign visual tokens for safety preservation, whereas RESTORE Cho et al. (2026) alleviates positional and attentional distortion in compressed representations. Despite these advances, existing methods mainly determine recovery using predefined criteria such as relevance, importance, safety, or representation quality. The resulting composition is generally adopted without explicitly comparing its decision-level effect with the original pruned representation. As additional or misaligned evidence can disturb already reliable predictions, post-pruning recovery should consider not only what information to restore, but also whether the restored composition provides stronger decision support.

3 Preliminary

Let vision-token set $\mathcal{V} = \{v_i\}_{i=1}^{N}$ denote the complete N visual tokens and q the current query. Given a target budget K, an existing visual-token pruner P produces an initial retained set $\mathcal{K}_0$ and its discarded complement $\mathcal{D}_0$. Accordingly, the initial partition is written as:

$$\mathcal{K}_0 = P(\mathcal{V}, q; K), \qquad \mathcal{D}_0 = \mathcal{V} \setminus \mathcal{K}_0, \qquad |\mathcal{K}_0| = K. \tag{1}$$

Post-pruning recovery operates on this partition by selectively recovering useful information from $\mathcal{D}_0$ to compensate for evidence lost during pruning. In our setting, recovery modifies the retained composition under the same budget K, while keeping both the original pruner and VLLM frozen.

4 Method

CARE performs post-pruning recovery under the same token budget through two components. **Diverse High-Recall Exchange Proposal** constructs one-to-one exchanges between replaceable retained tokens and query-relevant discarded evidence. **Counterfactual Decision-Guided Selection** then evaluates alternative exchange subsets against the original pruned representation, prioritizing prediction-preserving compositions and requiring positive counterfactual support for prediction changes. The original representation remains an explicit no-change option.

4.1 Diverse High-Recall Exchange Proposal

Starting from the initial pruned representation, we first construct a diverse set of recovery candidates that covers useful discarded evidence. To broaden the candidate space without increasing the

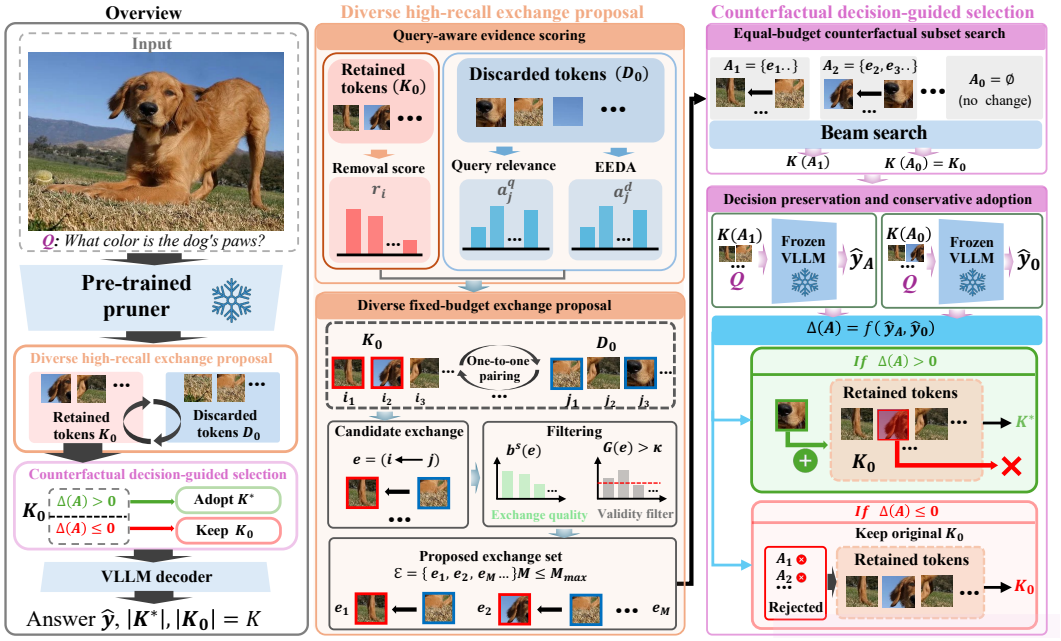


Figure 2: Overview of CARE. Diverse High-Recall Exchange Proposal constructs one-to-one token exchanges using query relevance and EEDA (Excluded-Evidence Direction Alignment) under a fixed visual-token budget, while Counterfactual Decision-Guided Selection searches alternative compositions and conservatively adopts them according to their decision-level effects.

visual-token budget, we identify replaceable retained tokens and retrieve promising discarded tokens through complementary evidence signals. These candidates are subsequently combined into diverse exchange proposals for decision-level evaluation.

Query-aware evidence scoring. To identify visual evidence associated with the current query, we first estimate the relevance of every visual token v_i using frozen pre-pruning representations. Let $\bar{v}_i$ denote the projected representation of v_i and $\bar{q}$ the mean-pooled query representation in the same feature space. The query-conditioned relevance distribution is expressed as:

$$\alpha_i(q) = \frac{\exp(\cos(\bar{v}_i, \bar{q})/T)}{\sum_{n=1}^{N} \exp(\cos(\bar{v}_n, \bar{q})/T)}, \tag{2}$$

where $T > 0$ controls the concentration of the distribution, and larger $\alpha_i(q)$ indicates stronger query relevance.

To identify retained tokens that can be replaced with minimal information loss, we first measure their redundancy within the retained set. Let $\rho_i = \max_{v_\ell \in \mathcal{K}_0, \ell \neq i} \cos(\bar{v}_i, \bar{v}_\ell)$ denote the redundancy of retained token v_i. Since redundancy and relevance have different numerical ranges, additive terms are min–max normalized within their corresponding candidate sets, where a tilde denotes the normalized value. The resulting replaceability score r_i is:

$$r_i = \lambda_r \, \widetilde{\rho}_i - \widetilde{\alpha}_i(q), \qquad v_i \in \mathcal{K}_0, \tag{3}$$

where λ_r balances normalized redundancy and query relevance. A larger r_i indicates greater redundancy and weaker query relevance, making the retained token more suitable for replacement.

To characterize the query-related evidence excluded by the initial retained representation, we next aggregate relevant information remaining in $\mathcal{D}_0$. This excluded-evidence direction d_0 is computed as:

$$d_0 = \frac{\sum_{v_j \in \mathcal{D}_0} \alpha_j(q) \, \bar{v}_j}{\sum_{v_j \in \mathcal{D}_0} \alpha_j(q) + \epsilon}. \tag{4}$$

Here, ϵ is a small numerical constant, and d_0 summarizes the query-related visual evidence that remains outside the retained token set.

To retrieve discarded evidence with high coverage, we evaluate each discarded token from two complementary perspectives: its direct relevance to the query and its alignment with the excluded-evidence direction. Their corresponding retrieval scores are defined as:

$$a_j^q = \widetilde{\alpha}_j(q), \qquad a_j^d = \widetilde{\alpha}_j(q)\left[\cos(\bar{v}_j, \bar{d}_0)\right]_+, \qquad v_j \in \mathcal{D}_0, \tag{5}$$

where $\bar{d}_0 = d_0/(\|d_0\|_2 + \epsilon)$ denotes the normalized excluded-evidence direction and $[\cdot]_+$ retains only positive alignment. The two scores provide complementary candidate pools: a_j^q directly retrieves query-relevant evidence, whereas a_j^d emphasizes query-relevant tokens consistent with evidence excluded from the retained representation. Candidates are ranked independently within the two sources, after which the source-specific scores are normalized and the retrieved candidates are merged to preserve diversity across both retrieval criteria.

Diverse fixed-budget exchange proposal. To convert the retrieved evidence into fixed-budget recovery candidates, we pair a highly replaceable retained token $u \in \mathcal{K}_0$ with a promising discarded token $w \in \mathcal{D}_0$ retrieved from one of the two evidence sources. Each pair forms an exchange candidate $e = (u \to w)$, where source $s \in \{q, d\}$ indicates whether w is retrieved by direct query relevance or excluded-evidence alignment. We then define the proposal priority and query-evidence change of each exchange as:

$$b^s(e) = \widetilde{r}_u + \widetilde{a}_w^s, \qquad G(e) = \alpha_w(q) - \alpha_u(q). \tag{6}$$

The proposal score $b^s(e)$ jointly reflects the replaceability of the retained token and the recovery quality of the discarded token, while $G(e)$ measures the resulting change in query relevance. Candidates satisfying $G(e) > \kappa$ undergo greedy non-overlapping selection without reusing retained or discarded tokens, where $\kappa < 0$ is a permissive lower bound that filters strongly unfavorable exchanges while preserving a high-recall candidate space. The resulting proposal set $\mathcal{E} = \{e_1, \ldots, e_M\}$ contains at most $M_{\max}$ exchanges, i.e., $M \le M_{\max}$. Since every $e_m = (u_m \to w_m)$ removes one token and inserts one token, the visual-token budget remains exactly K. Final adoption is determined by the subsequent decision-guided selection.

4.2 Counterfactual Decision-Guided Selection

Broad proposal generation improves access to query-relevant evidence, but evidence gain does not always translate into stronger decision support. CARE therefore uses the pre-recovery prediction as a original decision and adopts an alternative composition only when it receives stronger model support; otherwise, the original pruned representation is retained.

Equal-budget counterfactual subset search. A fixed progressive exchange order overlooks useful non-prefix combinations. We therefore perform beam search over $\mathcal{E}$ in descending proposal priority. Starting from $\mathcal{A}_0 = \varnothing$, each retained subset is expanded by excluding or including the current exchange, while $\varnothing$ is always retained as the no-change reference. Candidate subsets are explored through beam search, while only non-empty subsets with positive aggregate query-evidence gain are retained for decision-level evaluation by the frozen VLLM.

For any explored subset $\mathcal{A} \subseteq \mathcal{E}$, the corresponding fixed-budget composition is:

$$\mathcal{K}(\mathcal{A}) = (\mathcal{K}_0 \setminus \{u_m \mid e_m \in \mathcal{A}\}) \cup \{w_m \mid e_m \in \mathcal{A}\}, \qquad |\mathcal{K}(\mathcal{A})| = K, \tag{7}$$

where $\mathcal{K}(\{\varnothing\}) = \mathcal{K}_0$. Since all exchanges are one-to-one and non-overlapping, every explored composition preserves the original budget. To avoid favoring larger subsets, we define $R(\mathcal{A}) = |\mathcal{A}|^{-1} \sum_{e_m \in \mathcal{A}} G(e_m)$ for non-empty subsets, with $R(\varnothing) = 0$. Because the permissive proposal threshold $\kappa < 0$ retains near-neutral and mildly negative individual exchanges, $R(\mathcal{A}) > 0$ ensures that an evaluated subset still provides a positive aggregate gain in query-relevant evidence.

Decision preservation and conservative adoption. Let $\hat{y}_0$ denote the pre-recovery prediction induced by $\mathcal{K}_0$ and $\hat{y}_\mathcal{A}$ the prediction induced by $\mathcal{K}(\mathcal{A})$. We use $\hat{y}_0$ as a original decision rather than ground truth and measure whether an alternative composition provides stronger decision support than the original representation:

$$\Delta(\mathcal{A}) = \ell(\hat{y}_\mathcal{A}; \mathcal{K}(\mathcal{A})) - \ell(\hat{y}_0; \mathcal{K}_0), \tag{8}$$

where $\ell(\cdot)$ denotes the length-normalized log-probability computed by the frozen VLLM under the same query. A positive $\Delta(\mathcal{A})$ indicates that the candidate prediction receives stronger model-internal support than the pre-recovery original decision, without implying ground-truth correctness.

Let $\mathcal{S}$ denote non-empty subsets satisfying $R(\mathcal{A}) > 0$. The final subset is selected as:

$$\mathcal{A}^* = \begin{cases} \arg\max_{\mathcal{A}\in\mathcal{S}} \Delta(\mathcal{A}), & \max_{\mathcal{A}\in\mathcal{S}} \Delta(\mathcal{A}) > 0, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{9}$$

The final composition is $\mathcal{K}^* = \mathcal{K}(\mathcal{A}^*)$. CARE therefore adopts an evidence-improving alternative only when it receives stronger decision support than the original pruned representation; otherwise, it retains $\mathcal{K}_0$. The framework remains training-free, keeps the original pruner and VLLM frozen, and satisfies $|\mathcal{K}^*| = |\mathcal{K}_0| = K$.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on four vision-language benchmarks: GQA Hudson & Manning (2019), POPE Li et al. (2023), VizWiz Gurari et al. (2018), and SEED-Bench Li et al. (2024). GQA evaluates compositional visual reasoning, POPE focuses on object hallucination, VizWiz provides challenging real-world visual question answering, and SEED-Bench evaluates general multimodal understanding. For task performance, we use exact-match accuracy on GQA and VizWiz, accuracy on POPE, and the overall score on SEED-Bench.

Metrics. In addition to task performance, we evaluate recovery reliability, evidence localization, and intervention magnitude using the number of damaged samples (#Dmg), Recovery-Induced Error (RIE), Pointing Game (PG) Zhang et al. (2018), and Recovery Intervention Ratio (RIR). Let $\mathcal{S}$ denote samples correctly predicted by the base pruner and $\mathcal{D} \subseteq \mathcal{S}$ those changed to incorrect predictions after recovery. RIE is defined as

$$\text{RIE} = \frac{|\mathcal{D}|}{|\mathcal{S}|} \times 100\%, \tag{10}$$

where $|\mathcal{D}|$ corresponds to #Dmg. PG measures whether the highest-response location falls within the query-relevant ground-truth region. To quantify intervention magnitude, RIR measures the fraction of newly introduced visual tokens relative to the original pruning budget:

$$\text{RIR} = \frac{1}{|\mathcal{T}|} \sum_{i\in\mathcal{T}} \frac{|\mathcal{K}_i^{\text{rest}} \setminus \mathcal{K}_{0,i}|}{K} \times 100\%, \tag{11}$$

where $\mathcal{K}_{0,i}$ and $\mathcal{K}_i^{\text{rest}}$ denote the token compositions before and after recovery, respectively, and $\mathcal{T}$ is the evaluation set. Lower #Dmg and RIE indicate more reliable recovery, lower RIR indicates more selective intervention, and higher PG indicates better localization of query-relevant visual evidence.

Implementation Details. We use LLaVA-1.5-7B Grattafiori et al. (2024) as the backbone VLLM and keep all pretrained parameters frozen throughout evaluation. CARE is applied to two representative visual-token pruners, VisPruner Zhang et al. (2025) and OC-VTP Li et al. (2026), and is compared with Recoverable Compression (REC.) Chen et al. (2025) and RESTORE Cho et al. (2026). Main task performance is evaluated with $K \in \{48, 64\}$ visual tokens, while recovery reliability additionally includes the more aggressive $K = 32$ setting. Unless otherwise specified, the maximum number of candidate exchanges is set to $M_{\max} = 8$, and the beam size for counterfactual subset search is set to $B = 4$. CARE requires no additional training and performs one-to-one token exchanges, keeping the visual-token budget fixed for every sample.

Table 1: Task performance across base pruners and visual-token budgets.

Method	Retain 48 Tokens (8.3%)				Retain 64 Tokens (11.1%)			
	GQA	POPE	VizWiz	SEED	GQA	POPE	VizWiz	SEED
VisPruner Zhang et al. (2025)	53.12	77.26	52.41	51.88	55.40	80.40	53.30	53.70
+ REC. Chen et al. (2025)	53.78	78.10	52.95	52.36	56.05	81.25	53.95	54.40
+ RESTORE Cho et al. (2026)	58.34	84.26	54.38	55.82	59.20	85.10	54.85	56.50
+ CARE	**58.61**	**84.55**	**54.72**	**56.09**	**59.48**	**85.36**	**55.22**	**56.78**
OC-VTP Li et al. (2026)	55.12	81.34	55.91	56.73	55.70	83.00	56.30	57.90
+ REC. Chen et al. (2025)	55.51	81.78	56.22	57.08	56.18	83.42	56.68	58.21
+ RESTORE Cho et al. (2026)	55.73	82.04	56.45	57.31	56.46	83.66	56.92	58.44
+ CARE	**55.98**	**82.31**	**56.71**	**57.59**	**56.72**	**83.93**	**57.18**	**58.71**

Table 2: Backbone generalization and component ablation of CARE. (a) Generalization across different VLLM backbones on GQA. (b) Contributions of proposal components and counterfactual decision selection.

Method	Acc. ↑	#Dmg ↓	RIE ↓	RIR ↓	PG ↑
Qwen2.5-VL-7B					
REC.	59.67	381	1.72	12.44	19.4
RESTORE	59.83	289	1.30	12.34	19.9
CARE	**60.16**	**143**	**0.64**	**12.29**	**20.2**
LLaVA-1.5-13B					
REC.	65.34	318	1.43	12.45	20.8
RESTORE	65.61	241	1.08	12.34	21.2
CARE	**65.96**	**106**	**0.47**	**12.30**	**21.6**

(a) Backbone generalization on GQA.

Proposal Variant	Selection	#Dmg ↓	RIE ↓	RIR ↓	PG ↑	Acc. ↑
w/o EEDA	Proposal only	13	2.03	13.08	13.8	55.30
	+ CF Selection	8	1.25	12.74	15.6	55.60
w/o $G(e)$	Proposal only	11	1.72	12.96	14.5	55.10
	+ CF Selection	7	1.10	12.68	16.3	55.60
w/o Redundancy ρ_i	Proposal only	15	2.35	13.42	13.2	55.50
	+ CF Selection	11	1.72	13.01	15.1	55.60
Full Proposal	Proposal only	11	1.72	12.83	17.4	55.50
	+ CF Selection	**4**	**0.63**	**12.41**	**20.3**	**56.10**

(b) Ablation of proposal and counterfactual selection.

5.2 Performance Evaluation

Overall Performance. We first evaluate whether CARE improves post-pruning prediction quality across different pruning mechanisms and compression levels. As shown in Table 1, CARE achieves the highest task performance within each evaluated pruner–budget setting. The consistent improvements across VisPruner and OC-VTP indicate that CARE is not tied to a specific token-selection strategy.

Backbone Generalization. We further evaluate whether CARE generalizes across different VLLM backbones on GQA. Table 2a compares Qwen2.5-VL-7B Wu et al. (2025a) and LLaVA-1.5-13B Liu et al. (2024) under the same OC-VTP pruning strategy and 64-token budget. CARE consistently achieves higher GQA accuracy and PG while reducing #Dmg, RIE, and RIR across both backbones. These results demonstrate that the proposed recovery strategy generalizes across different VLLM architectures and model scales.

5.3 Component Analysis

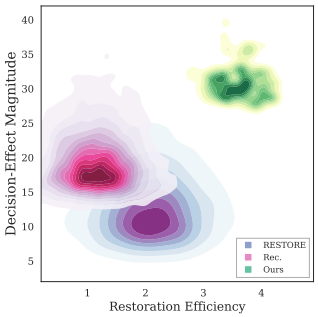
We ablate the major proposal components and counterfactual decision selection to examine their contributions to reliable post-pruning recovery. Under an experimental setting in which recovery is applied to 500 samples, Table 2b shows that removing EEDA, the exchange-gain criterion $G(e)$, or redundancy modeling ρ_i consistently degrades recovery reliability and evidence localization. Counterfactual selection further reduces recovery-induced errors and intervention magnitude across all proposal variants, while the full framework achieves the lowest RIE and the highest PG score. These results demonstrate the complementary roles of diverse evidence-aware proposal generation and decision-level intervention selection.

5.4 Further Analysis

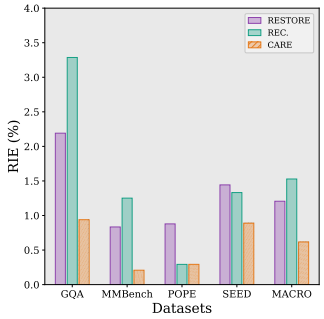
Recovery Reliability. We investigate the recovery–risk trade-off by examining whether post-pruning recovery damages predictions that are already correct. Under an experimental setting in which recovery is applied to 1,500 samples, Table 3 shows that CARE consistently achieves the lowest RIE across pruning budgets and benchmarks. This demonstrates that post-pruning recovery requires not only access to discarded evidence, but also explicit control over the decision risk introduced by unnecessary intervention on already reliable samples.

Table 3: Recovery reliability under different token budgets.

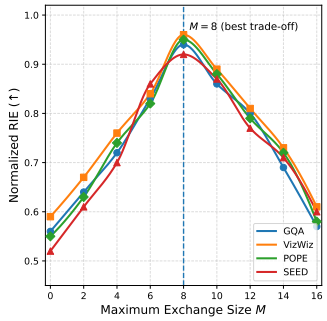
Method	GQA			VizWiz			POPE			SEED		
	#Dmg ↓	RIE ↓	RIR ↓	#Dmg ↓	RIE ↓	RIR ↓	#Dmg ↓	RIE ↓	RIR ↓	#Dmg ↓	RIE ↓	RIR ↓
Retain 32 Tokens (5.6%)												
Base Pruner	–	–	–	–	–	–	–	–	–	–	–	–
+ REC. Chen et al. (2025)	221	4.05	25.00	75	4.00	24.96	107	1.76	24.92	303	4.87	24.88
+ RESTORE Cho et al. (2026)	101	1.85	24.81	44	2.34	**24.37**	71	1.16	24.73	190	3.05	24.65
+ Ours	**81**	**1.48**	**24.76**	**24**	**1.28**	24.42	**52**	**0.85**	**24.69**	**118**	**1.90**	**24.58**
Retain 48 Tokens (8.3%)												
Base Pruner	–	–	–	–	–	–	–	–	–	–	–	–
+ REC. Chen et al. (2025)	154	2.64	16.67	35	1.79	16.62	58	0.89	16.58	202	2.96	16.54
+ RESTORE Cho et al. (2026)	128	2.20	16.49	19	0.97	**16.21**	43	0.66	16.45	146	2.14	16.40
+ Ours	**53**	**0.91**	**16.46**	**17**	**0.87**	16.27	**24**	**0.37**	**16.43**	**85**	**1.25**	**16.36**
Retain 64 Tokens (11.1%)												
Base Pruner	–	–	–	–	–	–	–	–	–	–	–	–
+ REC. Chen et al. (2025)	124	2.03	12.50	31	1.53	12.47	38	0.55	12.44	161	2.23	12.41
+ RESTORE Cho et al. (2026)	106	1.73	12.46	23	1.14	**12.19**	28	0.40	12.39	113	1.56	12.35
+ Ours	**47**	**0.77**	**12.43**	**8**	**0.40**	12.21	**13**	**0.19**	**12.37**	**60**	**0.83**	**12.32**



(a) Decision efficiency and impact of decision-effect.



(b) Recovery–risk trade-off under different recovery settings.



(c) Sensitivity to the maximum exchange size $M_{\max}$.

Figure 3: Decision-level behavior, recovery reliability, and exchange-capacity sensitivity. (a) Restoration efficiency versus decision impact. (b) RIE across benchmarks. (c) Sensitivity to the maximum exchange size $M_{\max}$.

Decision Behavior. We further analyze whether the proposed recovery mechanism produces selective interventions that meaningfully affect the current model decision. As illustrated in Fig. 3(a), CARE occupies the highest decision-impact region among the compared recovery methods. This indicates that query-adaptive exchange concentrates intervention on visual evidence that is more consequential to the current prediction.

Recovery Error Rate Analysis. To evaluate the error rates introduced by different recovery methods across multiple datasets, we compare their recovery-induced errors in Fig. 3(b). Our method consistently achieves the lowest error rate across all evaluated datasets, demonstrating its superior reliability during post-pruning recovery.

Sensitivity Analysis. We vary the maximum exchange size $M_{\max}$ to examine how the allowed intervention capacity affects recovery behavior. As shown in Fig. 3(c), performance follows a non-monotonic trend as the exchange range increases, with an intermediate capacity providing the most favorable behavior. This indicates that effective recovery requires sufficient flexibility to recover missing evidence while avoiding overly aggressive changes to the retained representation.

Selective and Query-Adaptive Recovery Behavior. We qualitatively examine CARE under different post-pruning states and query targets. As shown in Fig. 4 (a), conventional recovery can overturn an originally correct prediction, whereas CARE preserves the decision through selective token modification. In Fig. 4 (b), we construct a controlled correction setting with predefined ground-truth categories to compare recovery ability after pruning-induced errors. CARE restores more target-relevant evidence and yields stronger prediction recovery. The ground truth is used

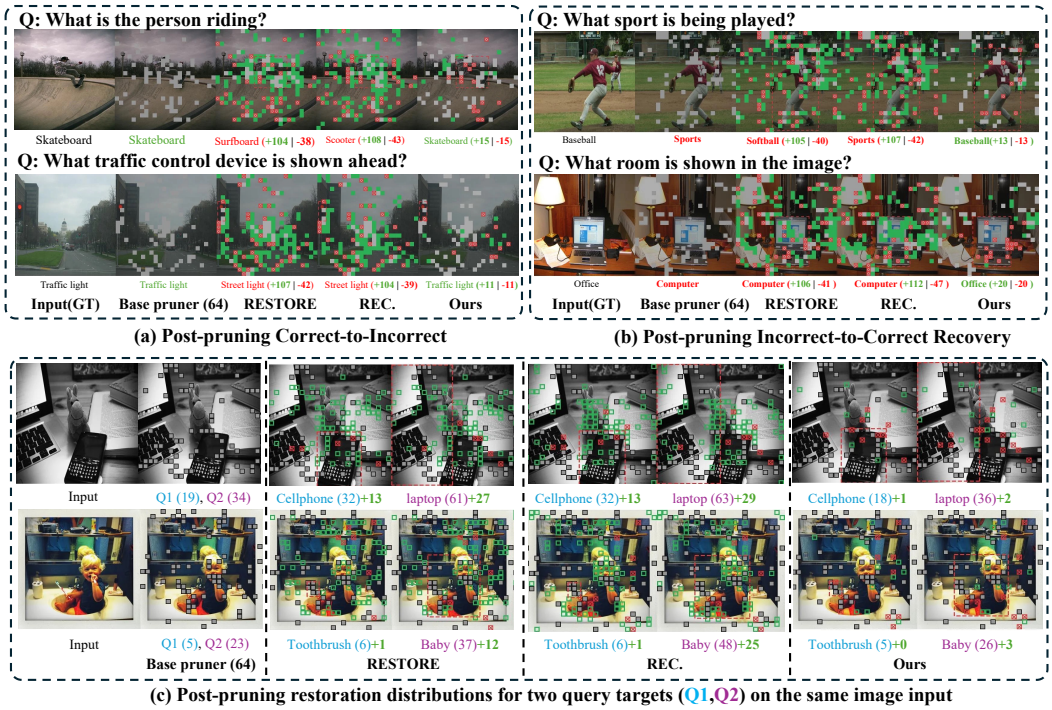


Figure 4: Qualitative comparison of post-pruning recovery and query-specific evidence recovery.

Table 4: Inference efficiency and evidence localization of recovery methods.

Method	Latency ↓ (ms)	Peak Mem. ↓ (MB)	Img/(ms)	Restore Img/(ms)	PG ↑ (%)
Full	129.12	23,221	–	–	–
Base Pruner Li et al. (2026)	105.43	15,113	**171.1**	–	21.0
+RESTORE Cho et al. (2026)	123.33	15,191	203.4	32.3	20.2
+REC. Chen et al. (2025)	136.50	15,880	205.8	34.7	19.7
+Ours	**105.16**	**15,114**	182.8	**11.7**	**20.3**

only for constructing and evaluating this experiment, not during normal inference. Fig. 4 (c) further shows that different queries produce distinct recovery distributions over corresponding visual regions, demonstrating query-adaptive token composition under the same visual-token budget.

Computational Efficiency. We further examine whether post-pruning recovery preserves the efficiency advantage of visual-token reduction while maintaining query-relevant visual evidence. As summarized in Table 4, CARE remains close to the base pruner in latency and peak memory while introducing substantially lower recovery overhead than competing methods. Its PG score also remains competitive with the base pruner and higher than the compared recovery methods, indicating effective localization of query-relevant evidence under fixed-budget intervention.

6 Conclusion and Limitations

We study post-pruning visual-token recovery as a recovery–risk trade-off, where restoring missing evidence strengthens affected decisions, while unnecessary intervention disrupts reliable predictions. We propose CARE, a training-free and fixed-budget framework that combines query-adaptive token exchange with counterfactual decision verification to separate recovery proposal from intervention adoption. Experiments across multiple pruners, benchmarks, and token budgets demonstrate consistent performance gains, fewer recovery-induced errors, and preserved pruning efficiency. A limitation is that counterfactual selection requires additional VLLM evaluations over candidate compositions, and our evaluation mainly focuses on single-image vision-language understanding. Future work will explore more efficient decision evaluation and broader multimodal settings.

AI USE STATEMENT

We used generative AI tools (ChatGPT) only for improving the grammar and wording of the manuscript. We did not use generative AI tools for developing the proposed method, designing or implementing experiments; all of these were carried out entirely by the authors. We reviewed every AI-suggested edit to ensure that the technical content remained unchanged, and we take full responsibility for all content in this paper.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings, including datasets, evaluation metrics, model configurations, pruning methods, token budgets, and implementation details. Appendix B further describes the complete counterfactual beam-search procedure and its complexity to facilitate reproduction of CARE and the reported results.

REFERENCES

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10044–10054, 2020.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Kaitong Cai, Yijia Fan, Jian Wang, Keze Wang, et al. Cf-vlm: Counterfactual vision-language fine-tuning. *Advances in Neural Information Processing Systems*, 38:58056–58094, 2026.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10797–10806. IEEE, 2020.

Yi Chen, Jian Xu, Xu-Yao Zhang, Wen-Zhuo Liu, Yang-Yang Liu, and Cheng-Lin Liu. Recoverable compression: A multimodal vision token recovery mechanism guided by text information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2293–2301, 2025.

Hyeonwoo Cho, Donghyeon Baek, Yewon Kim, and Bumsub Ham. Improving visual token reduction via rectifying distortions for efficient multimodal llm inference. *arXiv preprint arXiv:2606.01711*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5):2900–2919, 2023.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11975–11985. IEEE, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

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

Guangyuan Li, Rongzhen Zhao, Jinhong Deng, Yanbo Wang, and Joni Pajarinen. Object-centric vision token pruning for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7040–7049, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, 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*, pp. 292–305, 2023.

Yuxuan Liang, Fan Shi, Rui Zhu, Xu Li, Xiaolei Chen, Zhe Liu, Bin Li, and Xiangyang Xue. Envision, attend, then respond: Counterfactual hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18261–18272, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Junzhu Mao, Yang Shen, Jinyang Guo, Yazhou Yao, Xiansheng Hua, and Hengtao Shen. Prune and merge: Efficient token compression for vision transformer with spatial information preserved. *IEEE Trans. Multimedia*, 2025.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705. IEEE, 2021.

Pei Peng, Ming-Kun Xie, Hang Hao, Tong Jin, and Sheng-Jun Huang. Representation-level counterfactual calibration for debiased zero-shot recognition. *Advances in Neural Information Processing Systems*, 38:134547–134584, 2026.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Zixun Sun, Yubo Dong, Hehe Fan, and Yi Yang. Hi-lo prune: Look at what you'll lose before pruning with hierarchical token selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31941–31951, 2026.

Shuailong Wang, Xinyu Lyu, Shengming Yuan, Jingkuan Song, Heng Tao Shen, and Lianli Gao. Understanding and mitigating token-pruning-induced vulnerabilities in vlms. In *Forty-third International Conference on Machine Learning*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Yuanchen Wu, Lu Zhang, Hang Yao, Junlong Du, Ke Yan, Shouhong Ding, Yunsheng Wu, and Xiaoqiang Li. Antidote: A unified framework for mitigating lvlm hallucinations in counterfactual presupposition and object perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14646–14656. IEEE, 2025b.

Guangzhi Xiong, Sanchit Sinha, Zhenghao He, and Aidong Zhang. Retrieving counterfactuals improves visual in-context learning. *arXiv preprint arXiv:2603.16737*, 2026.

Sihan Yang, Runsen Xu, Chenhang Cui, Tai Wang, Dahua Lin, and Jiangmiao Pang. Vflowopt: A token pruning framework for lmms with visual information flow-guided optimization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23924–23934. IEEE, 2025.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867. IEEE, 2025.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Appendix

Table A: Recovery benefit and intervention utility under different visual-token budgets. #Rec denotes base-pruner errors recovered after recovery, RR is the recovery rate over base-wrong samples, Net Gain is defined as $\#Rec - \#Dmg$, and BRR denotes the benefit-risk ratio $\#Rec/\#Dmg$. Results are aggregated over GQA, VizWiz, POPE, and SEED.

Tokens	Method	#Rec ↑	RR ↑	#Dmg ↓	Net Gain ↑	BRR ↑
32	REC. Chen et al. (2025)	**780**	**31.20**	706	74	1.10
	RESTORE Cho et al. (2026)	700	28.00	406	294	1.72
	CARE	675	27.00	**275**	**400**	**2.45**
48	REC. Chen et al. (2025)	**620**	**29.52**	449	171	1.38
	RESTORE Cho et al. (2026)	585	27.86	336	249	1.74
	CARE	570	27.14	**179**	**391**	**3.18**
64	REC. Chen et al. (2025)	**520**	**28.89**	354	166	1.47
	RESTORE Cho et al. (2026)	495	27.50	270	225	1.83
	CARE	485	26.94	**128**	**357**	**3.79**

A Additional recovery Utility Analysis

Recovery Utility Analysis. Table A further evaluates the benefit–risk trade-off of post-pruning recovery. Although REC achieves a higher #Rec and RR, it also introduces substantially more damaged predictions. In contrast, CARE consistently achieves the lowest #Dmg and the highest Net Gain and BRR across all token budgets. In particular, at 64 tokens, CARE reduces #Dmg from 354 to 128 compared with REC, while increasing Net Gain from 166 to 357 and BRR from 1.47 to 3.79. These results show that CARE favors more selective interventions that retain substantial recovery benefit while avoiding a larger number of newly introduced errors.

The main paper evaluates recovery reliability by measuring how often post-pruning intervention disrupts predictions that are already correct. We further analyze recovery utility by jointly considering recovered predictions and newly introduced errors.

For a given visual-token budget K, let $\mathcal{B}_K$ denote the samples incorrectly predicted by the base pruner and $\mathcal{C}_K$ the samples correctly predicted before recovery. The recovered and damaged samples are defined as

$$\begin{aligned} \#\mathrm{Rec} &= \left|\{x \in \mathcal{B}_K \mid \hat{y}_{\mathrm{rest}}(x) = y(x)\}\right|, \\ \#\mathrm{Dmg} &= \left|\{x \in \mathcal{C}_K \mid \hat{y}_{\mathrm{rest}}(x) \neq y(x)\}\right|, \end{aligned} \tag{A}$$

where $\hat{y}_{\mathrm{rest}}(x)$ denotes the prediction after recovery and $y(x)$ denotes the ground-truth answer. The corresponding recovery rate is

$$\mathrm{RR} = \frac{\#\mathrm{Rec}}{|\mathcal{B}_K|} \times 100\%. \tag{B}$$

To jointly quantify recovery benefit and intervention risk, we additionally report Net Gain and the benefit-risk ratio (BRR):

$$\mathrm{Net\ Gain} = \#\mathrm{Rec} - \#\mathrm{Dmg}, \qquad \mathrm{BRR} = \frac{\#\mathrm{Rec}}{\#\mathrm{Dmg}}. \tag{C}$$

A larger Net Gain indicates that the number of recovered predictions exceeds the number of newly introduced errors by a larger margin, while a larger BRR indicates a more favorable balance between recovery benefit and risk.

This analysis complements task performance by separating the beneficial and harmful effects of recovery. A large number of recovered predictions alone does not imply reliable recovery when the intervention simultaneously introduces many new errors. Net Gain and BRR therefore characterize the utility of recovery from a benefit-risk perspective.

B Details of Counterfactual Beam Search

The first CARE component generates a non-overlapping exchange proposal set. The second component searches combinations of these exchanges and evaluates their effects using the frozen VLLM. Exhaustively evaluating all subsets grows exponentially with the number of proposals. We therefore employ a bounded beam search that preserves promising counterfactual compositions while retaining the original pruned representation as an explicit no-change option.

B.1 Proposal Ordering and Search Initialization

Let the exchange proposal set be

$$\mathcal{E} = \{e_m\}_{m\in\mathcal{I}}, \qquad e_m = (u_m \to w_m), \qquad u_m \in \mathcal{K}_0,\ w_m \in \mathcal{D}_0, \tag{D}$$

where $\mathcal{I}$ indexes the M non-overlapping exchanges generated by Diverse High-Recall Exchange Proposal. Before subset search, exchanges are processed in descending order of their proposal priority $b^s(e)$.

Let B denote the beam width. The search begins from the no-change state, and processing an exchange e_t expands every current beam state through an exclusion branch and an inclusion branch:

$$\mathbb{B}_{\text{init}} = \{\varnothing\}, \qquad \mathcal{C}_t = \bigcup_{\mathcal{A}\in\mathbb{B}} \{\mathcal{A}, \mathcal{A}\cup\{e_t\}\}. \tag{E}$$

Here, $\mathbb{B}$ denotes the beam before processing e_t, and $\mathcal{C}_t$ denotes the expanded candidate set. This include-or-exclude expansion enables the search to explore non-prefix combinations rather than following only a fixed progressive recovery sequence.

B.2 Fixed-Budget Counterfactual Composition

For any explored exchange subset $\mathcal{A}$, the corresponding visual-token composition is

$$\mathcal{K}(\mathcal{A}) = (\mathcal{K}_0 \setminus \{u_m \mid e_m \in \mathcal{A}\}) \cup \{w_m \mid e_m \in \mathcal{A}\}, \qquad \mathcal{K}(\varnothing) = \mathcal{K}_0. \tag{F}$$

The proposal stage prevents reuse of both retained and discarded tokens. Consequently, every selected exchange removes one distinct retained token and introduces one distinct discarded token.

Proposition 1. Fixed-budget preservation. Every counterfactual composition preserves the original visual-token budget.

Proof. For any subset $\mathcal{A}$, the number of removed tokens and inserted tokens are both equal to $|\mathcal{A}|$. Therefore,

$$|\mathcal{K}(\mathcal{A})| = |\mathcal{K}_0| - |\mathcal{A}| + |\mathcal{A}| = K. \tag{G}$$

Hence CARE evaluates alternative token compositions without changing the visual-token budget.

B.3 Evidence-Level Admissibility

The proposal stage retains each exchange satisfying $G(e) > \kappa$, where $\kappa < 0$ is a permissive lower bound. Individual exchanges with near-neutral or mildly negative query-evidence change therefore remain available during high-recall proposal generation.

For a non-empty subset, CARE measures its aggregate query-evidence change and retains only subsets with positive aggregate evidence gain:

$$R(\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{e_m \in \mathcal{A}} G(e_m), \qquad \mathcal{C}_t^+ = \{\mathcal{A} \in \mathcal{C}_t \setminus \{\varnothing\} \mid R(\mathcal{A}) > 0\}. \tag{H}$$

The individual proposal threshold and the subset-level evidence constraint therefore serve different purposes: the former maintains proposal recall, whereas the latter requires the complete composition to provide positive aggregate query-relevant evidence.

Proposition 2. Non-redundancy of the subset constraint. The condition $G(e) > \kappa$ with $\kappa < 0$ does not imply $R(\mathcal{A}) > 0$.

Proof. Consider exchanges whose individual evidence changes lie between the permissive lower bound and the neutral point. Each exchange satisfies the proposal condition, while a subset composed entirely of such exchanges has a negative average evidence change:

$$\kappa < G(e_m) < 0 \;\; \forall e_m \in \mathcal{A} \quad \Longrightarrow \quad G(e_m) > \kappa \;\text{ and }\; R(\mathcal{A}) < 0. \tag{I}$$

Thus, the subset-level condition removes combinations that pass the permissive individual proposal threshold but collectively reduce query-relevant evidence.

B.4 Counterfactual Decision Evaluation

The frozen VLLM evaluates each evidence-admissible token composition relative to the original pruned representation. Let $\hat{y}_0$ denote the pre-recovery prediction and $\hat{y}_{\mathcal{A}}$ the prediction obtained from the counterfactual composition. Their model-internal support is measured using length-normalized log-probability:

$$\begin{aligned} \hat{y}_0 &= f_\theta(\mathcal{K}_0, q), & \hat{y}_{\mathcal{A}} &= f_\theta(\mathcal{K}(\mathcal{A}), q), \\ \ell(y; \mathcal{K}) &= \frac{1}{|y|} \sum_t \log p_\theta\left(y_t \mid y_{<t}, q, \mathcal{K}\right), & \Delta(\mathcal{A}) &= \ell(\hat{y}_{\mathcal{A}}; \mathcal{K}(\mathcal{A})) - \ell(\hat{y}_0; \mathcal{K}_0). \end{aligned} \tag{J}$$

The parameters θ remain frozen throughout the search. A positive $\Delta(\mathcal{A})$ indicates that the prediction induced by the alternative composition receives stronger model-internal support than the original pre-recovery decision. This score evaluates the effect of the complete token intervention rather than relying only on token-level relevance.

B.5 Beam Update and Candidate Preservation

After counterfactual evaluation, the search retains the B admissible non-empty subsets with the largest decision-support scores. The no-change state is explicitly preserved and does not occupy one of the non-empty beam positions. We additionally maintain an archive $\mathcal{H}$ of all evaluated admissible subsets:

$$\mathbb{B} \leftarrow \{\varnothing\} \cup \mathrm{TopB}_\Delta\left(\mathcal{C}_t^+\right), \qquad \mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{C}_t^+. \tag{K}$$

Here, $\mathrm{TopB}_\Delta(\cdot)$ returns at most B subsets with the largest $\Delta(\mathcal{A})$. The archive prevents a previously evaluated high-quality composition from being discarded permanently when later exchange proposals are processed.

B.6 Conservative Final Adoption

After all exchange proposals have been processed, the feasible set consists of the evidence-admissible compositions stored during beam search. CARE adopts the alternative with the strongest positive counterfactual support; otherwise, it returns to the original pruned representation:

$$\begin{aligned} \mathcal{S} &= \{\mathcal{A} \in \mathcal{H} \mid R(\mathcal{A}) > 0\}, \\ \mathcal{A}^* &= \begin{cases} \arg\max_{\mathcal{A}\in\mathcal{S}} \Delta(\mathcal{A}), & \mathcal{S} \neq \varnothing \text{ and } \max_{\mathcal{A}\in\mathcal{S}} \Delta(\mathcal{A}) > 0, \\ \varnothing, & \text{otherwise}, \end{cases} \\ \mathcal{K}^* &= \mathcal{K}(\mathcal{A}^*). \end{aligned} \tag{L}$$

This formulation also explicitly handles the case in which no admissible non-empty subset survives the evidence constraint.

Proposition 3. No-change fallback. If no feasible alternative receives positive counterfactual decision support, CARE exactly retains the original pruned composition.

Proof. The conservative adoption rule directly yields

$$\max_{\mathcal{A}\in\mathcal{S}} \Delta(\mathcal{A}) \leq 0 \quad \Longrightarrow \quad \mathcal{A}^* = \varnothing \quad \Longrightarrow \quad \mathcal{K}^* = \mathcal{K}_0. \tag{M}$$

Therefore, recovery is not mandatory: the original pruned representation remains the final output whenever the explored alternatives fail to provide stronger decision support.

B.7 Search Complexity

A complete subset enumeration requires a number of candidate compositions that grows exponentially with M. Beam search instead maintains at most B non-empty states together with the no-change state at each proposal step. Therefore, the number of VLLM-evaluated compositions grows linearly with both the number of proposals and the beam width.

Let C_{VLLM} denote the computational cost of evaluating one counterfactual composition with the frozen VLLM. The resulting search complexity is

$$N_{\mathrm{eval}} = \mathcal{O}(MB), \qquad C_{\mathrm{search}} = \mathcal{O}\left(MB\, C_{\mathrm{VLLM}}\right). \tag{N}$$

Thus, CARE replaces exhaustive subset enumeration with a bounded search over decision-relevant compositions while preserving the same visual-token budget throughout all evaluated interventions.

B.8 Summary of the Search Procedure

The counterfactual beam search separates three roles in the recovery process. The proposal priority $b^s(e)$ determines the order in which individual exchanges enter the search. The aggregate evidence score $R(\mathcal{A})$ removes subsets that collectively reduce query-relevant evidence. Finally, $\Delta(\mathcal{A})$ evaluates the actual effect of each remaining token composition on the frozen VLLM decision. The original pruned representation remains available throughout the search and is recovered exactly whenever no alternative obtains stronger decision support.