CALIBRATE BEFORE REASON: ROBUST VISUAL TOKEN REDUCTION AGAINST SEMANTIC DRIFT IN VLMS

Anonymous authors
Paper under double-blind review

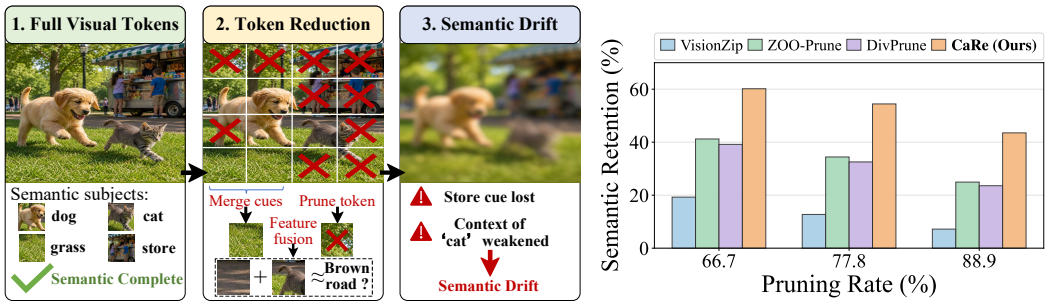

Figure 1: Illustration and comparison of semantic drift in visual token reduction methods.

ABSTRACT

Large Vision-Language Models (VLMs) suffer from prohibitive inference overhead due to long sequences of visual tokens. However, existing visual token reduction methods mainly improve efficiency by pruning or compressing redundant tokens without examining whether the resulting representation remains semantically consistent with the original representation. Mapping the original N-token visual sequence to K tokens may discard, dilute, or misassign critical visual cues, triggering severe **semantic drift** that deviates the VLM's understanding. In this paper, we first introduce the principle of "**Calibrate Before Reason**" to visual token reduction and propose **CaRe**, a training-free robust framework that calibrates compact visual representations before reasoning to preserve semantic fidelity in VLMs. CaRe consists of two mutually complementary modules: 1) Perturbation-Robust Calibration Anchoring, which identifies calibration anchors with stable model-side influence under multi-directional perturbations; 2) Confidence-Gated Token Calibration, which extracts reliable calibration signals from unselected tokens and injects them into anchors. Extensive evaluations across diverse VLM architectures and benchmarks verify that CaRe outperforms state-of-the-art token reduction baselines. While pruning **94.4%** of visual tokens, our method retains **96.4%** of the original full-token performance, delivering up to **2.30**$\times$ faster end-to-end inference speed relative to unpruned vanilla models.

1 INTRODUCTION

Large Vision Language Models (VLMs) Zhu et al. (2025); Bai et al. (2023) have rapidly evolved from projector-based architectures to high-resolution and unified image–video systems, such as LLaVA-NeXT Liu et al. (2024b), InternVL Chen et al. (2024b), and Qwen-VL Wang et al. (2024). While dense visual tokens improve perception, they also increase attention cost, KV-cache memory, and inference latency Yang et al. (2025a); Zhang et al. (2024a), making visual-token reduction essential for efficient VLM inference.

Recent visual token reduction methods follow two routes Lee et al. (2026): pruning tokens according to task- or model-specific importance scores Kim et al. (2026); Gao et al. (2026); Zhang et al. (2026a) and merging or compressing redundant tokens into compact representations Han et al.

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.

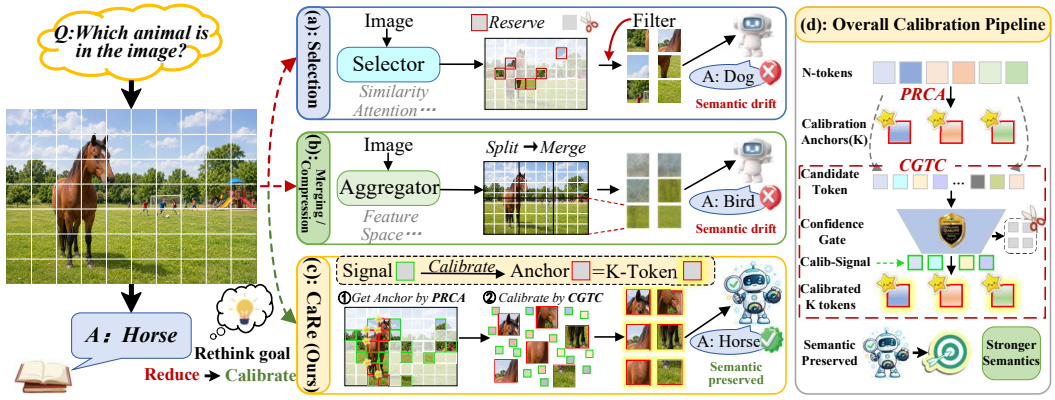

Figure 2: From Visual Token Reduction to Visual Token Calibration. Conventional token reduction methods rely on (a) token selection or (b) merging/compression, which cause semantic drift by losing or diluting critical visual cues. (c) CaRe introduces the principle of "Calibrate Before Reason" to visual token reduction to better preserve visual semantics. (d) Our method selects robust anchors and utilizes a confidence gate to inject reliable calibration signals, thereby producing K semantically calibrated anchors that better preserve task-relevant visual information.

(2026); Yang et al. (2025b); Zhang et al. (2026b). Although effective for VLM efficiency, both routes remain token-centric and overlook whether the reduced representation preserves the semantics of the full visual input. We define this failure as **semantic drift**. As shown in the upper panel of Figure 1 and Figure 2 (a) and (b), token selection may discard essential cues, whereas compression may entangle redundant information within retained tokens, weakening their semantic specificity.

Prior studies show that pruning inevitably loses part of the original visual information, while visual tokens remain critical to language reasoning Zhang et al. (2025e); Yang et al. (2025b); Hu et al. (2026); Huang et al. (2026). Accordingly, token reduction should not end with fewer tokens; the compact representation should be calibrated before reasoning to preserve semantic fidelity and mitigate semantic drift. Therefore, we introduce the principle of "**Calibrate Before Reason**" to visual token reduction and propose **CaRe**, a training-free robust visual token reduction framework. The lower panel of Figure 1 shows that CaRe outperforms selection and compression baselines in semantic retention, which measures preserved task-relevant semantics; higher values indicate less semantic drift. See Appendix A for details of calculation and analysis.

As shown in Figure 2 (c) and (d), CaRe proceeds in two core stages. In the first stage, the **Perturbation-Robust Calibration Anchoring** module applies multi-directional perturbations before the projector and evaluates the corresponding perturbation responses to estimate the influence of each token on the visual mapping on the model-side inspired by recent studies Kim et al. (2026); Sogi et al. (2024). We further introduce a directional risk correction mechanism to select anchors with stable model-side influence from projector-response sensitivity, and incorporate a diversity constraint to prevent the anchors from over-concentrating on the single salient region.

Second, the **Confidence-Gated Token Calibration** module determines how unselected tokens contribute to calibration. The confidence gate evaluates candidate signals through semantic-spatial affinity, best-anchor matching quality, and assignment concentration, suppressing residual tokens with ambiguous attribution before any information is transferred. Qualified signals are softly assigned to matched anchors without increasing the token budget. Thus, CaRe uses reliable calibration signals to enhance and positively recalibrate anchor semantics.

Our main highlights are as follows:

- Building on prior work on sensitivity-based token selection and information recycling, we address two specific questions: how to reduce variation in anchor scores across perturbation directions, and which residual tokens to admit for anchor updates.
- We propose **Perturbation-Robust Calibration Anchoring**, which identifies stable, diverse calibration anchors by measuring robust model-side influence.

- We design **Confidence-Gated Token Calibration**, which extracts reliable calibration signals from unselected tokens and injects them into anchors to enrich and calibrate the retained anchor representations.
- Extensive experiments across multiple VLM architectures and benchmarks show that CaRe outperforms existing token reduction methods across the considered token-reduction settings and budgets, preserving **96.4%** of full-token performance while pruning **94.4%** of visual tokens and achieving up to **2.30**× end-to-end speedup.

2 Related Work

2.1 Vision-Language Models

Large Vision-Language Models (VLMs) Zhang et al. (2026c); Ji et al. (2025); Dai et al. (2023); Wang et al. (2026a) have achieved remarkable progress in multimodal perception and reasoning by aligning visual encoders with large language models. Early representative systems, such as LLaVA Liu et al. (2023), connect a pre-trained vision encoder to an instruction-tuned language model through a lightweight projector, establishing a simple yet effective paradigm for visual instruction tuning. This architecture has been further extended by recent VLM families, including LLaVA-NeXT Liu et al. (2024b), InternVL Chen et al. (2024b), and Qwen-VL Wang et al. (2024), which improve visual understanding through stronger vision backbones, higher-resolution inputs, and more powerful multimodal alignment Zhang et al. (2025d); Dong et al. (2026). Despite their effectiveness, these models usually represent each image with hundreds or even thousands of visual tokens Zhang et al. (2024b), leading to substantial computation and memory costs during inference.

2.2 Visual Token Reduction

Visual token reduction has become an important direction for efficient VLM inference, aiming to reduce the number of visual tokens fed into the language model. Existing methods can be broadly categorized into two paradigms: token selection and token merging/compression Lee et al. (2026). The first paradigm estimates token importance and selects informative tokens based on attention scores, token similarity, diversity, region relevance, or adaptive strategies Kim et al. (2026); Wang et al. (2026b). FastV Chen et al. (2024a) and FasterVLM Zhang et al. (2025b) exploit attention statistics for token pruning, SparseVLM Zhang et al. (2025e) introduces text-guided attention for sparsification, and MADTP Cao et al. (2024) leverages multimodal alignment-guided dynamic pruning. HiRED Arif et al. (2025), PyramidDrop Xing et al. (2025), and DivPrune Alvar et al. (2025) further explore hierarchical redundancy, diversity preservation, and dynamic selection. The second paradigm reduces tokens through merging or compression by aggregating similar tokens or transforming local visual features into compact representations Wu et al. (2026); Tong et al. (2025); Fu et al. (2025). Representative methods include LLaVA-PruMerge Shang et al. (2025), VisionZip Yang et al. (2025b), TokenPacker Li et al. (2025), and LLaVA-Mini Zhang et al. (2025c), which explore adaptive merging, dominant token preservation, visual projection and modality fusion.

Despite their effectiveness, existing methods remain reduction-oriented: they obtain compact visual representations through selection, merging, or compression Han et al. (2026). Critical visual cues may still be discarded, diluted, or incorrectly fused, leading to semantic drift. To address this issue, we introduce the principle of "Calibrate Before Reason" and propose CaRe, a robust training-free framework that reformulates visual token reduction as visual token calibration. Unlike existing methods that directly remove or compress visual tokens, CaRe reformulates token reduction around stable anchor construction, confidence-qualified signal admission, and the calibration objective. It identifies perturbation-robust anchors, admits reliably attributable signals, and injects them to recalibrate anchor semantics, thereby mitigating semantic drift under aggressive token reduction.

3 Methodology

In this section, we present CaRe, a training-free framework for fixed-budget visual token calibration in VLMs. As illustrated in Figure 3, CaRe constructs perturbation-robust calibration anchors, then selects reliable calibration signals from unselected tokens, and finally injects them into the anchors

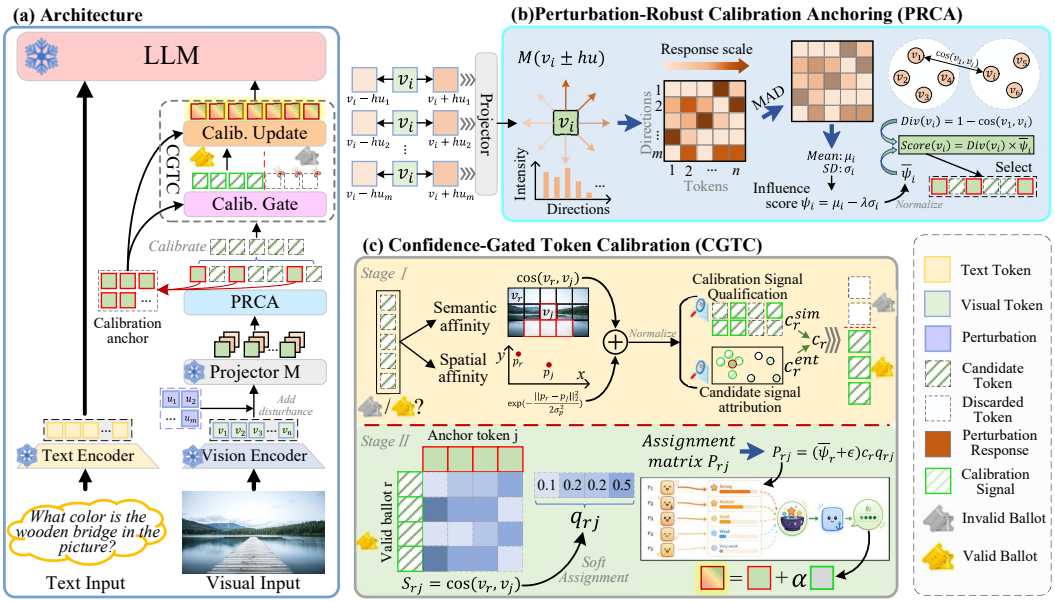

Figure 3: Overview of the proposed CaRe framework. (a) The overall architecture illustrates the complete token calibration pipeline, from anchor selection to final token calibration, and the relationship between PRCA and CGTC. (b) PRCA selects robust calibration anchors by measuring multi-directional perturbation responses and diversity. (c) CGTC contains two stages: Stage I qualifies reliable calibration signals via confidence gating, and Stage II softly assigns and injects calibration signals into anchors to obtain calibrated visual tokens.

to obtain calibrated visual tokens. Here, "robust" refers specifically to score stability across sampled perturbation directions.

3.1 Preliminary and Problem Formulation

Given an input image, the vision encoder produces a sequence of visual tokens $V = [v_1, \ldots, v_N]^\top \in R^{N \times d_v}$, where N is the number of visual tokens and d_v is the token dimension. For a fixed retention ratio ρ, the target budget is $K = \max(1, \lfloor \rho N \rfloor), K \geq 2$. Instead of directly selecting or compressing the original sequence into K tokens, our goal is to generate a calibrated visual representation $\hat{V}_S = \{\hat{v}_j \mid j \in S\}$, where $S \subset \{1, \ldots, N\}$ and cardinality $|S| = K$. The calibrated tokens preserve usable semantics from the full visual input while satisfying the same ratio ρ.

Prior sensitivity-based pruning studies Kim et al. (2026); Sogi et al. (2024); Huang et al. (2026) suggest that perturbing visual tokens before the multimodal projector reflects their influence on model-side visual mappings. Following this observation, we utilize the projector $M(\cdot)$ as a lightweight response function. CaRe needs to address two key questions: which visual tokens serve as stable semantic reference points, and how unselected visual tokens should contribute to calibration. These questions are addressed by Perturbation-Robust Calibration Anchoring (PRCA) and Confidence-Gated Token Calibration (CGTC), respectively.

3.2 Perturbation-Robust Calibration Anchoring

As shown in Figure 3(b), for visual tokens V, we sample the same m unit perturbation directions $\{u_x\}_{x=1}^m$, obtained by sampling $g_x \sim \mathcal{N}(0, I_{d_v})$ and normalizing it as $u_x = g_x / \|g_x\|_2$, and apply symmetric perturbations before the projector with step size h. The response of token v_i along direction x is computed as:

$$C_{x,i} = \frac{\|M(v_i + hu_x) - M(v_i - hu_x)\|_2}{2h}, \tag{1}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm.

This response measures how strongly the model-side visual representation changes when a token is locally perturbed. However, different perturbation directions may have different response scales. Directly averaging raw responses may overemphasize a few high-scale directions. Therefore, we perform direction-wise robust normalization with the median and median absolute deviation:

$$Z_{x,i} = \frac{C_{x,i} - \mathrm{med}_j(C_{x,j})}{\mathrm{med}_j|C_{x,j} - \mathrm{med}_q(C_{x,q})| + \epsilon}, \tag{2}$$

where ϵ is a small positive constant, and $\mathrm{med}_j(\cdot)$ denotes the median taken over index j.

For each token, we compute the mean response $\mu_i = \frac{1}{m}\sum_x Z_{x,i}$ and the directional standard deviation $\sigma_i = \left(\frac{1}{m}\sum_x (Z_{x,i} - \mu_i)^2\right)^{1/2}$. If token with strong responses in only a few perturbation directions may be an unstable anchor. We therefore define a perturbation-robust influence score as $\psi_i = \mu_i - \lambda\sigma_i$, where λ controls the penalty for directional risk. Larger ψ_i indicates that the token consistently influences the model-side mapping across perturbation directions. We then normalize ψ_i to $\bar{\psi}_i \in [0, 1]$ before anchor selection.

Selecting anchors solely by $\bar{\psi}_i$ may lead to many anchors concentrated in the same salient object or local region. To avoid this, PRCA further incorporates diversity. We define the visual distance between two tokens as $D(v_i, v_j) = 1 - \cos(v_i, v_j)$. The anchor set is initialized with the token of the largest $\bar{\psi}_i$. At step t, for each unselected token v_i, we compute its marginal diversity to the current anchor set as $\delta_i^{(t)} = \min_{j \in S_t} D(v_i, v_j)$, and select the next anchor by

$$i_{t+1} = \arg\max_{i \notin S_t} \bar{\psi}_i \delta_i^{(t)}, \quad S_{t+1} = S_t \cup \{i_{t+1}\}. \tag{3}$$

This process is repeated until $|S| = K$. The selected set S forms the calibration anchors, and the remaining tokens $R = \{i \in \{1, \ldots, N\} \mid i \notin S\}$ are treated as candidate tokens for subsequent calibration, subject to confidence-based screening and selective use.

3.3 Confidence-Gated Calibration Signals

CGTC determines which unselected tokens provide reliable calibration signals. Not every residual token should be fused into anchors: some may correspond to noisy background, or regions that cannot be confidently assigned. Therefore, the first stage of CGTC qualifies candidate calibration signals through a confidence gate, as shown in Figure 3(c).

For each residual token $\{v_r | r \in R\}$ and anchor token $\{v_j | j \in S\}$, we compute a semantic-spatial affinity:

$$A_{rj} = \cos(v_r, v_j) + \eta \exp\left(-\frac{\|p_r - p_j\|_2^2}{2\sigma_p^2}\right), \tag{4}$$

where p_r and p_j are the 2D patch coordinates of tokens v_r and v_j, η controls the spatial term, σ_p controls the spatial decay. The cosine term measures visual semantic similarity, while the spatial term encourages locally coherent assignments.

For each residual token, we take its top-k_f ($k_f = \min(K, \max(2, \lceil 0.1K \rceil))$) anchors according to A_{rj} and compute a normalized assignment distribution π_r with temperature τ_g. Then evaluate two confidence scores. The first is the best-anchor matching confidence, defined by $c_r^{\mathrm{sim}} = \mathrm{sigmoid}\left(\frac{A_r^{(1)} - \theta_s}{\tau_c}\right)$, where $A_r^{(1)}$ is the maximum affinity, θ_s is affinity threshold. The second is the assignment concentration confidence, which penalizes ambiguous assignments:

$$c_r^{\mathrm{ent}} = 1 + \frac{\sum_{\ell=1}^{k_f} \pi_{r\ell} \log \pi_{r\ell}}{\log k_f}. \tag{5}$$

If a residual token is evenly distributed over multiple anchors, its entropy is high and c_r^{ent} becomes small. We combine these two views as:

$$c_r = c_r^{\mathrm{sim}} \frac{1 + c_r^{\mathrm{ent}}}{2}, \quad R^+ = \{r \in R \mid c_r \geq \theta_c\}, \tag{6}$$

where θ_c is confidence threshold. Only tokens in R^+ are regarded as valid calibration signals. Their source weights are defined as $w_r = (\bar{\psi}_r + \epsilon)c_r$, where $\bar{\psi}_r$ measures model-side influence and c_r measures attribution reliability.

Table 1: Performance comparison on **LLaVA-1.5-7B**.

Method	TFLOPs	GQA	MMB	MME	POPE	SQA	VQA^{v2}	VQA^{T}	MMMU	SEED	Avg.
Total 576 Tokens											
LLaVA-1.5-7B	8.54	61.90	64.70	1862.00	85.90	69.50	78.50	58.20	36.30	58.60	100%
Retain 192 Tokens ↓ 66.7%											
FastV	3.44	52.70	61.20	1612.00	64.80	67.30	67.10	52.50	34.30	57.10	89.6%
SparseVLM	3.46	57.60	62.50	1721.00	83.60	69.10	75.60	56.10	33.80	55.80	95.5%
VisionZip	3.93	59.30	63.00	1782.00	85.30	68.90	76.80	57.30	36.60	56.40	97.9%
DivPrune	-	59.97	62.54	1762.30	87.00	68.91	76.87	56.97	35.44	58.71	98.0%
FiCoCo-V	3.20	58.51	61.74	1755.61	81.47	67.42	74.22	55.50	36.11	57.95	96.0%
V^{2}Drop	3.42	58.60	62.37	1789.76	85.10	69.30	74.92	55.62	35.87	56.74	97.0%
ZOO-Prune	3.98	60.03	62.57	1781.66	87.24	69.01	75.35	57.10	35.44	58.78	98.0%
***CaRe* (Ours)**	3.47	**60.57**	**63.41**	**1809.47**	**88.94**	**70.24**	**77.42**	**58.01**	**37.33**	**58.94**	**99.9%**
Retain 128 Tokens ↓ 77.8%											
FastV	2.62	49.60	56.10	1490.00	59.60	61.80	61.80	50.60	34.90	55.90	84.7%
SparseVLM	2.90	56.00	60.00	1696.00	80.50	67.10	73.80	54.90	33.80	53.40	93.0%
VisionZip	3.15	57.60	62.00	**1761.70**	83.20	68.90	75.60	56.80	**37.90**	54.90	96.8%
DivPrune	-	59.25	62.03	1718.22	86.72	68.96	75.96	56.06	35.56	56.98	96.9%
FiCoCo-V	2.79	57.59	60.34	1718.06	82.24	68.73	72.97	54.10	35.44	54.98	94.5%
V^{2}Drop	2.59	56.19	59.28	1712.59	80.90	68.83	73.11	53.80	35.12	55.82	93.9%
ZOO-Prune	3.13	59.43	61.86	1741.23	87.13	68.86	74.55	56.79	35.67	57.39	97.2%
***CaRe* (Ours)**	3.02	**60.15**	**62.11**	1731.46	**87.79**	**69.89**	**76.73**	**57.96**	36.22	**57.73**	**98.3%**
Retain 64 Tokens ↓ 88.9%											
FastV	1.98	46.10	48.00	1256.00	48.00	51.10	55.00	47.80	34.00	51.90	75.5%
SparseVLM	2.11	52.70	56.20	1505.00	75.10	62.20	68.20	51.80	32.70	51.10	87.0%
VisionZip	2.25	55.10	60.10	**1690.00**	77.00	69.00	72.40	55.50	35.20	52.20	92.8%
DivPrune	-	57.78	59.28	1674.40	85.56	68.17	74.11	54.69	35.56	55.13	94.8%
FiCoCo-V	1.68	53.43	55.09	1653.67	75.29	67.92	71.63	51.88	34.22	53.14	90.1%
V^{2}Drop	1.76	53.52	55.21	1641.23	75.14	68.92	71.03	51.83	34.29	53.11	90.1%
ZOO-Prune	2.29	58.36	60.22	1626.80	85.56	68.17	72.97	55.35	35.21	**55.84**	94.8%
***CaRe* (Ours)**	2.18	**59.12**	**60.58**	1664.51	**86.98**	**69.80**	**75.81**	**56.32**	**36.11**	55.78	**96.5%**

3.4 Anchor Calibration with Signals

The semantic–spatial affinity screens residual tokens for admission. Accepted tokens are distributed over all anchors using cosine-based soft assignment. Thus, the gate controls which source tokens contribute without imposing a hard spatial constraint on their destinations. CGTC's second stage injects these qualified signals into anchors while keeping the final token budget unchanged. For each valid residual token $\{v_r | r \in R^+\}$, we compute its soft assignment to calibration anchors by:

$$q_{rj} = \frac{\exp(\cos(v_r, v_j)/\tau_s)}{\sum_{l \in S} \exp(\cos(v_r, v_l)/\tau_s)}, \quad j \in S. \tag{7}$$

The assignment matrix is then weighted by the source reliability, $P_{rj} = w_r q_{rj}$. For each anchor token v_j, the aggregated calibration signal is computed as a normalized weighted average:

$$g_j = \frac{\sum_{r \in R^+} P_{rj} v_r}{\sum_{r \in R^+} P_{rj} + \epsilon}. \tag{8}$$

Finally, the anchor representation is updated by $v'_j = v_j + \alpha g_j$, where α controls the calibration strength. To reduce distribution drift at the model interface, we preserve the original anchor norm:

$$\hat{v}_j = \frac{\|v_j\|_2}{\|v'_j\|_2 + \epsilon} v'_j. \tag{9}$$

The resulting $\hat{V}_S = \{\hat{v}_j \mid j \in S\}$ contains exactly K calibrated visual tokens. These calibrated tokens are then combined with text tokens and fed into the LLM for subsequent reasoning. In this way, CaRe maintains a fixed visual-token budget while leveraging informative signals from unselected tokens to enhance and positively recalibrate anchors.

4 Experiments

We evaluate CaRe on LLaVA-1.5-7B Liu et al. (2024a), LLaVA-NeXT-7B Liu et al. (2024b) across nine benchmarks, and in eight benchmarks for Qwen2.5-VL-7B Bai et al. (2025). LLaVA-1.5-7B encodes images into 576 tokens. LLaVA-NeXT represents each image with 2,880 tokens, and

Table 2: Performance comparison on **LLaVA-NeXT-7B**. Details are provided in Appendix C.4

Method	GQA	MMB	MME	POPE	SQA	VQA^{v2}	VQA^{T}	MMMU	SEED	Avg.
Total 2880 Tokens										
LLaVA-NeXT-7B	64.20	67.90	1842.00	86.40	70.20	80.10	61.30	35.10	70.20	100%
Retain 640 Tokens ↓ 77.8%										
VisionZip	61.30	**66.30**	1787.00	86.30	68.10	**79.10**	**60.20**	34.70	66.70	97.5%
FiCoCo-V	60.40	64.18	1780.22	86.04	67.94	75.77	54.89	35.91	65.77	95.8%
ZOO-Prune	62.19	65.10	1802.47	86.75	68.02	78.13	57.98	36.82	67.59	97.9%
***CaRe* (Ours)**	**63.87**	65.98	**1831.14**	**88.19**	**69.41**	78.02	58.39	**37.89**	**67.92**	**99.4%**
Retain 320 Tokens ↓ 88.9%										
VisionZip	59.30	63.10	1702.00	82.10	67.30	76.20	**58.90**	35.30	63.40	94.5%
FiCoCo-V	59.26	63.30	1721.74	85.18	67.51	74.08	54.67	34.33	64.51	93.9%
ZOO-Prune	60.83	64.60	1787.68	85.47	67.08	76.04	57.26	36.78	66.18	96.5%
***CaRe* (Ours)**	**61.98**	**64.92**	**1791.90**	**87.58**	**68.86**	**77.76**	57.67	**37.33**	**66.28**	**97.9%**
Retain 160 Tokens ↓ 94.4%										
VisionZip	55.50	60.10	1630.00	74.80	68.30	71.40	**56.20**	36.10	58.30	90.4%
FiCoCo-V	58.11	60.32	1657.81	84.27	68.01	73.22	55.10	33.98	63.06	92.4%
ZOO-Prune	59.48	64.18	1706.08	83.05	67.72	74.07	55.42	**36.67**	64.04	94.5%
***CaRe* (Ours)**	**59.97**	**64.47**	**1706.96**	**86.77**	**69.70**	**75.81**	55.77	36.56	**67.87**	**96.4%**

Qwen2.5-VL adopts a dynamic-resolution ViT encoder. We assess CaRe and competing methods by LMMs-Eval Zhang et al. (2025a). All experiments are conducted on a single NVIDIA GeForce RTX 5090 D v2 GPU with $h = 5 \times 10^{-4}$, $m = 64$, $\lambda = 0.5$, $\eta = 0.45$, $\theta_s = 0.4$, $\theta_c = 0.9$, and $\alpha = 0.15$. The compared methods span two categories of token reduction. Token selection methods: FastV Chen et al. (2024a), SparseVLM Zhang et al. (2025e), DivPrune Alvar et al. (2025), V^{2}Drop Chen et al. (2026) and ZOO-Prune Kim et al. (2026); and token compression methods: VisionZip Yang et al. (2025b) and FiCoCo Han et al. (2026). Details are provided in Appendix C.1 & C.2.

4.1 Comparison on Diverse Tasks

Results on LLaVA-1.5-7B. As shown in Table 1, our CaRe consistently outperforms existing selection- and compression-based pruning methods. With only 192 tokens, it retains 99.9% of the unpruned performance. And it surpasses ZOO-Prune Kim et al. (2026) by 1.9, 1.1, and 1.7 percentage points at three pruning rates, while maintaining low TFLOPs. Further analysis and TFLOPs details of calculation are provided in Appendix B & C.4.

Results on LLaVA-NeXT-7B. As shown in Table 2, we further evaluate CaRe on LLaVA-NeXT-7B for high-resolution images. Even at 94.4% pruning rate, it preserves 96.4% performance and surpasses the second-best method by 1.5, 1.4, and 1.9 percentage points across three pruning rates, demonstrating the superiority of token calibration over conventional token reduction methods.

Results on Qwen2.5-VL-7B. To verify the generalization capability of CaRe, we further evaluate it on the new Qwen-series model, Qwen2.5-VL-7B, as shown in Table 3. With only 20% and 10% of the tokens retained, it achieves 96.3% and 89.3% performance retention, respectively, consistently outperforming Comparison methods Kim et al. (2026); Han et al. (2026). This further shows superior performance over conventional token reduction methods with long, variable-length token sequences. We present partial results here; the complete results are available in Appendix C.4.

Table 3: Performance comparison on **Qwen2.5-VL-7B**.

Method	GQA	MMB	MME	POPE	Avg.-4
Full Tokens					
Qwen2.5-VL-7B	60.84	82.82	2310	86.30	100%
Retain 20% Tokens ↓ 80.0%					
FiCoCO-V	55.28	76.47	2120	81.29	92.3%
ZOO-Prune	55.92	77.41	2139	81.71	93.2%
***CaRe* (Ours)**	**58.11**	**77.91**	**2274**	**83.94**	**96.3%**
Retain 10% Tokens ↓ 90.0%					
FiCoCO-V	52.79	71.95	1967	77.61	87.2%
ZOO-Prune	53.28	72.51	1962	78.46	87.7%
***CaRe* (Ours)**	**54.45**	**73.37**	**2017**	**79.23**	**89.3%**

Unlike ZOO-Prune Kim et al. (2026), which ranks tokens by zeroth-order sensitivity and diversity, CaRe aggregates multi-directional responses into robust calibration anchors. Unlike FiCoCO-V's Han et al. (2026) correlation-driven compression, CGTC filters residual signals through confidence gating before softly injecting them into selected anchors. Together with module ablations (Table 4 and Appendix D.1), gains across three backbones and high-pruning regimes support semantic calibration as a distinct mechanism beyond combining selection and compression.

Table 4: Module ablation on LLaVA-1.5-7B. Further details are provided in Appendix D.1.

PRCA	CGTC	GQA	MMB	MME	POPE	SQA	VQA^{v2}	VQA^{T}	MMMU	SEED	Avg.
Retain 192 Tokens ↓ 66.7%											
-	-	60.03	62.57	1781.66	87.24	69.01	75.35	57.10	35.44	58.78	98.0%
✓	-	60.11	63.01	1784.31	87.78	69.80	77.01	57.19	36.56	58.89	99.0%
-	✓	60.52	63.23	1801.97	88.72	68.77	75.46	57.49	35.98	58.87	98.8%
✓	✓	**60.57**	**63.41**	**1809.47**	**88.94**	**70.24**	**77.42**	**58.01**	**37.33**	**58.94**	**99.9%**
Retain 128 Tokens ↓ 77.8%											
-	-	59.43	61.86	1741.23	87.13	68.86	74.55	56.79	35.67	57.39	97.2%
✓	-	59.72	61.36	1733.30	87.50	69.42	**76.73**	56.62	36.11	57.46	97.6%
-	✓	59.87	62.00	**1747.59**	87.58	68.52	74.76	56.96	35.89	57.55	97.5%
✓	✓	**60.15**	**62.11**	1731.46	**87.79**	**69.89**	**76.73**	**57.96**	**36.22**	**57.73**	**98.3%**
Retain 64 Tokens ↓ 88.9%											
-	-	58.36	60.22	1626.80	85.56	68.17	72.97	55.35	35.21	**55.84**	94.8%
✓	-	58.61	60.23	1631.64	86.28	68.71	75.61	56.26	35.44	55.46	95.6%
-	✓	58.95	60.57	1648.61	86.53	68.52	73.25	55.71	35.56	55.71	95.4%
✓	✓	**59.12**	**60.58**	**1664.51**	**86.98**	**69.80**	**75.81**	**56.32**	**36.11**	55.78	**96.5%**

4.2 Ablation Studies and Further Analyses

Ablation Studies. We conduct ablation studies on the two modules of CaRe, PRCA and CGTC, across nine benchmarks, as shown in Table 4. When PRCA is removed, we adopt the existing token selection method ZOO-Prune Kim et al. (2026) as the baseline. Table 4 clearly shows that adding either PRCA or CGTC consistently improves performance, while combining both leads to more substantial performance gains over the baseline. These findings show that the two modules are both effective and complementary for visual token calibration. Their combination achieves the best results. Further analysis is provided in Appendix D.1.

Impact of h and m. We analyze the perturbation step size h and number of perturbation directions m on POPE with LLaVA-1.5-7B. Across three cases, $h = 5 \times 10^{-4}$ and $m = 64$ yield high F1 scores. Since lower values of m reduce computation, we adopt these settings for accuracy–efficiency balance, as shown in Figure 4.

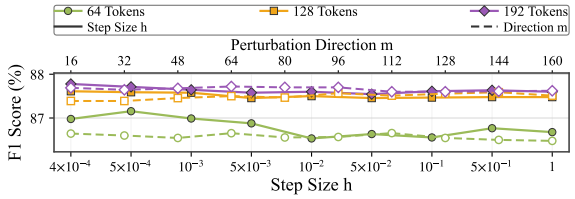

Figure 4: Ablation study of the h and m.

Impact of θ_s and θ_c. As shown in Figure 5, we perform an ablation in POPE with LLaVA-1.5-7B for the affinity threshold θ_s and the confidence threshold θ_c. These parameters jointly determine calibration-signal selection and thus affect token calibration. The left heatmap reports F1 scores, where higher values indicate better combinations. The right heatmap shows signal acceptance rate: excessively high rates introduce redundant signals, whereas excessively low rates may exclude critical ones. We set $\theta_s = 0.4$ and $\theta_c = 0.9$. Further analyses are provided in Appendix D.2.

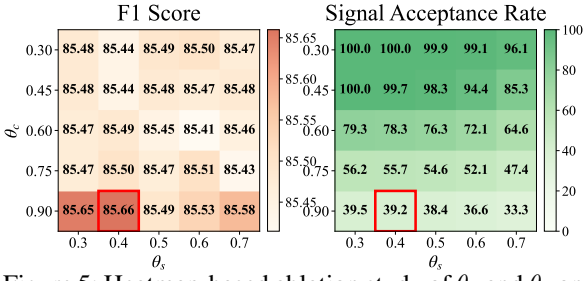

Figure 5: Heatmap-based ablation study of θ_s and θ_c on POPE with LLaVA-1.5-7B at a pruning rate of 94.4%.

Impact of λ, η and α. As shown in Figure 6, we conduct ablation studies on LLaVA-1.5-7B under three pruning ratios across six benchmarks: GQA, MMB, MME, $POPE$, SQA, and VQA^T. We evaluate the robustness regularization strength λ used in Calib-Anchor computation, the spatial weight η in semantic–spatial affinity estimation under confidence gating, and the calibration-signal weight α during token calibration. The optimal values are $\lambda = 0.5$, $\eta = 0.45$, $\alpha = 0.15$. Details are provided in the Appendix D.2.

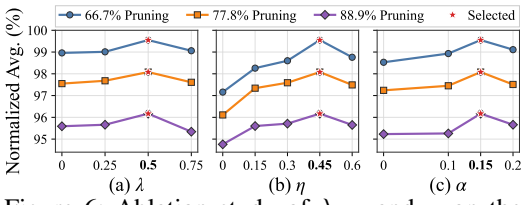

Figure 6: Ablation study of λ, η and α on the normalized average over six benchmarks.

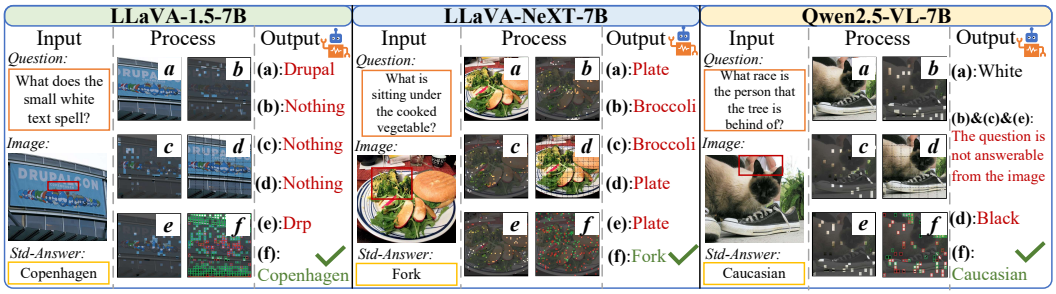

Figure 7: Qualitative visualizations. (a) Baseline; (b) sensitivity-based token selection; (c) similarity-based token selection; (d) token compression via feature merging; (e) PRCA-only retaining selected Calib-Anchors; (f) **CaRe (Ours)**. "Process" shows image processing; red and green boxes in (f) mark Calib-Anchors and calibration signals. See Appendix E.1 & E.2 for details.

Qualitative Visualization As shown in Figure 7, we evaluate three tasks (VQA^T, GQA and VQA^{v2}) across three models and present a subset of representative results. The baseline reveals that redundant visual information interfere with prediction, while (b–d) exhibit substantial semantic drift. Moreover, (e) shows that Calib-Anchors alone are insufficient, highlighting the necessity of calibration signals. These results further validate the superiority of our CaRe method, which effectively preserves critical visual semantics and improves prediction reliability.

Inference Efficiency. To evaluate efficiency, Figure 8 compares the prefilling time and end-to-end latency of CaRe with the selection method ZOO-Prune Kim et al. (2026) and the compression method VisionZip Yang et al. (2025b). Prefilling is dominated by visual-token processing and is critical for pruning-based acceleration. Our method achieves better performance while maintaining low latency, consistently offering a superior accuracy–efficiency trade-off across different token budgets and pruning ratios.

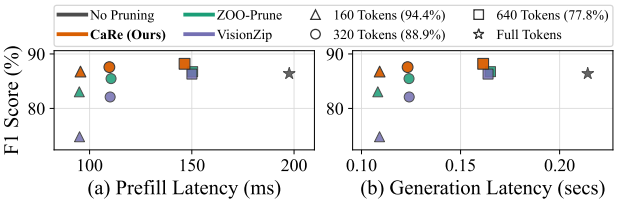

Figure 8: Inference efficiency comparison between CaRe and competing methods on LLaVA-NeXT-7B across different pruning ratios on POPE.

Under unified inference and timing protocols on the same hardware, we measure prefilling time and end-to-end latency for three acceleration methods and the unpruned model. CaRe achieves a $2.30\times$ speedup over unpruned inference. Its calibration adds only lightweight tensor operations without additional LLM forward passes; therefore, the associated overhead remains small and does not offset the acceleration gains. More details are provided in the Appendix C.

5 Conclusion and Limitations

This paper introduces the principle of "**Calibrate Before Reason**" into visual token reduction, thereby preventing efficiency gains from coming at the cost of semantic fidelity, against semantic drift. And we propose CaRe, a training-free robust visual token reduction framework, which first identifies robust calibration anchors, then extract calibration signals from the remaining tokens and use them to calibrate the anchor representations. Extensive experiments across diverse VLM architectures show that CaRe formulation consistently improves normalized performance under fixed token budgets while reducing prefill time and end-to-end latency. CaRe opens a promising direction for token reduction by shifting the focus from merely selecting tokens to calibrating retained representations for greater semantic completeness.

CaRe remains sensitive to visual evidence that is weakly encoded, highly ambiguous, or severely degraded. The calibration process refines and redistributes semantics already contained in the visual tokens, yet unable to fully recover information lost during visual encoding. Therefore, extending token calibration to the visual encoding stage for more comprehensive calibration represents a natural direction for future research.

AI USE STATEMENT

Generative AI tools were used to assist with translation, English drafting and polishing, organizing the presentation of author-provided experimental analyses, providing reviewer-style feedback on the exposition of the method and experiments, and LaTeX formatting. Generative AI was not used to generate datasets, execute experiments, or produce the reported numerical results. All AI-assisted content and suggestions were manually reviewed, revised, and checked by the authors against the implementation, experimental records, and relevant sources. The authors made all final methodological and interpretive decisions and take full responsibility for the accuracy, originality, claims, and artifacts of this work.

ETHICS STATEMENT

This work uses publicly released pretrained vision-language models and established evaluation benchmarks. It does not introduce a new dataset, collect personal data, or involve human subjects or human annotation. CaRe is a training-free, inference-time visual token reduction method and does not modify the parameters or knowledge of the underlying models. Nevertheless, aggressive token reduction may suppress fine-grained or weakly encoded visual evidence and consequently alter model predictions. Moreover, CaRe inherits the biases, hallucination risks, privacy limitations, and other safety concerns of the underlying models and benchmark data. Therefore, the method should not be deployed in safety-critical or high-stakes applications without task-specific validation and appropriate safeguards. Representative failure cases and current limitations are discussed in Appendix E.2 and Appendix F.

REPRODUCIBILITY STATEMENT

The complete formulation and algorithmic workflow of CaRe are presented in Section 3. Appendix B provides the theoretical derivation of the computational cost and TFLOPs calculation. Appendix C documents the evaluated backbones, image preprocessing, implementation environment, hardware configuration, visual token budgets, hyperparameters, random seed, evaluation framework, benchmark protocols, metrics, and complete comparison results. Appendix D reports the module-wise ablations and hyperparameter analyses, while Appendix E presents additional qualitative results and representative failure cases. CaRe is entirely training-free and therefore requires no method-specific trained checkpoints. An anonymized source-code archive is included in the supplementary materials for review and will be made publicly available upon publication.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1773–1781, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15710–15719, 2024.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. Variation-aware vision token dropping for faster large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3489–3499, 2026.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36:49250–49267, 2023.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. In *International Conference on Learning Representations*, volume 2026, pp. 48075–48099, 2026.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22654–22663, 2025.

Tianxiao Gao, Shanwei Zhao, Shuo Fang, Shiai Zhu, and Chenguang Ma. Quietprune: Query-guided early token pruning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3553–3562, 2026.

Yuhang Han, Xuyang Liu, Zihan Zhang, Pengxiang Ding, Junjie Chen, Honggang Chen, Donglin Wang, Qingsen Yan, and Siteng Huang. Filter, correlate, compress: Training-free token reduction for mllm acceleration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4601–4609, 2026.

Lianyu Hu, Liqing Gao, Fanhua Shang, Liang Wan, and Wei Feng. illava: An image is worth fewer than 1/3 input tokens in large multimodal models. In *International Conference on Learning Representations*, volume 2026, pp. 73724–73737, 2026.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong YU, Laizhong Cui, and Qi Tian. Nüwa: Mending the spatial integrity torn by VLM token pruning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=C9yclwdquU.

Zhong Ji, Zhihao Li, Yan Zhang, Yanwei Pang, and Xuelong Li. Visual semantic contextualization network for multi-query image retrieval. *IEEE Transactions on Multimedia*, 2025.

Youngeun Kim, Youjia Zhang, Huiling Liu, Aecheon Jung, Sunwoo Lee, and Sungeun Hong. Zooprune: Training-free token pruning via zeroth-order gradient estimation in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39572–39582, 2026.

DongJae Lee, Jiwan Hur, Jaehyun Choi, Jaemyung Yu, and Junmo Kim. Frequency-aware token reduction for efficient vision transformer. *Advances in Neural Information Processing Systems*, 38:34058–34083, 2026.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

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

Naoya Sogi, Takashi Shibata, and Makoto Terao. Object-aware query perturbation for cross-modal image-text retrieval. In *European Conference on Computer Vision*, pp. 447–464. Springer, 2024.

Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, and Rongrong Ji. Flashsloth: Lightning multimodal large language models via embedded visual compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14570–14581, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Ruiqi Wang, Dezhong Zhao, Ziqin Yuan, Tianyu Shao, Guohua Chen, Dominic Kao, Sungeun Hong, and Byung-Cheol Min. Primt: Preference-based reinforcement learning with multimodal feedback and trajectory synthesis from foundation models. *Advances in Neural Information Processing Systems*, 38:153095–153135, 2026a.

Ziyang Wang, Mengwei Li, Hao Yin, Wenhao Liu, and Zilei Wang. Posprune: Visual token pruning with positional bias correction for efficient large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10485–10493, 2026b.

Zhuojie Wu, Shijie Wang, and Xin Yu. Metom: Metadata-guided token merging for efficient video llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10441–10450, 2026.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19803–19813, 2025a.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025b.

Evelyn Zhang, Fufu Yu, Aoqi Wu, Zichen Wen, Ke Yan, Shouhong Ding, Biqing Qi, and Linfeng Zhang. D^2pruner: Debiased importance and structural diversity for mllm token pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12412–12420, 2026a.

Jusheng Zhang, Xiaoyang Guo, Kaitong Cai, Qinhan Lv, Yijia Fan, Wenhao Chai, Jian Wang, and Keze Wang. Hybrid token compression for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31889–31899, 2026b.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025a.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster, 2025b. URL https://arxiv.org/abs/2412.01818.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, volume 2025, pp. 53285–53310, 2025c.

Yan Zhang, Zhong Ji, Yanwei Pang, Jungong Han, and Xuelong Li. Modality-experts coordinated adaptation for large multimodal models. *Science China Information Sciences*, 67(12):220107, 2024a.

Yan Zhang, Zhong Ji, Di Wang, Yanwei Pang, and Xuelong Li. User: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing*, 33: 595–609, 2024b.

Yan Zhang, Zhong Ji, Changxu Meng, and Yanwei Pang. iebaker: Improved remote sensing image-text retrieval framework via eliminate before align and keyword explicit reasoning. *Expert Systems with Applications*, pp. 128968, 2025d.

Yan Zhang, Qin Zhijin, Xu Feng, Ding Guiguang, and Han Jungong. CofactGVR: Counterfactual intervention for grounded visual reasoning. In *Forty-third International Conference on Machine Learning*, 2026c.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, pp. 74840–74857. PMLR, 2025e.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

CONTENTS OF TECHNICAL APPENDICES

A Detailed Discussion of the Motivation **15**

A.1 Additional Motivation 15

A.2 Calibrated Semantic Retention 15

B Theoretical Derivation of TFLOPs **16**

C Experiment Setup **17**

C.1 Model Settings 17

C.2 Implementation Details 18

C.3 Evaluation Details and Benchmarks 19

C.4 Comparative Experiment Analysis 21

D Ablation Studies and Hyperparameter Analysis **22**

D.1 Ablation Studies with Module 22

D.2 Hyperparameter Analysis 23

E Additional Qualitative Results **25**

E.1 More Visualization Examples 25

E.2 Failure Case Examples 26

F Discussion and Future Work **26**

A Detailed Discussion of the Motivation

This section expands the motivation for CaRe by examining why fixed-budget visual token reduction can induce semantic drift and by elaborating on the failure modes illustrated in Figure 1 of the main paper. We then introduce Calibrated Semantic Retention (CSR), detail its question-conditioned formulation and matched-budget evaluation protocol, and clarify its role as a representation-level diagnostic. Finally, we use Figure 2 of the main paper to explain how visual token calibration preserves complementary semantics while maintaining a fixed token budget.

A.1 Additional Motivation

Fixed-budget visual token reduction aims to reduce visual tokens while preserving the information required for multimodal reasoning. Existing methods mainly rely on token selection, token merging and compression. However, satisfying a token budget does not guarantee semantic consistency with the full visual sequence. Task-relevant tokens may be pruned, semantically incompatible regions may be fused, and foreground evidence may be diluted by dominant background features. We refer to this discrepancy as **semantic drift**. The objective should therefore be not merely to retain fewer tokens, but to preserve the semantics required by the current question, which is crucial for model reasoning.

The upper panel of Figure 1 in the main paper illustrates these failure modes. The full sequence jointly represents the dog, cat, grass, and store. Selection-based reduction may remove less salient contextual tokens, causing the "store" semantics to disappear. Merging or compression may instead combine visually similar yet semantically different regions. For example, cat features may be mixed with road or background features, weakening the evidence for "cat." The retained tokens may remain visually plausible, but their combined representation no longer covers the complete scene semantics.

A.2 Calibrated Semantic Retention

To quantify this effect, the lower panel of Figure 1 in the main paper reports **Calibrated Semantic Retention (CSR)**, abbreviated as Semantic Retention in the bar chart. CSR is a question-conditioned representation-coverage metric that estimates how much source information remains recoverable from the reduced visual sequence. We use calibration to denote a gated residual update of retained visual features; it does not refer to probabilistic calibration of predictive confidence. For each image–question pair, let $V = \{v_i\}_{i=1}^{N}$, $Q = \{t_l\}_{l=1}^{T}$, and $V_S = \{v_j\}_{j=1}^{K}$ denote the full visual tokens, non-special question-token embeddings, and reduced visual tokens, respectively. Both V and V_S are mapped through the same frozen multimodal projector, while Q is obtained from the frozen LLM embedding layer. We first measure the relevance of each source token to the question as follows:

$$r_i = \max_{1 \le l \le T} \left[\bar{v}_i^{\top} \bar{t}_l \right]_+, \qquad [x]_+ = \max(x, 0). \tag{10}$$

The resulting relevance scores are normalized over the full visual sequence as follows:

$$w_i = \begin{cases} \dfrac{r_i}{\sum_{\ell=1}^{N} r_\ell}, & \sum_{\ell=1}^{N} r_\ell > 0, \\ \dfrac{1}{N}, & \text{otherwise.} \end{cases} \tag{11}$$

The uniform fallback avoids an undefined score when none of the visual tokens has positive similarity to the question. Importantly, w_i is computed once from the same uncompressed reference sequence and then reused for every reduction method. For each source token, we next measure whether its representation remains covered by any reduced token:

$$c_i = \max_{1 \le j \le K} \left[\bar{v}_i^{\top} \bar{v}_j \right]_+. \tag{12}$$

Here, $c_i \in [0, 1]$ represents source-to-output coverage: a high value indicates that the semantics carried by v_i still have a compatible carrier in the reduced sequence, irrespective of whether that carrier

Table 5: TFLOPs comparison between CaRe and the baseline across three backbones. Qwen2.5-VL-7B processes images at variable resolutions, its TFLOPs are calculated using a representative sample, which is encoded 1,110 tokens.

Backbone	method	Rate	Tokens	TFLOPs
LLaVA-1.5-7B	baseline	0	576	8.54
LLaVA-1.5-7B	CaRe	66.7%	192	3.47
LLaVA-1.5-7B	CaRe	77.8%	128	3.02
LLaVA-1.5-7B	CaRe	88.9%	64	2.18
LLaVA-NeXT-7B	baseline	0	2880	42.7
LLaVA-NeXT-7B	CaRe	77.8%	640	12.5
LLaVA-NeXT-7B	CaRe	88.9%	320	8.2
LLaVA-NeXT-7B	CaRe	94.4%	160	6.1
Qwen2.5-VL-7B	baseline	0	1110	17.3
Qwen2.5-VL-7B	CaRe	80.0%	222	5.4
Qwen2.5-VL-7B	CaRe	90.0%	111	3.8

was produced by selection, merging, or residual calibration. CSR is then defined as the question-weighted coverage of the full source sequence:

$$\mathrm{CSR} = \sum_{i=1}^{N} w_i c_i. \tag{13}$$

The bar chart reports $100 \times \mathrm{CSR}$, averaged over the same 1,000 evaluated image-question pairs with the backbone of LLaVA-1.5-7B across GQA, MME, SQA and VQA^T. All methods are evaluated at the same output-token budget and with the same reference tokens, representation layer, image-question samples, and scoring rule; no method-specific retention mask, confidence gate, merge assignment, or perturbation score enters the calculation. CSR is used only as an auxiliary representation-level diagnostic rather than a causal measure of answer preservation.

Figure 2 in the main paper further contrasts conventional reduction with visual token calibration. Selection may preserve insufficient evidence and produce "dog", whereas compression may over-mix foreground and background information and produce "bird". CaRe instead identifies robust **Calib-Anchors**, filters reliable residual signals, and routes them to compatible anchors. It therefore recovers useful semantics while maintaining the fixed output budget, providing a more reliable visual basis for predicting "horse".

B Theoretical Derivation of TFLOPs

This section provides a detailed derivation of the computational cost reported in the main paper. We decompose the overall cost into perturbation-robust calibration-anchor selection, residual-signal calibration and fusion, and LLM prefilling, and derive the corresponding expressions in Eqs 14-20. Table 5 then applies this formulation to LLaVA-1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B under their respective token budgets.

We estimate the single-sample computational cost of CaRe by decomposing inference into three stages: robust calibration-anchor selection, residual-signal calibration and fusion, and LLM prefill after token reduction. The visual encoder is excluded because it processes the complete image before token reduction and therefore introduces the same overhead for all compared methods. Let N and K denote the numbers of original and retained visual tokens, respectively, and let T denote the number of text tokens. The number of discarded tokens N_d, retention ratio ρ, and sequence length W entering the LLM are defined as:

$$N_d = N - K, \qquad \rho = \frac{K}{N}, \qquad W = T + K. \tag{14}$$

In PRCA, CaRe estimates the stable model-side influence of each visual token using m symmetric perturbation directions and jointly considers token diversity during anchor selection. Its computational cost is approximated by

$$\mathcal{F}_{\text{score}} = 2mF_M(N) + O(mNH + N^2 d_v + KN), \tag{15}$$

where d_v and H are the dimension of the visual characteristics and the hidden dimension of LLM, $F_M(N)$ denotes the FLOPs of the multimodal projector in the actual backbone when processing N tokens. The first term measures positive and negative perturbation responses through the projector, assuming that one multiply-add operation contributes two FLOPs. The remaining terms correspond to the estimation of the response-magnitude, the computation of pairwise diversity, and the iterative selection of calibration anchors K.

CGTC further examines the N_d unselected tokens and extracts reliable complementary signals. Its cost is:

$$\mathcal{F}_{\text{res}} = O(N_d K d_v + N_d K). \tag{16}$$

Here, $N_d K d_v$ measures the affinity between residual tokens and calibration anchors, while $N_d K$ represents scalar-level operations in gated mechanisms, weight normalization, and residual signal weighted fusion. The final term covers confidence gating, weight normalization, and weighted signal fusion. The complete visual-token processing cost is:

$$\mathcal{F}_{\text{CaRe–select}} = \mathcal{F}_{\text{score}} + \mathcal{F}_{\text{res}}. \tag{17}$$

After calibration, the resulting W-token sequence is fed into the language model. For one decoder-only Transformer layer, the prefill cost is estimated as:

$$\mathcal{F}_{\text{layer}} = 8WH^2 + 6WHI + 4W^2 H, \tag{18}$$

where I is the FFN intermediate dimension. The three terms respectively represent the Q/K/V/output projections, the up, gate, down projections in the gated FFN, and the two matrix multiplications in self-attention. For an LLM containing L Transformer layers, the total prefill cost becomes

$$\mathcal{F}_{\text{prefill}} = L\left(8WH^2 + 6WHI + 4W^2 H\right). \tag{19}$$

Finally, the total computational cost and its TFLOPs representation are given by

$$\begin{aligned} \mathcal{F}_{\text{total}}^{\text{CaRe}} &= \mathcal{F}_{\text{CaRe–select}} + \mathcal{F}_{\text{prefill}}, \\ \text{TFLOPs} &= \frac{\mathcal{F}_{\text{total}}^{\text{CaRe}}}{10^{12}}. \end{aligned} \tag{20}$$

This formulation separates the calibration overhead from the token-budget-dependent LLM prefill cost, enabling a consistent comparison across different retained-token settings. The comparison of TFLOPs between our method and the baseline is presented in Table 5.

C Experiment Setup

This section documents the complete experimental protocol and extends the comparative evaluation presented in the main paper. We describe the model and visual-token configurations, implementation environment, pruning ratios, hyperparameter settings, evaluation pipeline, and benchmark-specific metrics, together with the definition of normalized average performance. Tables 6–8 report the complete comparison results across three backbones, while Figure 9 presents the corresponding prefilling-time and end-to-end latency analyses.

C.1 Model Settings

We evaluate CaRe on three backbones: LLaVA-1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B. Both LLaVA-based models employ CLIP as the vision encoder and apply their default image preprocessing pipelines before visual encoding. Specifically, LLaVA-1.5-7B resizes each input image to 336×336, which is subsequently encoded into 576 visual tokens. LLaVA-NeXT-7B is designed



Figure 9: Inference efficiency comparison between CaRe and competing methods on LLaVA-1.5-7B, LLaVA-NeXT-7B and Qwen2.5-VL-7B across different pruning ratios on POPE with all 9,000 samples.

to accommodate higher-resolution inputs and processes each image at a resolution of 672×672, producing 2,880 visual tokens after CLIP encoding.

For the Qwen series, we adopt Qwen2.5-VL-7B, which is built upon the Qwen2.5-7B language model and employs a dynamic-resolution ViT encoder equipped with window attention. Consequently, the number of visual tokens varies according to the resolution of each input image. To ensure a fair comparison, we retain the original image preprocessing pipeline of each backbone without modification. All experiments are conducted on the same server under consistent hardware and software settings.

C.2 Implementation Details

All experiments are conducted on a single NVIDIA GeForce RTX 5090 D v2 GPU with batch size 1 and random seed 42. CaRe is entirely training-free and does not rely on attention scores for token selection or calibration. Its main computation is concentrated in the perturbation operations performed around the lightweight projector layer. Consequently, the method requires neither backbone-specific retraining nor sample-specific optimization. Although the absolute computational cost varies with the number of input visual tokens, the overall calibration procedure can be applied directly to different backbones. Moreover, because most intermediate quantities required for calibration-signal selection have already been obtained during calibration-anchor construction, the additional overhead introduced by CGTC is relatively small. As shown in Figure 9, CaRe achieves a favorable performance–efficiency trade-off across all three backbones compared with the baseline and other conventional token reduction methods.

To provide a comprehensive evaluation under different token budgets, we adopt three pruning ratios of 66.7%, 77.8% and 88.9% for LLaVA-1.5-7B; 77.8%, 88.9% and 94.4% for LLaVA-NeXT-7B; and 80% and 90% for Qwen2.5-VL-7B. We set hyperparameter $h = 5 \times 10^{-4}$, $m = 64$, $\lambda = 0.5$, $\eta = 0.45$, $\theta_s = 0.4$, $\theta_c = 0.9$, $\alpha = 0.15$, $\sigma_p = 0.2$, $\tau_g = 0.07$, $\tau_c = 0.1$, $\tau_s = 0.07$ and $\epsilon = 1e-8$. Overall, the ablation studies demonstrate that CaRe remains robust across a broad range of hyperparameter settings. All evaluations are conducted using the LMMs-Eval framework, following the official evaluation protocols and metrics for each benchmark.

C.3 Evaluation Details and Benchmarks

We conduct a comprehensive evaluation of **CaRe** on nine widely used vision-language benchmarks, covering four complementary capability dimensions: visual question answering, advanced multimodal reasoning, object hallucination evaluation, and comprehensive multimodal assessment. All experiments follow the official evaluation protocols and data splits of the corresponding benchmarks. We evaluate all methods under the same model configuration, visual-token budget, and decoding settings to ensure a fair comparison. All evaluations are conducted in a single-model, zero-shot setting without task-specific fine-tuning.

For evaluation metrics, we report Accuracy (Acc.) for VQA^{v2}, GQA, ScienceQA-IMG, TextVQA, MMBench, MMMU and SEED-Bench. For POPE, we report the F1-score, which balances precision and recall in object-existence prediction. For MME, we report the sum of the Perception and Cognition scores, denoted as P+C. In addition to the individual benchmark results, we report a normalized average score to provide an overall comparison across benchmarks with different numerical scales. Specifically, the score on each benchmark is normalized by the performance of the corresponding unpruned model and the resulting ratios are averaged across all benchmarks:

$$\text{Avg.} = \frac{1}{N} \sum_{i=1}^{N} \frac{s_i}{s_i^{\text{full}}} \times 100\%, \tag{21}$$

where s_i denotes the model performance on the i-th benchmark, while s_i^{full} denotes the performance of the corresponding unpruned model and N is the number of diverse benchmarks evaluated.

Visual Question Answering. This category evaluates whether a model can understand visual content and answer questions grounded in the input image. We adopt four representative benchmarks:

- **VQA**^{v2} is a general-purpose visual question answering benchmark containing open-ended questions about real-world images. Its balanced data construction reduces language priors and requires models to jointly understand visual content and natural-language questions.
- **GQA** focuses on compositional reasoning over diverse scene graphs. It contains structured questions involving object attributes, spatial relations, multi-step visual reasoning, and so on.
- **ScienceQA-IMG** evaluates multimodal scientific reasoning using questions accompanied by images or diagrams. It requires both domain knowledge and the ability to interpret visual evidence.
- **TextVQA** assesses the ability to recognize and reason over textual information embedded in natural images. Successful performance therefore depends on both optical character recognition and contextual reasoning.

Advanced Multimodal Reasoning. Beyond conventional visual question answering, we evaluate the model's ability to perform fine-grained perception, knowledge-intensive reasoning, and complex cross-modal inference on three challenging benchmarks:

- **MMBench** evaluates a broad range of perception and reasoning capabilities across multiple fine-grained skill dimensions, including attribute recognition, spatial understanding, logical reasoning, knowledge-based inference, and so on.
- **MMMU** is a multidisciplinary benchmark containing expert-level questions from more than 30 academic subjects, which are grouped into six major disciplines. Many questions involve complex diagrams, charts, tables, and domain-specific knowledge.

Table 6: Complete performance comparison on **LLaVA-1.5-7B**.

Method	GQA	MMB	MME	POPE	SQA	VQA^{v2}	VQA^{T}	MMMU	SEED	Avg.
Total 576 Tokens										
LLaVA-1.5-7B	61.90	64.70	1862.00	85.90	69.50	78.50	58.20	36.30	58.60	100%
Retain 192 Tokens ↓ 66.7%										
ToMe	54.30	60.50	1563.00	72.40	65.20	68.00	52.10	-	-	-
FastV	52.70	61.20	1612.00	64.80	67.30	67.10	52.50	34.30	57.10	89.6%
HiRED	58.70	62.80	1737.00	82.80	68.40	74.90	47.40	-	-	-
SparseVLM	57.60	62.50	1721.00	83.60	69.10	75.60	56.10	33.80	55.80	95.5%
PyramidDrop	57.10	63.20	1766.00	82.30	68.80	75.10	56.10	-	-	-
VisionZip	59.30	63.00	1782.00	85.30	68.90	76.80	57.30	36.60	56.40	97.9%
DivPrune	59.97	62.54	1762.30	87.00	68.91	76.87	56.97	35.44	58.71	98.0%
FiCoCo-V	58.51	61.74	1755.61	81.47	67.42	74.22	55.50	36.11	57.95	96.0%
V^2Drop	58.60	62.37	1789.76	85.10	69.30	74.92	55.62	35.87	56.74	97.0%
ZOO-Prune	60.03	62.57	1781.66	87.24	69.01	75.35	57.10	35.44	58.78	98.0%
***CaRe* (Ours)**	**60.57**	**63.41**	**1809.47**	**88.94**	**70.24**	**77.42**	**58.01**	**37.33**	**58.94**	**99.9%**
Retain 128 Tokens ↓ 77.8%										
ToMe	52.40	53.30	1343.00	62.80	59.60	63.00	49.10	-	-	-
FastV	49.60	56.10	1490.00	59.60	61.80	61.80	50.60	34.90	55.90	84.7%
HiRED	57.20	61.50	1490.00	59.60	60.20	61.80	50.60	-	-	-
SparseVLM	56.00	60.00	1696.00	80.50	67.10	73.80	54.90	33.80	53.40	93.0%
PyramidDrop	56.00	61.00	1644.00	82.30	68.30	72.90	55.10	-	-	-
VisionZip	57.60	62.00	**1761.70**	83.20	68.90	75.60	56.80	**37.90**	54.90	96.8%
DivPrune	59.25	62.03	1718.22	86.72	68.96	75.96	56.06	35.56	56.98	96.9%
FiCoCo-V	57.59	60.34	1718.06	82.24	68.73	72.97	54.10	35.44	54.98	94.5%
V^2Drop	56.19	59.28	1712.59	80.90	68.83	73.11	53.80	35.12	55.82	93.9%
ZOO-Prune	59.43	61.86	1741.23	87.13	68.86	74.55	56.79	35.67	57.39	97.2%
***CaRe* (Ours)**	**60.15**	**62.11**	1731.46	**87.79**	**69.89**	**76.73**	**57.96**	36.22	**57.73**	**98.3%**
Retain 64 Tokens ↓ 88.9%										
ToMe	48.60	43.70	1138.00	52.50	50.00	57.10	45.30	-	-	-
FastV	46.10	48.00	1256.00	48.00	51.10	55.00	47.80	34.00	51.90	75.5%
HiRED	54.60	60.20	1599.00	73.60	68.20	69.70	44.20	-	-	-
SparseVLM	52.70	56.20	1505.00	75.10	62.20	68.20	51.80	32.70	51.10	87.0%
PyramidDrop	41.90	33.30	1092.00	55.90	68.60	69.20	45.90	-	-	-
VisionZip	55.10	60.10	**1690.00**	77.00	69.00	72.40	55.50	35.20	52.20	92.8%
DivPrune	57.78	59.28	1674.40	85.56	68.17	74.11	54.69	35.56	55.13	94.8%
FiCoCo-V	53.43	55.09	1653.67	75.29	67.92	71.63	51.88	34.22	53.14	90.1%
V^2Drop	53.52	55.21	1641.23	75.14	68.92	71.03	51.83	34.29	53.11	90.1%
ZOO-Prune	58.36	60.22	1626.80	85.56	68.17	72.97	55.35	35.21	**55.84**	94.8%
***CaRe* (Ours)**	**59.12**	**60.58**	1664.51	**86.98**	**69.80**	**75.81**	**56.32**	**36.11**	55.78	**96.5%**

- **SEED-Bench** evaluates multimodal large language models through diverse visually grounded multiple-choice questions. It covers a wide range of perception and reasoning abilities and provides a unified assessment of general multimodal understanding.

Object Hallucination Evaluation. We employ **POPE** to evaluate object hallucination, a critical failure mode in which a model incorrectly claims that an absent object appears in an image. POPE formulates object-existence verification as binary questions and measures whether the model can distinguish objects that are present from those that are absent. We report the F1-score to jointly account for precision and recall, thereby providing a relatively balanced evaluation of hallucination resistance.

Comprehensive Multimodal Assessment. We further adopt **MME** to assess the model's overall multimodal capability. MME covers a broad collection of perception-oriented and cognition-oriented tasks, including object recognition, optical character recognition, spatial understanding, commonsense reasoning, and numerical reasoning. It reports separate scores for Perception and Cognition, and we use their sum, denoted as P+C, as the final evaluation metric.

Table 7: Complete performance comparison on **LLaVA-NeXT-7B**.

Method	GQA	MMB	MME	POPE	SQA	VQA^{v2}	VQA^{T}	MMMU	SEED	Avg.
Total 2880 Tokens										
LLaVA-NeXT-7B	64.20	67.90	1842.00	86.40	70.20	80.10	61.30	35.10	70.20	100%
Retain 640 Tokens ↓ 77.8%										
SparseVLM	60.30	65.70	1772.00	-	67.70	77.10	57.80	34.60	-	-
VisionZip	61.30	**66.30**	1787.00	86.30	68.10	**79.10**	60.20	34.70	66.70	97.5%
DivPrune	61.58	65.38	1773.04	85.51	67.82	78.94	55.41	36.89	67.56	97.1%
FiCoCo-V	60.40	64.18	1780.22	86.04	67.94	75.77	54.89	35.91	65.77	95.8%
ZOO-Prune	62.19	65.10	1802.47	86.75	68.02	78.13	57.98	36.82	67.59	97.9%
***CaRe* (Ours)**	**63.87**	65.98	**1831.14**	**88.19**	**69.41**	78.02	**58.39**	**37.89**	**67.92**	**99.4%**
Retain 320 Tokens ↓ 88.9%										
SparseVLM	57.70	64.30	1694.00	-	67.30	73.40	55.90	34.40	-	-
VisionZip	59.30	63.10	1702.00	82.10	67.30	76.20	**58.90**	35.30	63.40	94.5%
DivPrune	59.63	63.66	1731.04	83.47	67.82	76.64	53.84	37.11	65.35	95.1%
FiCoCo-V	59.26	63.30	1721.74	85.18	67.51	74.08	54.67	34.33	64.51	93.9%
ZOO-Prune	60.83	64.60	1787.68	85.47	67.08	76.04	57.26	36.78	66.18	96.5%
***CaRe* (Ours)**	**61.98**	**64.92**	**1791.90**	**87.58**	**68.86**	**77.76**	57.67	**37.33**	**66.28**	**97.9%**
Retain 160 Tokens ↓ 94.4%										
SparseVLM	51.20	63.10	1542.00	-	67.50	66.30	46.40	32.80	-	-
VisionZip	55.50	60.10	1630.00	74.80	68.30	71.40	**56.20**	36.10	58.30	90.4%
DivPrune	57.79	62.29	1658.25	79.36	68.02	73.92	52.42	36.44	62.54	92.4%
FiCoCo-V	58.11	60.32	1657.81	84.27	68.01	73.22	55.10	33.98	63.06	92.4%
ZOO-Prune	59.48	64.18	1706.08	83.05	67.72	74.07	55.42	**36.67**	64.04	94.5%
***CaRe* (Ours)**	**59.97**	**64.47**	**1706.96**	**86.77**	**69.70**	**75.81**	55.77	36.56	**67.87**	**96.4%**

C.4 Comparative Experiment Analysis

Unlike previous token reduction approaches that either select visual tokens according to different importance scores or compress them through token merging, CaRe formulates token reduction as a token calibration problem. We analyze its performance on three different backbones below.

Results on LLaVA-1.5-7B. As shown in Table 6, **CaRe** consistently outperforms both score-based token selection and token compression methods across all three token budgets. With 192, 128, and 64 retained tokens, it achieves normalized average performance of 99.9%, 98.3%, and 96.5%, exceeding the strongest prior method by 1.9, 1.1 and 1.7 percentage points, respectively. At 192 tokens, CaRe even surpasses the full-token model on POPE, SQA, MMMU, and SEED, while preserving nearly all performance on the remaining benchmarks. Its advantage becomes more pronounced under aggressive reduction: at 64 tokens, it obtains the best POPE, SQA, VQA^{v2} and VQA^{T} results, indicating that calibration recovers task-relevant information that conventional selection or compression may discard or dilute. Isolated gaps remain on MME and MMMU at 128 tokens, and on MME and SEED at 64 tokens. These cases mainly involve dense global evidence or fine-grained textual cues, yet the gaps are limited and CaRe still delivers the strongest overall performance–efficiency trade-off.

Results on LLaVA-NeXT-7B. As shown in Table 7, **CaRe** shows stable superiority under increasingly aggressive token reduction. It retains 99.4%, 97.9%, and 96.4% of full-model performance with 640, 320 and 160 tokens, outperforming the strongest competing method by 1.5, 1.4 and 1.9 percentage points, respectively. At 640 tokens, it nearly matches the full model on MME and exceeds it on POPE and MMMU, demonstrating that token calibration can remove substantial redundancy without weakening and sometimes even improving reliability of reasoning. With 320 tokens, CaRe ranks first on eight of the nine reported benchmarks. In particular, its advantages on POPE and SEED remain clear at the most aggressive setting, showing strong robustness in hallucination control and general multimodal understanding. VQA^{T} is the only consistent exception, where VisionZip performs slightly better, while VisionZip also leads marginally on VQA^{v2} at 640 tokens. These small deficits may reflect the difficulty of preserving OCR-sensitive local details, but the substantially higher normalized averages confirm the overall advantage of the visual token calibration formulation.

Table 8: Complete performance comparison on **Qwen2.5-VL-7B**.

Method	GQA	MMB	MME	POPE	SQA	VQA^T	MMMU	SEED	Avg.-8
Full Tokens									
Qwen2.5-VL-7B	60.84	82.82	2310.00	86.30	88.30	77.69	50.22	74.21	100%
Retain 20% Tokens ↓ 80.0%									
VisionZip	57.27	76.72	2221.00	81.67	-	-	-	-	-
DivPrune	**60.05**	75.87	2173.00	83.11	-	-	-	-	-
FiCoCo-V	55.28	76.47	2120.00	81.29	-	-	-	-	-
ZOO-Prune	55.92	77.41	2139.14	81.71	82.40	68.80	47.11	69.67	92.8%
***CaRe* (Ours)**	58.11	**77.91**	**2274.28**	**83.94**	**84.43**	**70.85**	**47.68**	**73.16**	**95.7%**
Retain 10% Tokens ↓ 90.0%									
VisionZip	54.09	72.16	1937.00	78.97	-	-	-	-	-
DivPrune	**55.49**	71.95	2004.00	79.05	-	-	-	-	-
FiCoCo-V	52.79	71.95	1967.00	77.61	-	-	-	-	-
ZOO-Prune	53.28	72.51	1961.89	78.46	79.92	61.99	45.33	66.86	87.7%
***CaRe* (Ours)**	54.45	**73.37**	**2016.92**	**79.23**	**81.54**	**65.33**	**46.11**	**69.30**	**89.9%**

Results on Qwen2.5-VL-7B. Unlike Table 3 in the main paper, Table 8 reports the performance of CaRe and the main competing methods across eight benchmarks. Consequently, the "Avg." values differ slightly from those reported in the main paper. Nevertheless, **CaRe** consistently surpasses both ZOO-Prune, a score-based selection method, and FiCoCO-V, a token compression method. Retaining only 20% and 10% of visual tokens, it preserves 95.7% and 89.9% of the full-model performance, improving over ZOO-Prune by 2.9 and 2.2 percentage points in normalized average score. At the 20% budget, CaRe remains superior on every metric reported by ZOO-Prune and FiCoCO-V. VQA^T shows one of the most visible degradations relative to the full-token model among the accuracy-based benchmarks, reflecting its dependence on fine-grained OCR cues. Nevertheless, CaRe achieves a notable improvement on TextVQA, indicating that calibration better preserves and restores critical textual evidence under severe token reduction.

D Ablation Studies and Hyperparameter Analysis

This section provides a more comprehensive analysis of the components and hyperparameters underlying CaRe. Table 9 extends the module-wise ablation study by separately examining anchor selection, simple calibration, confidence-gated calibration, and the complete PRCA–CGTC pipeline. We further analyze the roles and sensitivity of h, m, λ, η, θ_s, θ_c, and α through Tables 10–12 and the corresponding hyperparameter studies reported in the main paper.

D.1 Ablation Studies with Module

As shown in Table 9, we conduct more comprehensive module-wise ablation studies on LLaVA-1.5-7B under three token-reduction ratios, namely 66.7%, 77.8% and 88.9%, to examine the contribution of each component in CaRe. The unpruned LLaVA-1.5-7B serves as the full-token reference. **Only Sense-Diversity** retains the original score-based selection strategy, which selects visual tokens according to sensitivity and diversity scores without applying token calibration. **Only Calib-Anchor** introduces PRCA to identify perturbation-robust calibration anchors, but does not perform subsequent information calibration. **Only Simple-Calib** applies the original selector and directly transfers residual information to the selected tokens without confidence gating. **Only Confidence-Calib** first selects tokens using the original strategy and then calibrates them through CGTC. **Calib-Anchor+Simple-Calib** combines PRCA-based anchor selection with ungated candidate-to-anchor calibration. Finally, **Full CaRe** denotes the complete CaRe framework integrating both PRCA and CGTC.

The results reveal a clear overall trend: performance improves as the calibration pipeline becomes more complete. Compared with Only Sense-Diversity, introducing PRCA alone increases the normalized average from 98.0% to 99.0%, from 97.2% to 97.6%, and from 94.8% to 95.6% under the three respective token budgets. Both "Simple-Calib" and "Confidence-Calib" also consistently improve upon the selection-only baseline, confirming that recovering residual information is beneficial

Table 9: Module ablation on LLaVA-1.5-7B.

Module	GQA	MMB	MME	POPE	SQA	VQA^{v2}	VQA^{T}	MMMU	SEED	Avg.
Total 576 Tokens										
LLaVA-1.5-7B	61.90	64.70	1862.00	85.90	69.50	78.50	58.20	36.30	58.60	100%
Retain 192 Tokens ↓ 66.7%										
Only Sense-diversity	60.03	62.57	1781.66	87.24	69.01	75.35	57.10	35.44	58.78	98.0%
Only Calib-Anchor	60.11	63.01	1784.31	87.78	69.80	77.01	57.19	36.56	58.89	99.0%
Only Simple-Calib	60.18	62.98	1787.18	88.63	69.06	75.51	57.29	37.03	58.82	98.9%
Only Confidence-Calib	60.52	63.23	1801.97	88.72	68.77	75.46	57.49	35.98	58.87	98.8%
Calib-Anchor+Simple-Calib	60.21	63.34	**1809.51**	88.76	69.85	77.12	57.80	37.02	58.93	99.6%
***Full CaRe* (Ours)**	**60.57**	**63.41**	1809.47	**88.94**	**70.24**	**77.42**	**58.01**	**37.33**	**58.94**	**99.9%**
Retain 128 Tokens ↓ 77.8%										
Only Sense-diversity	59.43	61.86	1741.23	87.13	68.86	74.55	56.79	35.67	57.39	97.2%
Only Calib-Anchor	59.72	61.36	1733.30	87.50	69.42	**76.73**	56.62	36.11	57.46	97.6%
Only Simple-Calib	59.50	61.89	1719.87	87.59	68.91	74.81	56.85	36.11	57.67	97.4%
Only Confidence-Calib	59.87	62.00	**1747.59**	87.58	68.52	74.76	56.96	35.89	57.55	97.5%
Calib-Anchor+Simple-Calib	59.92	61.77	1733.59	87.47	69.72	76.21	57.53	**36.22**	57.58	97.9%
***Full CaRe* (Ours)**	**60.15**	**62.11**	1731.46	**87.79**	**69.89**	**76.73**	**57.96**	**36.22**	**57.73**	**98.3%**
Retain 64 Tokens ↓88.9%										
Only Sense-diversity	58.36	60.22	1626.80	85.56	68.17	72.97	55.35	35.21	**55.84**	94.8%
Only Calib-Anchor	58.61	60.23	1631.64	86.28	68.71	75.61	56.26	35.44	55.46	95.6%
Only Simple-Calib	58.51	60.54	1644.63	86.68	68.77	73.29	55.75	35.33	55.62	95.3%
Only Confidence-Calib	58.95	60.57	1648.61	86.53	68.52	73.25	55.71	35.56	55.71	95.4%
Calib-Anchor+Simple-Calib	58.98	60.57	1657.81	86.54	69.60	75.76	56.15	35.56	55.62	96.1%
***Full CaRe* (Ours)**	**59.12**	**60.58**	**1664.51**	**86.98**	**69.80**	**75.81**	**56.32**	**36.11**	55.78	**96.5%**

even without robust anchor identification. Combining PRCA with "Simple-Calib" further raises the average performance to 99.6%, 97.9% and 96.1%, demonstrating that robust anchors provide more reliable destinations for information transfer.

The complete model achieves the best normalized averages of 99.9%, 98.3% and 96.5%. Relative to "Calib-Anchor+Simple-Calib", confidence-gated calibration yields additional gains of 0.3, 0.4 and 0.4 percentage points, indicating that indiscriminate information transfer is suboptimal and that CGTC effectively suppresses unreliable calibration signals. At 64 tokens, it remains best on eight of the nine benchmarks. The only exception is SEED, where Only Sense-Diversity is marginally higher by 0.06 points. Similarly, at 192 tokens, "Calib-Anchor + Simple-Calib" slightly outperforms the full model on MME. These isolated fluctuations suggest that benchmark-specific optima may occasionally differ, but they do not affect the consistent advantage of the complete framework. Together with the comparative results, the ablations verify that PRCA and CGTC are complementary and jointly enable CaRe to outperform previous token reduction methods.

D.2 Hyperparameter Analysis

Impact of h and m. In the Perturbation-Robust Calibration Anchoring (PRCA) module, the perturbation step size and the number of perturbation directions are set to $h = 5 \times 10^{-4}$ and $m = 64$, respectively. The step size h controls the magnitude of the symmetric perturbations applied to visual tokens before they enter the projector. An excessively large perturbation may move the token representation away from its local neighborhood, whereas an overly small perturbation may produce responses that are too weak to reliably characterize the projector's sensitivity. We therefore select an intermediate value that yields stable and distinguishable perturbation responses when selecting calibrated anchors.

The hyperparameter m specifies the number of perturbation directions used to estimate the model-side influence of each visual token. We organize these perturbations into paired positive and negative directions, enabling a symmetric estimation of the projector response. Increasing m generally produces a more reliable estimate by covering a broader set of local directions, but it also introduces additional computation. Considering both accuracy and efficiency, we set the number of perturbation directions $m = 64$.

Impact of λ. The robustness coefficient λ controls the contribution of response variation across perturbation directions to anchor scoring. When $\lambda = 0$, anchor selection degenerates into ranking tokens solely according to their mean response magnitude. However, a token may exhibit a strong response along only one direction while remaining nearly insensitive to the others. Such a token has a large response in a limited local direction but does not exert a stable influence on the model. Incorporating directional variation therefore helps PRCA prioritize tokens with consistently strong model-side effects. As shown in Table 10, ablation results identify $\lambda = 0.5$ as the optimal setting.

Table 10: Ablation study of λ on LLaVA-1.5-7B.

λ	GQA	MMB	MME	POPE	SQA	VQA^T	Avg.
Total 576 Tokens							
-	61.90	64.70	1862	85.90	69.50	58.20	100%
Retain 192 Tokens ↓ 66.7%							
0.00	60.42	63.06	1786	87.68	69.51	57.88	98.8%
0.25	60.41	63.26	1798	87.68	69.94	57.90	99.0%
0.50	**60.57**	**63.41**	**1809**	**88.94**	**70.24**	**58.01**	**99.6%**
0.75	60.42	63.06	1797	88.23	69.93	57.88	99.1%
Retain 128 Tokens ↓ 77.8%							
0.00	60.04	61.94	1709	87.60	69.60	57.44	97.6%
0.25	60.14	61.94	1721	87.62	69.66	57.40	97.7%
0.50	**60.15**	**62.11**	**1731**	**87.79**	**69.89**	**57.96**	**98.1%**
0.75	60.00	61.86	1719	87.58	69.60	57.44	97.6%
Retain 64 Tokens ↓ 88.9%							
0.00	58.81	60.39	1652	86.51	69.06	56.10	95.6%
0.25	58.82	**60.61**	1651	86.53	69.11	56.14	95.7%
0.50	**59.12**	60.58	**1665**	**86.98**	**69.80**	**56.32**	**96.2%**
0.75	58.82	60.31	1656	86.52	68.02	56.04	95.3%

Impact of η, θ_s and θ_c. After PRCA identifies the calibration anchors, CGTC calibrates them through a two-stage procedure. The first stage applies a confidence gate to select reliable calibration signals from the residual, unselected tokens. This stage is controlled by three hyperparameters: the spatial-proximity weight η in the semantic-spatial affinity, the best-affinity threshold θ_s, and the confidence threshold θ_c. The coefficient η balances semantic similarity and spatial proximity when matching a residual token to an anchor. Semantically similar tokens often correspond to the same object or region, but visually similar features may also arise from different objects. Spatial proximity is therefore incorporated to reduce incorrect cross-object assignments. In practice, an anchor tends to absorb semantically compatible signals from its local neighborhood, thereby enriching its representation without introducing unrelated content. Based on the Table 11, we set $\eta = 0.45$.

Table 11: Ablation study of η on LLaVA-1.5-7B.

η	GQA	MMB	MME	POPE	SQA	VQA^T	Avg.
Total 576 Tokens							
-	61.90	64.70	1862	85.90	69.50	58.20	100%
Retain 192 Tokens ↓ 66.7%							
0.00	59.89	61.77	1698	87.41	68.96	57.36	97.2%
0.15	60.21	62.11	1763	88.13	69.51	57.61	98.3%
0.30	60.26	62.23	1772	88.48	69.80	57.89	98.6%
0.45	**60.57**	**63.41**	**1809**	**88.94**	**70.24**	**58.01**	**99.6%**
0.60	60.27	62.54	1795	88.26	69.60	57.75	98.8%
Retain 128 Tokens ↓ 77.8%							
0.00	59.39	60.65	1649	87.52	69.11	56.53	96.1%
0.15	59.55	62.00	1708	87.68	69.35	57.28	97.3%
0.30	59.92	61.83	1719	87.62	69.60	57.43	97.6%
0.45	**60.15**	**62.11**	**1731**	**87.79**	**69.89**	**57.96**	**98.1%**
0.60	59.85	61.86	1724	87.51	69.40	57.22	97.5%
Retain 64 Tokens ↓ 88.9%							
0.00	59.41	59.36	1626	86.72	68.52	55.65	95.0%
0.15	58.56	59.97	1650	86.82	69.60	56.16	95.6%
0.30	58.63	60.14	1650	86.94	69.75	56.13	95.7%
0.45	**59.12**	**60.58**	**1665**	**86.98**	**69.80**	**56.32**	**96.2%**
0.60	58.65	60.05	1644	86.96	69.80	56.09	95.7%

Not every residual token should be used for calibration. Indiscriminately injecting all residual information may introduce excessive redundancy and reproduce the semantic contamination observed in conventional token compression. The threshold θ_s first requires a candidate token to exhibit a sufficiently strong preference for at least one calibration anchor. The confidence threshold θ_c then evaluates the overall reliability of the assignment by jointly considering affinity quality and assignment certainty. These two thresholds interact to control the amount and quality of information admitted by the confidence gate. As shown in Figure 5 of the main paper, we set $\theta_s = 0.4$ and $\theta_c = 0.9$. Under this configuration, approximately 39.2% of the residual tokens are retained as calibration signals. Using all residual tokens instead leads to inferior performance, confirming that only a subset contains reliable complementary semantics.

Impact of α. In the second stage of CGTC, the selected signals are fused into their matched calibration anchors. The coefficient α controls the strength of this residual injection. Calibration signals are intended to complement and reinforce the anchor representation rather than dominate it. A small α prevents useful residual semantics from being effectively transferred, whereas an excessively large

Table 12: Ablation study of α on LLaVA-1.5-7B.

α	GQA	MMB	MME	POPE	SQA	VQA^T	Avg.
Total 576 Tokens							
-	61.90	64.70	1862	85.90	69.50	58.20	100%
Retain 192 Tokens ↓ 66.7%							
0.00	60.11	63.01	1784	87.78	69.80	57.19	98.5%
0.10	60.20	62.72	1792	88.70	69.83	57.87	98.9%
0.15	**60.57**	**63.41**	**1809**	**88.94**	**70.24**	**58.01**	**99.6%**
0.20	60.45	63.36	1789	88.60	69.86	57.81	99.1%
Retain 128 Tokens ↓ 77.8%							
0.00	59.72	61.36	1733	87.50	69.42	56.62	97.2%
0.10	59.63	61.86	1743	87.51	69.06	56.99	97.5%
0.15	**60.15**	**62.11**	**1731**	**87.79**	**69.89**	**57.96**	**98.1%**
0.20	59.83	62.11	1724	87.50	69.42	56.62	97.4%
Retain 64 Tokens ↓ 88.9%							
0.00	58.61	60.23	1632	86.28	68.71	56.26	95.2%
0.10	58.47	**60.65**	1626	86.56	68.67	56.14	95.3%
0.15	**59.12**	60.58	**1665**	**86.98**	**69.80**	**56.32**	**96.2%**
0.20	58.89	60.57	1654	86.59	68.67	56.33	95.7%

value may overwrite the original anchor representation and make the procedure resemble conventional token merging. According to the Table 12, we set $\alpha = 0.15$. Overall, the ablation studies demonstrate that CaRe remains robust across a broad range of hyperparameter settings.

E Additional Qualitative Results

This section complements the quantitative evaluation with additional qualitative results across LLaVA-1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B on four vision-language tasks. Figures 10, 12, and 13 compare the full-token baseline, conventional selection and compression strategies, calibration anchors alone, and the complete CaRe framework, while explicitly visualizing the selected anchors and calibration signals. Figure 11 further presents representative failure cases involving ambiguous or low-resolution textual evidence, illustrating the remaining dependence on OCR fidelity and token-level evidence quality.

E.1 More Visualization Examples

Figures 10, 12 and 13 provide additional qualitative comparisons on LLaVA-1.5-7B, LLaVA-NeXT-7B and Qwen2.5-VL-7B, respectively. For each backbone, we visualize representative examples from four tasks, including GQA, VQA^T, VQA^{v2} and MME. The leftmost part of each example presents the original question, input image, and ground-truth answer. The remaining columns show the detailed token-processing results and all model predictions produced by six different configurations.

"**Baseline**" denotes the original model that performs inference using all visual tokens without token reduction. "**Sensitivity Select**" retains tokens solely according to their sensitivity scores, whereas "**Similarity Select**" performs token selection based on semantic similarity. These two settings represent conventional score-based token selection strategies. "**Compress**" represents feature-compression methods that aggregate the original visual tokens into a fixed number of compact tokens before subsequent inference. "**Calib-Anchors**" corresponds to an incomplete variant of our framework, in which only perturbation-robust calibration anchors are retained for inference without incorporating additional calibration signals. Finally, "**Ours**" denotes the complete CaRe framework. It first identifies stable calibration anchors and then selects informative tokens from the remaining token set as calibration signals, which are used to refine the semantic representations of the anchor tokens.

In the visualizations of CaRe, tokens enclosed by red boxes denote calibration anchors, while those enclosed by green boxes represent the selected calibration signals. Tokens without either annotation are discarded. This decomposition illustrates that CaRe does not indiscriminately retain or merge visual information. Instead, it preserves a compact set of semantically stable anchors and selectively exploits complementary evidence from the remaining tokens to correct and enrich their representations.

The qualitative examples illustrate the limitations of conventional token reduction. Sensitivity- and similarity-based selection can overlook sparse but decisive evidence, such as object materials, scene text, numerical identifiers, colors, spatial relations and clock readings. Feature compression may retain broader image coverage, but aggregating heterogeneous regions can dilute fine-grained semantics and produce incorrect answers. These failures indicate that simply retaining high-scoring tokens or merging nearby features does not guarantee preservation of task-relevant semantics.

A comparison between "**Calib-Anchors**" and "**Ours**" highlights the importance of calibration. While robust anchors retain salient regions, using them alone can still cause semantic deviations. By incorporating reliable calibration signals, CaRe preserves and reassigns complementary evidence and corrects these errors, sometimes even outperforming the unpruned baseline. These results support modeling token reduction as token calibration and show its effectiveness in mitigating semantic drift.

E.2 Failure Case Examples

As shown in Figure 11, we further present several representative failure cases. These examples reveal a limitation in fine-grained visual grounding under ambiguous or low-resolution textual cues. Both require exact transcription of names or numbers rather than coarse semantic recognition. Although CaRe preserves task-relevant evidence, calibration cannot reconstruct information that is weakly encoded, visually corrupted, or inconsistently represented across tokens; language priors may then dominate decoding. Thus, performance remains sensitive to OCR fidelity and token-level evidence quality, reflecting a deeper information bottleneck rather than a simple selection error.

F Discussion and Future Work

CaRe reframes fixed-budget visual token reduction as a calibration problem rather than a purely selective one, suggesting that preserving semantic fidelity requires not only retaining informative tokens but also refining their representations before multimodal reasoning. Its training-free and evaluated across the three tested backbones formulation makes this principle broadly applicable to existing reduction pipelines, where discarded tokens can serve as auxiliary evidence for calibrating robust anchors. Nevertheless, CaRe remains constrained by the quality of the original visual representations. When critical evidence is weakly encoded, highly ambiguous, or absent from the token sequence, projector-level calibration cannot fully reconstruct the missing information. In our view, additional directions include adaptive calibration under varying token budgets and uncertainty-aware signal selection, which may further improve semantic completeness across heterogeneous architectures and multimodal inputs.

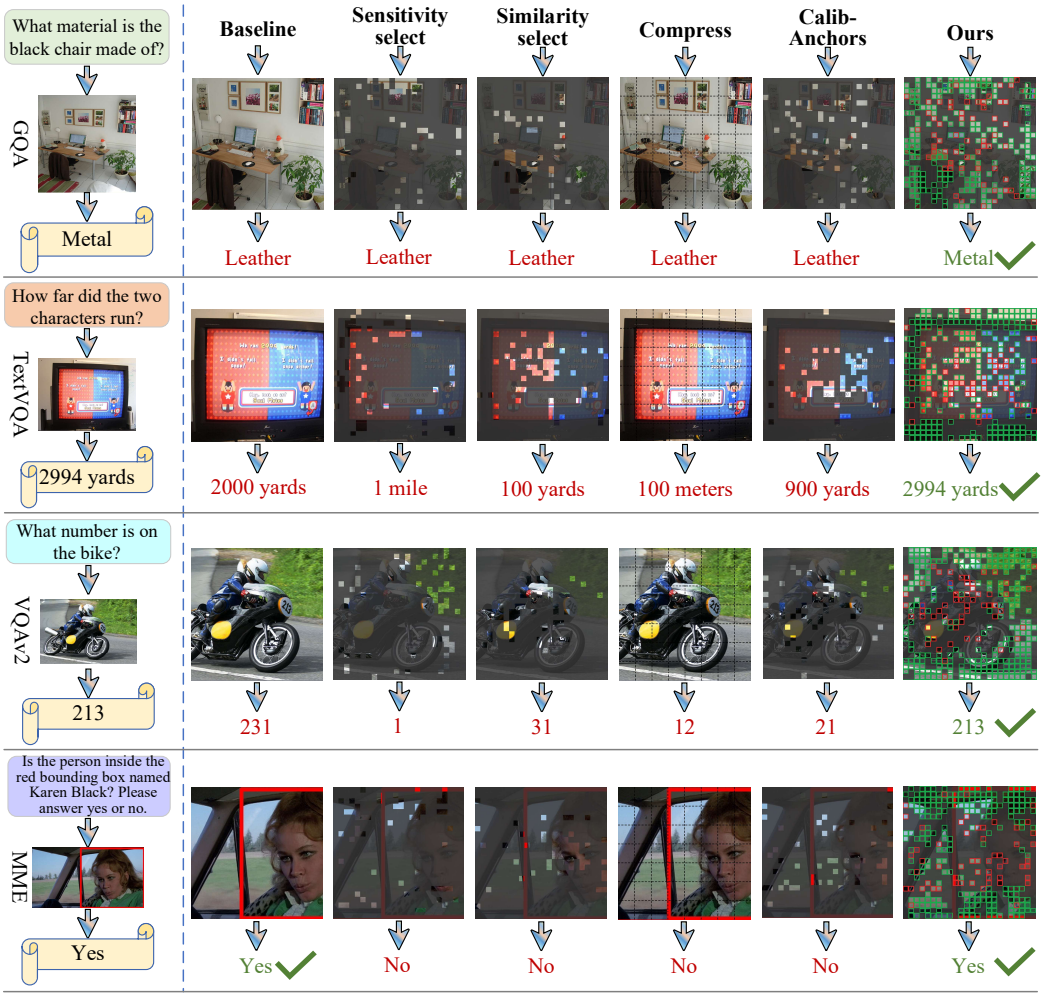

Figure 10: Qualitative visualizations on LLaVA-1.5-7B across four tasks, including GQA, VQA^T, VQA^{v2} and MME. Pruning ratio is 88.9%.

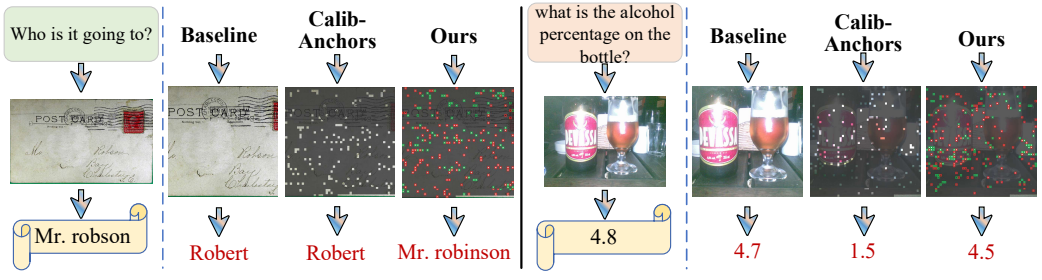

Figure 11: Failure cases of CaRe at a 94.4% pruning ratio.

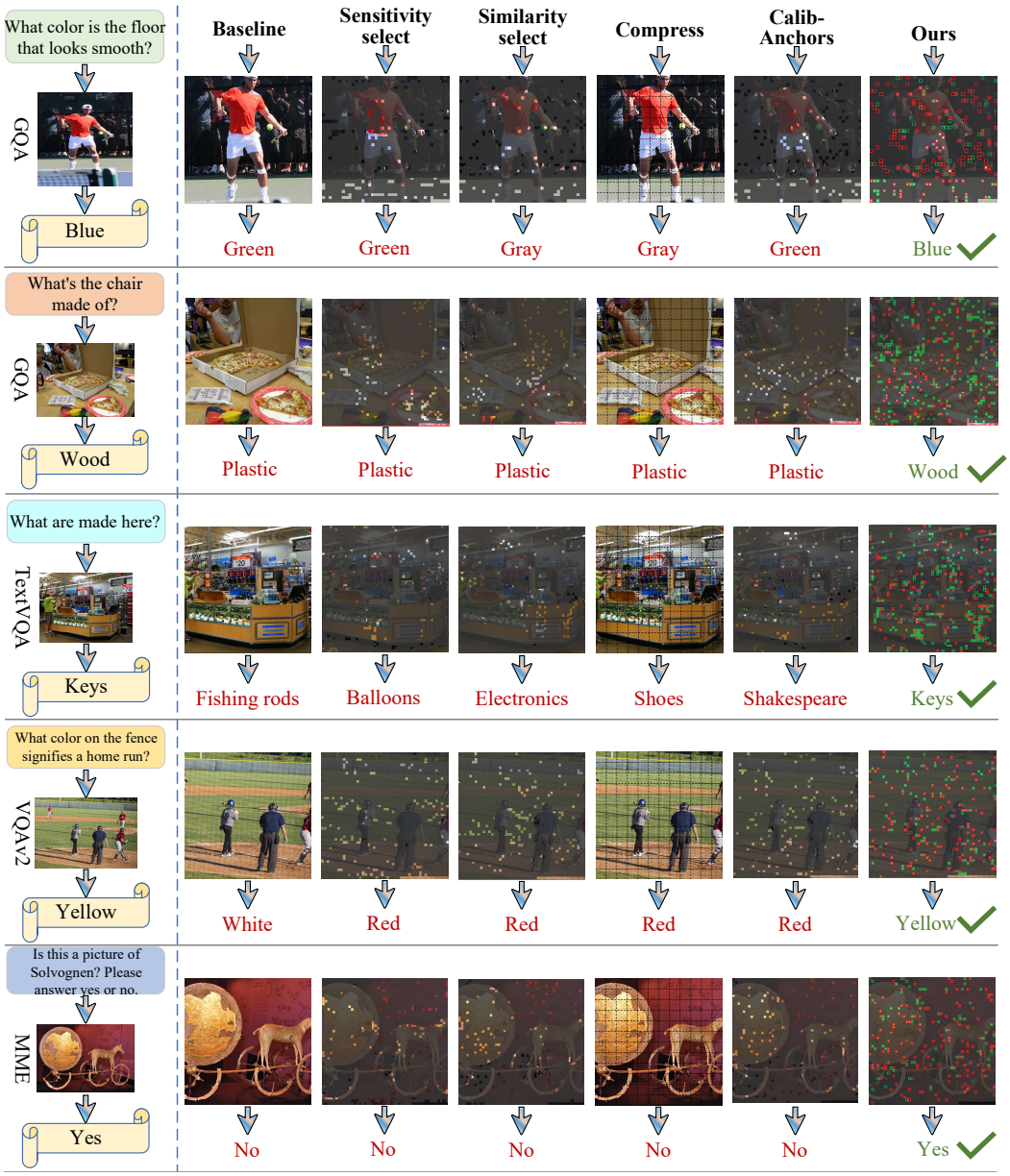

Figure 12: Qualitative visualizations on LLaVA-NeXT-7B across four tasks, including GQA, VQA^T, VQA^{v2} and MME. Pruning ratio is 94.4%.

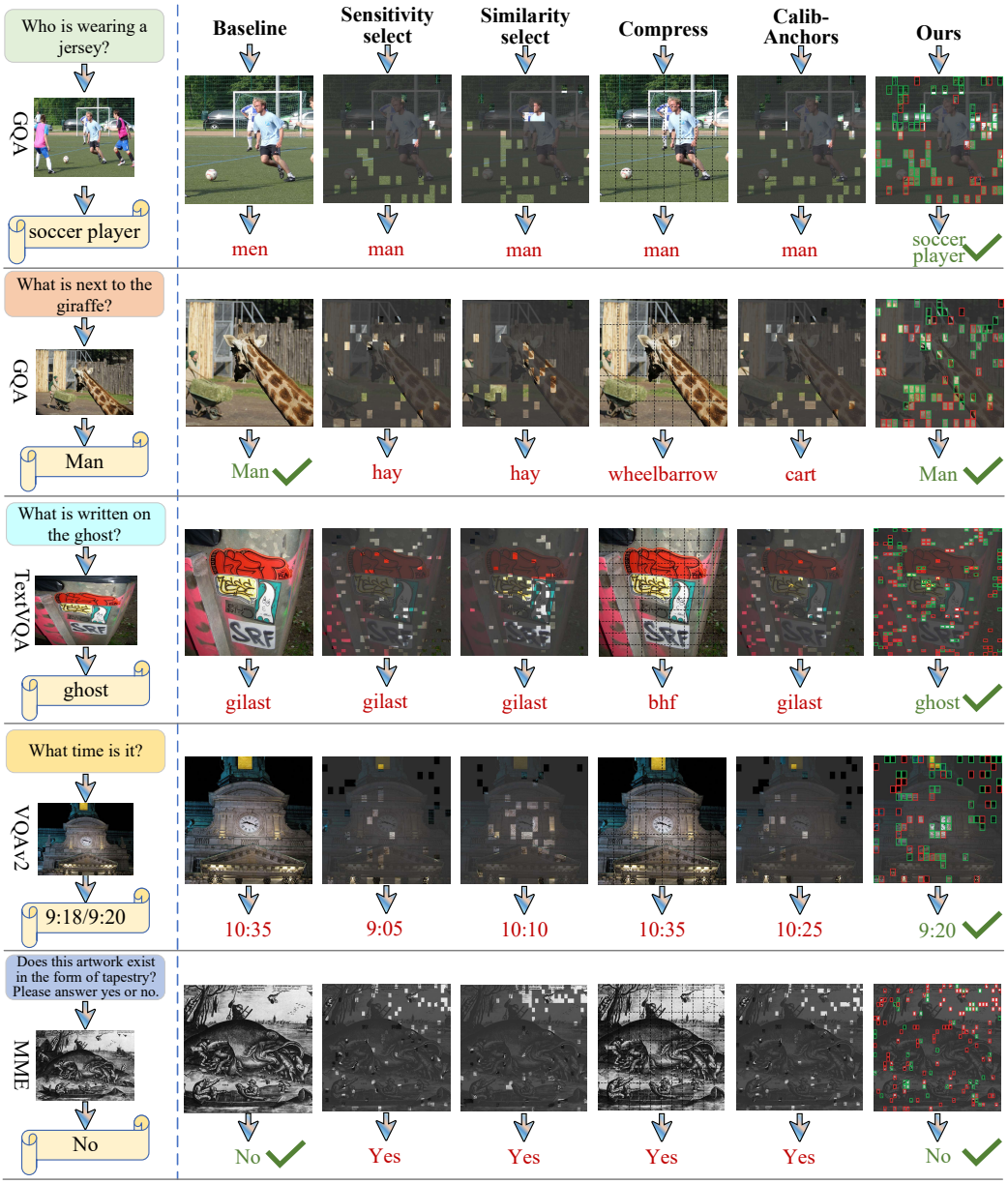

Figure 13: Qualitative visualizations on Qwen2.5-VL-7B across four tasks, including GQA, VQA^T, VQA^{v2} and MME. Pruning ratio is 90%.