RETAINED BUT UNDERUSED: MODALITY ATTENTION IMBALANCE AFTER VISUAL TOKEN PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual token pruning accelerates vision-language models (VLMs) by retaining only a small yet informative subset of image tokens. Most existing methods focus on token selection and assume that the decoder can effectively use the retained visual evidence. However, we find that pruning also changes the attention allocation between visual and textual modalities. As the token budget decreases, the average attention received by each retained visual token increases, while aggregate visual attention relative to textual attention decreases. We refer to this phenomenon as modality attention imbalance. Meanwhile, task accuracy and teacher-forced next-token prediction loss may even vary in the same direction. The output distribution also shifts toward the distribution produced without an image, indicating greater reliance on textual context. Simply rebalancing attention across all decoder layers fails to correct this imbalance. Through controlled causal interventions, we identify task-specific decoder regions where visual evidence is critical. These regions vary across models and tasks but transfer across pruning methods. Based on these findings, we propose Causal Attention Recalibration (CAR), a training-free post-pruning framework consisting of Causal Attention Profiling and Plateau Attention Recalibration. Causal Attention Profiling identifies decoder regions where visual evidence is causally important. Plateau Attention Recalibration applies a flat-top profile within these regions to rebalance visual and textual attention. Experiments across multiple VLMs, benchmarks, and representative pruning methods demonstrate performance gains of up to 2.38% with negligible computational overhead.

1 INTRODUCTION

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in language understanding and reasoning Yang et al. (2025a); Wei et al. (2022); Brown et al. (2020). Building on these advances, vision-language models (VLMs) equip LLMs with visual perception and have achieved strong performance across many tasks, including visual understanding, reasoning, and knowledge-intensive applications Liu et al. (2024a); Bai et al. (2025); Li et al. (2023a).

VLMs typically convert an input image into a sequence of visual tokens and feed them into the language model together with textual tokens Liu et al. (2023; 2024a); Bai et al. (2025); Cha et al. (2024). Hundreds of visual tokens per image substantially increase both the computational and memory cost of decoding Dhouib et al. (2025); Du et al. (2026); Li et al. (2026a).

To improve VLM efficiency, recent studies have explored visual token pruning. Visual token pruning reduces redundant image tokens while preserving a compact and informative subset, thereby shortening the visual sequence processed by the language decoder Chen et al. (2024); Rao et al. (2021). Existing methods can be broadly divided into four categories: attention-based methods select tokens according to attention scores Chen et al. (2024); Yang et al. (2025b); diversity-based methods reduce redundancy among visual features Alvar et al. (2025); Zhang et al. (2026); hybrid-based methods combine multiple selection criteria Kim et al. (2026b); Du et al. (2026); and learning-based methods train dedicated modules to determine which tokens to retain Li et al. (2026b); Ye et al. (2025).

Despite their effectiveness, existing pruning methods primarily focus on which visual tokens to retain, assuming that the decoder will continue to use the surviving visual evidence effectively. However, we find that pruning also changes how the decoder allocates attention between visual and textual information. We characterize this change using two metrics. Average Visual Token Attention

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

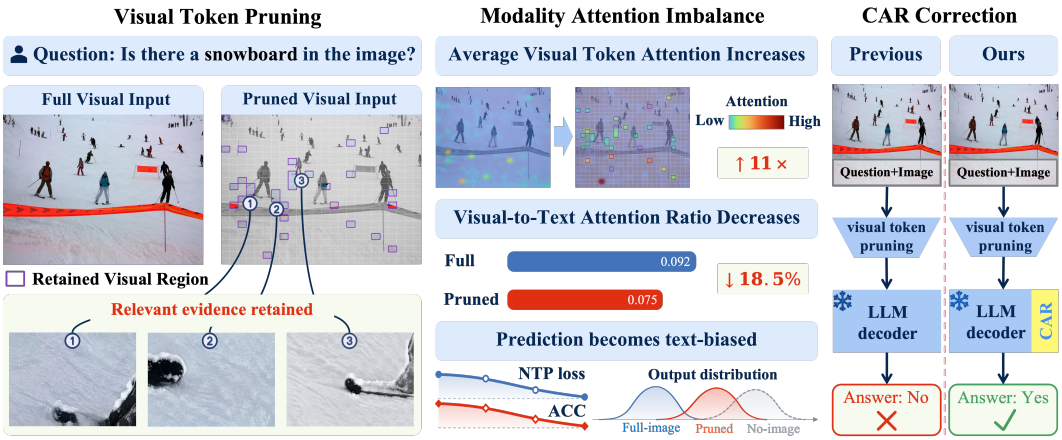


Figure 1: **Illustration of modality attention imbalance after visual token pruning.** With LLaVA-v1.5-7B Liu et al. (2024a), VisionZip Yang et al. (2025b) retains relevant evidence. Although average visual token attention rises by $11\times$, visual-to-text attention ratio still falls by 18.5% and prediction becomes text-biased. CAR recalibrates attention after pruning and restores the correct answer.

(AVTA) measures the average attention received by each retained visual token. The Visual-to-Text Attention Ratio (VTAR) measures aggregate visual attention relative to textual attention. As illustrated in Fig. 1, VisionZip retains the snowboard. Although the decoder partially compensates by increasing AVTA by $11\times$, VTAR still decreases by 18.5%, showing that this compensation is insufficient to preserve the balance between visual and textual attention. Despite retaining the relevant evidence, the model underuses it and produces an incorrect answer. We refer to this phenomenon as **modality attention imbalance**. Our analysis further shows that imbalance becomes more pronounced as pruning becomes more aggressive. Meanwhile, task accuracy and teacher-forced Next-Token Prediction (NTP) loss may even vary in the same direction. The output distribution shifts toward the distribution produced without an image, indicating greater reliance on textual context.

These observations motivate us to investigate how to mitigate modality attention imbalance after visual token pruning. A natural strategy is to rebalance attention between visual and textual tokens across the decoder. However, uniformly reweighting attention across all layers yields limited or even negative gains. To understand why, we examine how visual attention is allocated across decoder layers. We find that visual attention profiles vary more across prompts than images, suggesting that decoder behavior is task-specific. Through controlled causal interventions, we further identify task-specific decoder regions where visual evidence is critical. These regions differ across models and tasks but transfer well across different pruning methods.

Based on these findings, we propose **Causal Attention Recalibration (CAR)**, a training-free framework consisting of **Causal Attention Profiling** and **Plateau Attention Recalibration**. Causal Attention Profiling identifies task-specific decoder regions where visual evidence is causally important. Plateau Attention Recalibration then applies a flat-top profile within these regions to rebalance attention between visual and textual tokens. As illustrated in Fig. 1, CAR is a post-pruning framework and operates during subsequent decoding to mitigate modality attention imbalance.

Extensive experiments across multiple VLMs, benchmarks, and representative pruning methods demonstrate performance gains of up to 2.38%. CAR can be applied to different pruning methods, requires no additional KV-cache entries, and introduces negligible computational overhead.

2 Related Work

2.1 Vision-Language Models

Recent VLMs typically connect pretrained vision encoders to LLMs through lightweight interfaces Alayrac et al. (2022); Dai et al. (2023); Lin et al. (2024). BLIP-2 Li et al. (2023a) introduces a query-based transformer to bridge frozen vision and language models, while LLaVA Liu et al. (2023; 2024a) adopts a simpler projection-based architecture. Qwen-VL Bai et al. (2023) fur-

ther improves visual understanding and multimodal reasoning through stronger visual encoders and higher resolution inputs. Other works, such as Honeybee Cha et al. (2024), explore more efficient visual projectors for compressing and transferring visual representations. Despite these architectural differences, VLMs ultimately rely on the language decoder to integrate visual and textual information. Effective use of visual evidence within the decoder is therefore important Zhang et al. (2025a).

2.2 VISUAL TOKEN PRUNING

Visual token pruning reduces VLM inference cost by retaining only a subset of informative visual tokens. Existing methods differ mainly in how to select this subset. Attention-based methods, such as VisionZip Yang et al. (2025b) and SparseVLM Zhang et al. (2025b), rank tokens using attention signals. Diversity-based approaches, including DivPrune Alvar et al. (2025), reduce redundancy while preserving representative visual content. Hybrid methods, such as ZOO-Prune Kim et al. (2026b) and CoIn Du et al. (2026), combine signals such as importance, redundancy, and semantic relevance. Learning-based approaches, including ATP-LLaVA Ye et al. (2025) and OC-VTP Li et al. (2026b), train dedicated modules that adapt token retention to the input or task. Several studies further address information loss after pruning. Recoverable Compression Chen et al. (2025) uses textual guidance to recover task-relevant visual tokens. GAP Chien et al. (2025) corrects positional misalignment caused by pruning in visual grounding tasks. DSTP Kim et al. (2026a) tracks changes in visual information during decoding and dynamically updates the token set.

These methods mainly reduce performance degradation by recovering, correcting, or updating visual tokens. In contrast, we study how pruning changes the decoder's allocation of attention. We refer to this effect as modality attention imbalance and address it without changing the token selection process. CAR is therefore complementary to existing visual token pruning methods.

3 PRELIMINARIES

Multimodal Decoding. Given an input image I and a textual prompt X, a VLM encodes and maps I into N_0 visual tokens in the language model's embedding space, denoted by $\mathbf{X}_v = \{\mathbf{x}_i^v\}_{i=1}^{N_0}$. Together with the textual embeddings $\mathbf{X}_t$, they are processed by an L-layer autoregressive decoder F_θ. For a generated response $Y = (y_1, \ldots, y_T)$, the next-token distribution at decoding step t is

$$p_\theta(y_t \mid I, X, y_{<t}) = \mathrm{Softmax}\left(\mathbf{W}_o \mathbf{h}_t^{(L)}\right), \qquad \mathbf{h}_t^{(L)} = F_\theta(\mathbf{X}_v, \mathbf{X}_t, y_{<t}), \tag{1}$$

where $y_{<t} = (y_1, \ldots, y_{t-1})$ and $\mathbf{W}_o$ denotes the language model output head.

Visual Token Pruning. Visual token pruning reduces the visual sequence from N_0 tokens to a target budget $K < N_0$ while preserving informative visual evidence. A pruning method selects an index set $\mathcal{S} \subseteq \{1, \ldots, N_0\}$ with $|\mathcal{S}| = K$, producing $\widetilde{\mathbf{X}}_v = \{\mathbf{x}_i^v \mid i \in \mathcal{S}\}$.

Multimodal Attention Allocation. Within the decoder, visual and textual information interact through self-attention. For attention head $h \in \{1, \ldots, H\}$ at layer $l \in \{0, \ldots, L-1\}$, the attention weight from query position q to key position j is

$$Z_{qj}^{(l,h)} = \frac{\mathbf{q}_q^{(l,h)} \mathbf{k}_j^{(l,h)\top}}{\sqrt{d_h}}, \qquad A_{qj}^{(l,h)} = \mathrm{Softmax}_{j \in \mathcal{C}_q}\left(Z_{qj}^{(l,h)}\right), \tag{2}$$

where d_h is the dimension of each attention head and $\mathcal{C}_q$ denotes the set of key positions visible to query q under the causal attention mask. Let $\mathcal{V}$ and $\mathcal{T}$ denote the positions of visual and textual tokens, respectively. We define the attention mass allocated to each modality as

$$A_V^{(l,q)} = \frac{1}{H} \sum_{h=1}^{H} \sum_{j \in \mathcal{V}} A_{qj}^{(l,h)}, \qquad A_T^{(l,q)} = \frac{1}{H} \sum_{h=1}^{H} \sum_{j \in \mathcal{T}} A_{qj}^{(l,h)}. \tag{3}$$

Let $\mathcal{Q}$ denote the answer positions used for evaluation. By aggregating attention over decoder layers and positions in $\mathcal{Q}$, we define the **Average Visual Token Attention (AVTA)** as

$$\mathrm{AVTA} = \frac{1}{L|\mathcal{Q}||\mathcal{V}|} \sum_{l=0}^{L-1} \sum_{q \in \mathcal{Q}} A_V^{(l,q)}. \tag{4}$$

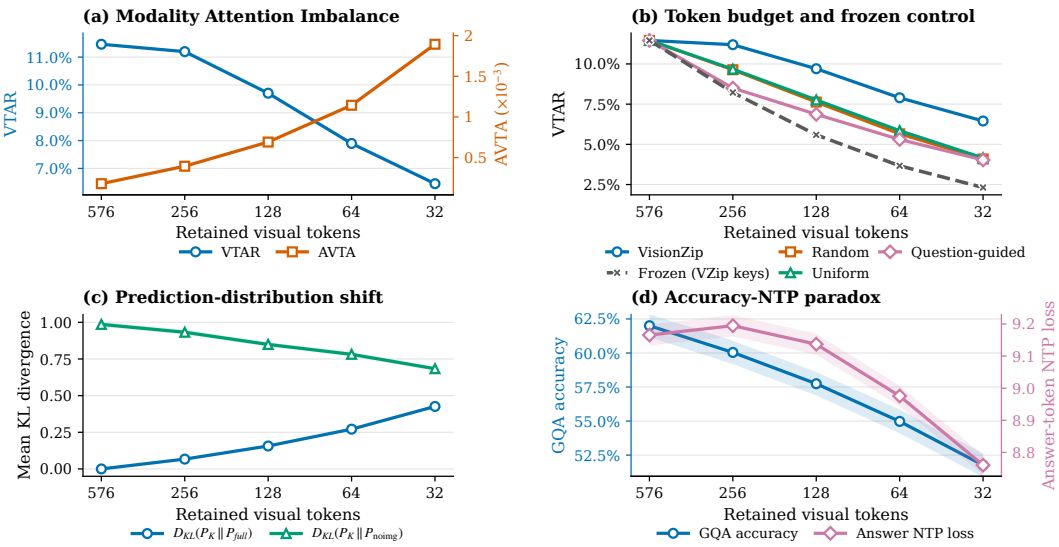


Figure 2: **Analysis of modality attention imbalance after visual token pruning. (a)** Lower token budgets increase AVTA but decrease VTAR. **(b)** VTAR declines across selectors, while VisionZip partially compensates relative to Frozen. **(c)** Predictions shift from the full-image toward the no-image condition. **(d)** Accuracy decreases despite a lower teacher-forced NTP loss.

AVTA measures the average attention received by each visual token. We further define the **Visual-to-Text Attention Ratio (VTAR)** as

$$\mathrm{VTAR} = \frac{\sum_{l=0}^{L-1} \sum_{q \in \mathcal{Q}} A_V^{(l,q)}}{\sum_{l=0}^{L-1} \sum_{q \in \mathcal{Q}} A_T^{(l,q)}}. \tag{5}$$

VTAR measures aggregate visual attention relative to textual attention. Together, AVTA and VTAR are used to characterize modality attention imbalance after visual token pruning.

4 Modality Attention Imbalance

Unless otherwise specified, we conduct the following analyses on LLaVA-v1.5-7B Liu et al. (2024a) with VisionZip Yang et al. (2025b) on the testdev-balanced split of GQA Hudson & Manning (2019).

Pruning induces modality attention imbalance. As shown in Fig. 2(a), VTAR consistently decreases as K shrinks, whereas AVTA increases. Thus, visual tokens are not simply receiving less attention. Instead, attention is redistributed toward the surviving tokens, but this compensation is insufficient to preserve the balance between visual and textual attention. This reflects a cross-modal competition imbalance, which we term modality attention imbalance.

Token budget drives the imbalance despite partial internal compensation. Theoretically, since $|\mathcal{S}| = K < N_0$, fixing the attention logits and recomputing the Softmax yields

$$\frac{\sum_{i \in \mathcal{S}} A_{qi}^{(l,h)}}{\sum_{j \in \mathcal{T}} A_{qj}^{(l,h)}} = \frac{\sum_{i \in \mathcal{S}} \exp\left(Z_{qi}^{(l,h)}\right)}{\sum_{j \in \mathcal{T}} \exp\left(Z_{qj}^{(l,h)}\right)} \leq \frac{\sum_{i=1}^{N_0} \exp\left(Z_{qi}^{(l,h)}\right)}{\sum_{j \in \mathcal{T}} \exp\left(Z_{qj}^{(l,h)}\right)}. \tag{6}$$

Since $A_V^{(l,q)}$ and $A_T^{(l,q)}$ aggregate these weights over visual and textual keys, respectively, Eq. 6 gives $\mathrm{VTAR}(K) \leq \mathrm{VTAR}(N_0)$. Empirically, Fig. 2(b) shows the same decline across selectors: Random samples tokens uniformly, Uniform uses deterministic farthest-point sampling to preserve spatial coverage, and Question-Guided ranks visual tokens by their maximum cosine similarity to question tokens. Together, the theoretical result and the consistent empirical trend across selection strategies identify token budget as the primary driver of modality attention imbalance.

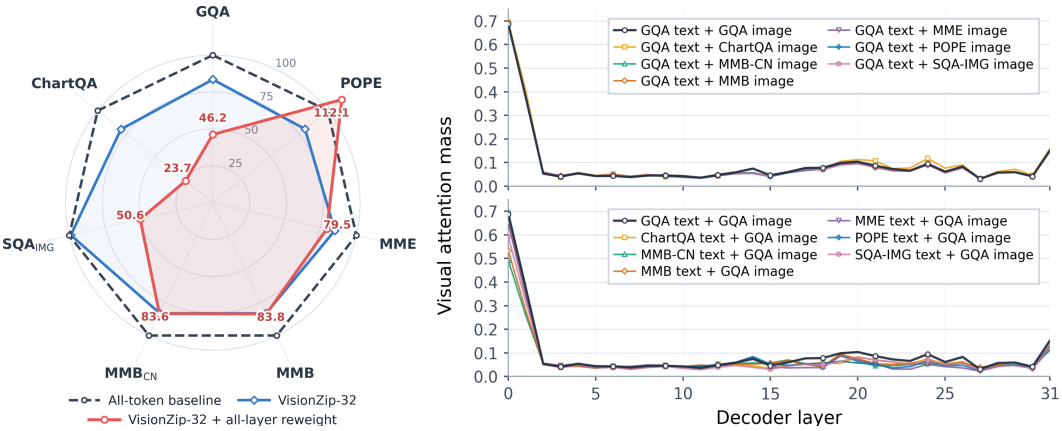


Figure 3: **Motivation for Causal Attention Profiling. Left:** All-layer reweighting provides inconsistent gains across benchmarks. VisionZip-32 denotes VisionZip retaining 32 visual tokens. **Upper right:** Changing images with fixed GQA questions produces nearly identical profiles. **Lower right:** Changing questions with fixed GQA images leads to substantially different profiles.

To further examine the model's response to the structural VTAR decline caused by token removal, we compare VisionZip with a Frozen control. Frozen retains only the visual keys selected by VisionZip but masks attention logits obtained from the full input, without propagating the resulting attention outputs. VisionZip instead propagates the reduced sequence, changing subsequent hidden states and attention logits, such that $\mathbf{h}^{(l)}_{\text{VZip}} \neq \mathbf{h}^{(l)}_{\text{full}}$ and $Z^{(l,h)}_{qj,\text{VZip}} \neq Z^{(l,h)}_{qj,\text{frozen}}$. Figure 2(b) shows $\text{VTAR}_{\text{frozen}}(K) < \text{VTAR}_{\text{VZip}}(K) < \text{VTAR}_{\text{full}}$. The gap between the full and Frozen conditions captures the structural decline caused by token removal. The higher VTAR under VisionZip shows that propagating the reduced sequence changes hidden representations and attention logits in a direction that increases visual attention. This internal response partially offsets the structural decline but does not restore the original VTAR, leaving a residual modality attention imbalance.

Predictions become text-biased. Let P_K, P_{full}, and P_{noimg} denote the output distributions under K retained tokens, full visual input, and no visual input, respectively. KL divergence measures the discrepancy between two distributions, with a smaller value indicating greater similarity. Fig. 2(c) shows that $D_{\text{KL}}(P_K \| P_{\text{full}})$ increases while $D_{\text{KL}}(P_K \| P_{\text{noimg}})$ decreases with stronger pruning. The decline in VTAR is therefore accompanied by a behavioral shift toward the condition without image input, supporting that the imbalance increases the relative influence of textual context on prediction.

Accuracy and NTP loss may even vary in the same direction. For a ground-truth answer $Y = (y_1, \ldots, y_T)$, we measure the teacher-forced NTP loss as

$$\mathcal{L}_{\text{NTP}}(K) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta \left(y_t \mid \widetilde{\mathbf{X}}_v^{(K)}, \mathbf{X}_t, y_{<t} \right). \tag{7}$$

A lower NTP loss means higher probability assigned to ground-truth tokens. Figure 2(d) shows a counterintuitive pattern: stronger pruning reduces both task accuracy and $\mathcal{L}_{\text{NTP}}$. This suggests greater reliance on textual regularities, making teacher-forced prediction easier while weakening visual grounding. Appendix E provides additional analyses across models and benchmarks.

5 Causal Attention Recalibration

Causal Attention Recalibration (CAR) consists of two components: Causal Attention Profiling determines where attention should be recalibrated, while Plateau Attention Recalibration determines how the recalibration should be performed.

5.1 Causal Attention Profiling

A natural way to mitigate modality attention imbalance is to increase visual attention across all decoder layers. However, as shown on the left of Fig. 3, this only improves POPE Li et al. (2023b) but

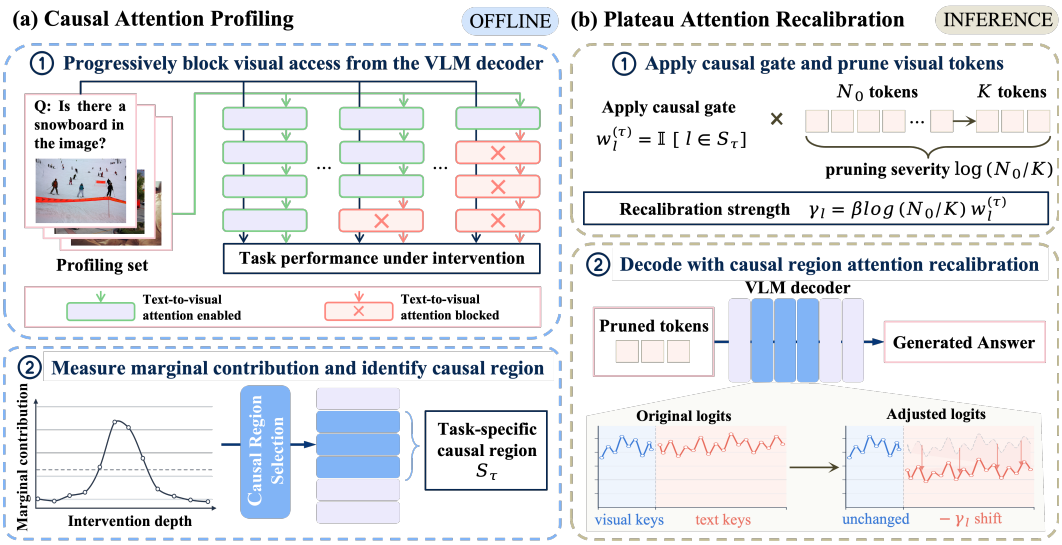


Figure 4: **Overview of Causal Attention Recalibration (CAR). (a) Causal Attention Profiling** identifies the causal region $\mathcal{S}_\tau$ by progressively blocking visual access from the decoder output side. **(b) Plateau Attention Recalibration** combines the resulting flat-top gate with pruning severity to set γ_l and adjust textual attention logits within the region.

degrades the other benchmarks, showing that uniform correction is unreliable. We therefore examine visual attention profiles using a profiling set $\mathcal{D}_\tau$, formed by randomly sampling 500 examples for each task τ. As shown on the right of Fig. 3, changing images with fixed GQA questions barely alters the attention profiles, whereas changing questions with fixed images produces substantial differences. This result suggests that visual attention across decoder layers depends more on the task than on the image, motivating profiling specific to each task.

Based on this observation, we introduce **Causal Attention Profiling**, an offline procedure that identifies decoder regions where visual evidence is causally important for task τ. As illustrated in Fig. 4(a), we first progressively remove visual access from the VLM decoder on the profiling set $\mathcal{D}_\tau$. For an L-layer decoder indexed by $l \in \{0, \ldots, L-1\}$, we consider intervention depths $\mathcal{K} = \{s, 2s, \ldots, L-s\}$, where $k \in \mathcal{K}$ denotes the number of consecutive layers intervened from the output side of the decoder. At depth k, textual-to-visual attention is blocked in the last k layers:

$$\mathcal{I}_k = \{L-k, \ldots, L-1\}, \qquad Z_{qj}^{(l,h)} \leftarrow -\infty, \quad l \in \mathcal{I}_k,\ q \in \mathcal{Q}_{\text{text}},\ j \in \mathcal{V}. \tag{8}$$

For each $k \in \mathcal{K}$, we measure the accuracy under intervention on the profiling set $\mathcal{D}_\tau$, denoted by $P_\tau(k)$. Combined with the baseline accuracy $P_\tau(0)$, these measurements form the task performance trajectory $\{P_\tau(k)\}_{k \in \{0\} \cup \mathcal{K}}$. Increasing k from $k-s$ to k additionally removes visual access from the next s decoder layers. We measure the marginal contribution of these newly intervened layers as

$$C_\tau(k) = P_\tau(k-s) - P_\tau(k). \tag{9}$$

A positive $C_\tau(k)$ indicates that additionally blocking visual access in these layers decreases task performance under the preceding cumulative intervention. The resulting marginal contribution curve is shown in the lower-left of Fig. 4(a). We extract the causal region from this curve using a **Causal Region Selection** algorithm. To reduce the influence of multimodal fusion in the first 20% of decoder layers and output processing in the final block, we exclude both from selection, yielding an interior candidate domain $\mathcal{K}_c \subset \mathcal{K}$. We compute the mean marginal contribution within this domain:

$$\bar{C}_\tau = \frac{1}{|\mathcal{K}_c|} \sum_{k \in \mathcal{K}_c} C_\tau(k), \tag{10}$$

and mark blocks with $C_\tau(k) > \bar{C}_\tau$ as active. We then search for contiguous non-recovery phases with $C_\tau(k) \geq 0$, retaining those supported by multiple active blocks and merging adjacent phases separated by weak recovery. The selected interval $[k_s, k_e]$ is mapped back to decoder depth as

$$\mathcal{S}_\tau = \{L - k_e, \ldots, L - k_s + s - 1\}. \tag{11}$$

As shown in the lower-right of Fig. 4(a), this produces the task-specific causal region used by the subsequent recalibration stage. The complete Causal Attention Profiling procedure and the detailed Causal Region Selection algorithm are provided in Appendix A. Profiling is performed once for each model and task pair, and the resulting region is reused across pruning methods and token budgets.

5.2 Plateau Attention Recalibration

Causal Attention Profiling determines where visual evidence is critical. **Plateau Attention Recalibration** then determines how to adjust attention within the identified region during inference.

For each input, visual token pruning first retains K of the original N_0 tokens. CAR then combines the resulting pruning severity with the task-specific causal region identified offline. As shown in Fig. 4(b), we convert the causal region $\mathcal{S}_\tau$ into a flat-top layer gate

$$w_l^{(\tau)} = \mathbb{I}[l \in \mathcal{S}_\tau]\,. \tag{12}$$

The recalibration strength for the current input is then computed as

$$\gamma_l = \beta \log\left(\frac{N_0}{K}\right) w_l^{(\tau)}, \tag{13}$$

where β is a task-specific scaling coefficient. Thus, γ_l remains zero elsewhere, increases with pruning severity, and vanishes when $K = N_0$. Within the identified layers, we rebalance visual evidence against text while preserving the attention pattern within each modality. A minimum-KL projection under a constraint on textual attention mass yields a shared offset to all textual logits. We provide the derivation in Appendix B. Accordingly, for textual queries, we apply

$$\widetilde{Z}_{qj}^{(l,h)} = Z_{qj}^{(l,h)} - \gamma_l \mathbb{I}[j \in \mathcal{T}], \qquad \widetilde{A}_{qj}^{(l,h)} = \mathrm{Softmax}_j\left(\widetilde{Z}_{qj}^{(l,h)}\right). \tag{14}$$

This preserves relative attention within each modality while increasing the VTAR:

$$\frac{\sum_{j\in\mathcal{V}} \widetilde{A}_{qj}^{(l,h)}}{\sum_{j\in\mathcal{T}} \widetilde{A}_{qj}^{(l,h)}} = e^{\gamma_l} \frac{\sum_{j\in\mathcal{V}} A_{qj}^{(l,h)}}{\sum_{j\in\mathcal{T}} A_{qj}^{(l,h)}} = \left(\frac{N_0}{K}\right)^{\beta w_l^{(\tau)}} \frac{\sum_{j\in\mathcal{V}} A_{qj}^{(l,h)}}{\sum_{j\in\mathcal{T}} A_{qj}^{(l,h)}}. \tag{15}$$

6 Experiments

6.1 Experimental Setup

We evaluate CAR on LLaVA-v1.5-7B Liu et al. (2024a), Qwen2.5-VL-7B Bai et al. (2025), and LLaVA-NeXT-7B Li et al. (2024). We consider four representative pruning methods: VisionZip Yang et al. (2025b), DivPrune Alvar et al. (2025), ZOO-Prune Kim et al. (2026b), and OC-VTP Li et al. (2026b), covering attention-, diversity-, hybrid-, and learning-based selection.

We evaluate on GQA Hudson & Manning (2019), POPE Li et al. (2023b), MME Fu et al. (2026), MMBench (MMB), MMBench-CN ($\mathrm{MMB^{CN}}$) Liu et al. (2024b), ScienceQA-Image ($\mathrm{SQA^{I}}$) Lu et al. (2022), and ChartQA (CQA) Masry et al. (2022). Since these benchmarks use different score scales, each score is normalized by its full-token baseline and averaged as AvgAcc. ΔAcc denotes the change in AvgAcc after applying CAR.

All experiments are conducted on four NVIDIA A100 GPUs. Additional experimental settings are provided in Appendix C, with further ablation and robustness studies in Appendix D.

6.2 Main Results

Tables 1 and 2 show that CAR consistently improves pruning on LLaVA-v1.5 and Qwen2.5-VL across different methods and token budgets. On LLaVA-v1.5, the gain increases from $+0.84\%$ at 128 tokens to $+2.38\%$ at 32 tokens for VisionZip, while on Qwen2.5-VL all eight settings improve, with gains up to $+1.38\%$. This trend is consistent with our earlier observation that more aggressive pruning induces a stronger modality attention imbalance, leaving more room for recalibration. In contrast, LLaVA-NeXT shows only modest gains at 320 tokens and even negative results at 160 tokens, including -0.31% for ZOO-Prune and -1.44% for OC-VTP. These results suggest that CAR is effective when pruning substantially disrupts modality attention balance, but its benefit is not universal. We further examine this applicability boundary in the next subsection.

Table 1: Comparison of different pruning methods with and without CAR on LLaVA-v1.5-7B.

Method	Type	GQA	POPE	MME	MMB	MMB^{CN}	SQA^{I}	CQA	AvgAcc.	ΔAcc.
Upper Bound, 576 Tokens (100%)										
Vanilla	-	61.94	85.84	1866	65.77	57.99	69.61	18.24	100%	-
Retain Averaged 128 Tokens (↓ 77.8%)										
VisionZip	Attention	57.52	83.25	1767	61.61	54.77	68.67	16.96	94.90%	**+0.84%**
+ CAR		57.54	84.05	1769	62.01	54.83	68.47	17.76	**95.74%**	
DivPrune	Diversity	59.29	86.72	1716	60.87	53.96	68.57	17.04	95.18%	**+0.56%**
+ CAR		59.21	87.52	1709	61.48	54.23	68.67	17.40	**95.74%**	
ZOO-Prune	Hybrid	59.55	86.91	1733	61.68	54.36	69.11	17.24	95.94%	**+0.42%**
+ CAR		59.36	87.26	1739	62.28	54.56	68.62	17.60	**96.36%**	
OC-VTP	Learning	57.55	82.91	1760	61.47	54.30	68.27	16.40	94.13%	**+1.41%**
+ CAR		57.43	84.02	1760	62.42	54.56	68.17	17.68	**95.54%**	
Retain Averaged 64 Tokens (↓ 88.9%)										
VisionZip	Attention	55.03	77.22	1718	60.07	52.95	68.86	15.88	91.36%	**+1.20%**
+ CAR		55.16	81.41	1710	60.54	52.89	69.51	16.28	**92.56%**	
DivPrune	Diversity	57.76	85.50	1639	58.99	51.68	67.92	15.84	91.99%	**+1.13%**
+ CAR		57.51	86.47	1629	59.60	52.01	68.27	16.88	**93.12%**	
ZOO-Prune	Hybrid	58.49	85.78	1657	59.66	52.75	68.17	16.76	93.52%	**+0.49%**
+ CAR		58.34	86.79	1662	60.74	52.82	68.47	16.76	**94.01%**	
OC-VTP	Learning	53.95	75.67	1647	58.12	50.54	68.47	16.24	89.49%	**+1.27%**
+ CAR		53.88	79.52	1648	58.52	50.40	68.72	16.92	**90.76%**	
Retain Averaged 32 Tokens (↓ 94.4%)										
VisionZip	Attention	51.73	68.76	1586	54.83	48.26	68.86	14.56	84.85%	**+2.38%**
+ CAR		51.46	78.37	1581	55.23	48.32	69.01	15.52	**87.23%**	
DivPrune	Diversity	55.06	81.54	1594	57.58	48.86	68.72	14.72	88.65%	**+1.34%**
+ CAR		54.87	84.80	1604	57.52	49.26	68.82	15.56	**89.99%**	
ZOO-Prune	Hybrid	55.79	83.11	1638	58.79	50.20	68.52	15.08	90.25%	**+1.10%**
+ CAR		55.64	85.67	1630	58.32	50.60	69.11	15.92	**91.35%**	
OC-VTP	Learning	49.66	61.04	1414	45.91	37.92	67.18	14.72	77.07%	**+1.80%**
+ CAR		49.75	71.04	1404	46.64	38.05	66.58	14.88	**78.87%**	

Table 2: Comparison of different pruning methods with and without CAR on Qwen2.5-VL-7B and LLaVA-NeXT-7B.

	Qwen2.5-VL-7B						LLaVA-NeXT-7B					
Method	GQA	POPE	MMB	SQA^{I}	AvgAcc.	ΔAcc.	GQA	POPE	MMB	SQA^{I}	AvgAcc.	ΔAcc.
	576 Tokens (100%)						*2880 Tokens (100%)*					
Vanilla	58.17	86.45	82.21	80.96	100%	-	64.24	86.45	69.53	70.20	100%	-
	Retain 20% Tokens (↓80%)						*Retain 320 Tokens (11.1%, ↓88.9%)*					
VisionZip	54.21	83.54	78.39	78.48	95.53%	**+1.17%**	58.98	82.02	62.62	67.28	93.15%	**+0.14%**
+ CAR	54.28	84.38	78.86	80.91	**96.70%**		59.01	82.17	62.82	67.33	**93.29%**	
DivPrune	54.38	80.71	74.63	74.02	92.26%	**+1.13%**	60.31	84.30	65.17	67.18	95.21%	**+0.04%**
+ CAR	54.45	81.72	74.90	76.35	**93.39%**		60.34	84.54	65.23	67.03	**95.25%**	
ZOO-Prune	55.50	85.32	76.51	76.70	95.48%	**+0.84%**	60.73	85.77	64.56	68.32	95.98%	**+0.28%**
+ CAR	55.59	85.87	76.71	78.58	**96.32%**		60.69	86.27	64.83	68.47	**96.26%**	
OC-VTP	52.91	79.44	74.03	75.06	91.40%	**+1.21%**	57.53	77.60	60.20	66.44	90.14%	**+0.18%**
+ CAR	52.98	80.80	74.03	77.59	**92.61%**		57.47	78.45	60.00	66.53	**90.32%**	
	Retain 10% Tokens (↓90%)						*Retain 160 Tokens (5.6%, ↓94.4%)*					
VisionZip	51.44	78.48	71.81	74.81	89.74%	**+1.23%**	55.49	74.83	59.80	68.42	89.10%	**+0.03%**
+ CAR	51.34	79.86	72.21	77.24	**90.97%**		55.27	75.17	59.93	68.32	**89.13%**	
DivPrune	50.74	75.05	68.79	71.39	86.47%	**+1.38%**	58.66	81.47	63.02	67.82	93.20%	**+0.17%**
+ CAR	50.98	76.65	68.93	73.87	**87.85%**		58.58	82.06	63.29	67.63	**93.37%**	
ZOO-Prune	52.83	81.93	70.60	73.82	90.66%	**+1.17%**	57.58	82.70	58.99	68.96	**92.09%**	-0.31%
+ CAR	52.93	82.58	70.60	76.85	**91.83%**		57.47	80.89	59.66	69.01	91.78%	
OC-VTP	49.40	72.46	67.58	72.58	85.15%	**+0.22%**	53.11	69.92	52.55	66.63	**83.51%**	-1.44%
+ CAR	49.43	69.34	67.65	76.10	**85.37%**		52.97	65.51	52.28	66.58	82.07%	

6.3 Analysis and Ablation

When is CAR effective? Section 4 shows that although changes in hidden representations and attention logits after pruning partially offset the VTAR decline, they are insufficient to eliminate modality attention imbalance. The limited gains on LLaVA-NeXT motivate us to examine whether the model's internal response to pruning varies across model families and affects the applicability of CAR. We therefore conduct a controlled comparison on SQA^{I} using VisionZip with 10% visual token retention, additionally including InternVL3-8B Zhu et al. (2025), Gemma-3-4B-IT Team et al. (2025), and MiniCPM-V 2.6 Yao et al. (2024). As shown in Table 3, all models exhibit increased

Table 3: Cross-model modality attention imbalance.

Model	ΔAVTA	ΔVTAR
LLaVA-NeXT-7B	+791.94%	-11.33%
LLaVA-v1.5-7B	+614.36%	-29.48%
InternVL3-8B	+602.08%	-30.37%
Gemma-3-4B-IT	+448.80%	-45.97%
Qwen2.5-VL-7B	+322.55%	-57.48%
MiniCPM-V 2.6	+231.77%	-68.14%

Table 4: Ablation of causal attention profiling.

Layer Selection	AvgAcc.	ΔAcc.
Pruned	84.85%	–
Random Region	77.90%	-6.95%
Attention Shift	74.34%	-10.51%
Ours	**87.23%**	**+2.38%**

Table 5: Inference efficiency of CAR on LLaVA-v1.5-7B with VisionZip-32.

Method	Prefill (ms)	E2E (ms)	Peak Mem. (GiB)	Prefill TFLOPs
VisionZip-32	39.28	420.10	13.52	1.632
VisionZip-32 + CAR	39.62	425.15	13.52	1.632

AVTA after pruning, but the resulting modality attention imbalance varies substantially. LLaVA-NeXT shows the strongest compensation, with VTAR decreasing by only 11.33%, while the other models show substantially larger VTAR reductions. This suggests that modality attention imbalance is prevalent but strongly model-dependent. We conjecture that architectural differences, including redundancy among visual tokens and the decoder's capacity to redistribute attention, contribute to this variation. A stronger internal response leaves less residual imbalance for CAR to correct and may increase the risk of overcorrection. CAR is therefore most effective when the model cannot sufficiently preserve the balance between visual and textual attention after pruning. Understanding the architectural factors governing this behavior remains an important direction for future work.

Effect of layer selection. We evaluate layer selection on LLaVA-v1.5-7B with VisionZip-32 over the same seven benchmarks used in our main evaluation. All variants use the same recalibration setting and equal-length regions, differing only in the selection criterion. For each task, Random Region randomly selects a contiguous decoder interval with the same number of layers as the region identified by Causal Attention Profiling. Attention Shift Region selects an equal-length contiguous interval exhibiting the largest average pruning-induced change in visual attention. As shown in Table 4, Random Region and Attention Shift Region reduce AvgAcc. to 77.90% and 74.34%, whereas Causal Attention Profiling improves it from 84.85% to 87.23%. This shows that large pruning-induced attention changes do not necessarily indicate causal importance, and recalibration is most effective when applied to causally identified regions. The full table is provided in Appendix D.1.

Efficiency. CAR introduces negligible inference overhead. As shown in Table 5, on LLaVA-v1.5-7B with VisionZip-32, CAR increases prefill and end-to-end latency by only 0.9% and 1.2%, respectively, while peak GPU memory and prefill computation remain unchanged. This is because CAR applies only a lightweight logit offset within selected decoder layers, without adding visual tokens, model parameters, or KV-cache entries. Causal Attention Profiling is performed offline once for each model–task pair and incurs no inference-time cost.

7 Conclusion

We show that visual token pruning affects not only which visual evidence remains, but also how the decoder balances visual and textual information. As pruning becomes stronger, average visual token attention increases while the visual-to-text attention ratio decreases, forming a modality attention imbalance. To address this, we propose Causal Attention Recalibration (CAR), a training-free framework that causally identifies task-specific decoder regions and selectively recalibrates attention within them according to pruning severity. CAR improves multiple pruning methods on LLaVA-v1.5 and Qwen2.5-VL with negligible inference overhead, while its limited gains on LLaVA-NeXT reveal that recalibration is most effective when pruning substantially disrupts the original modality balance. Our results suggest that future VLM compression methods should consider not only token selection, but also how retained visual evidence is used during decoding.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, draft portions of the manuscript, retrieve and identify potentially relevant literature, and assist in proving and checking mathematical claims. All suggested references were manually verified against the original publications for their existence, relevance, and accurate representation. In particular, the authors independently rederived and verified all mathematical arguments, including the relationship between the visual token budget and VTAR. AI-assisted drafts were critically reviewed, substantially revised, and incorporated only after their technical accuracy, consistency with the experimental evidence, and originality had been checked by the authors. Generative AI was not used to generate experimental data, execute experiments, or autonomously determine the reported conclusions. We take full responsibility for the final content of this work, including all text, claims, derivations, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a training-free method for improving the reliability of visually pruned vision-language models. Our experiments use publicly available pretrained models and established benchmarks. We do not collect new data or conduct studies involving human participants. By preserving model performance under aggressive visual token pruning with negligible overhead, CAR may support more resource-efficient VLM inference. However, CAR does not eliminate biases, hallucinations, privacy concerns, or other limitations inherited from the underlying models and evaluation data. Moreover, improved efficiency may facilitate broader deployment in settings where these risks remain. CAR should therefore be used in accordance with the licenses and intended uses of the underlying models and datasets, with application-specific evaluation and appropriate human oversight in high-stakes scenarios.

REPRODUCIBILITY STATEMENT

We provide the formulation and implementation details of CAR in the main text. Appendix A provides the complete Causal Attention Profiling procedure, Appendix B derives the recalibration rule, Appendix C details the experimental setup and task-specific configurations, and Appendix D provides additional ablation and robustness studies. Additional analyses and qualitative examples are provided in Appendices E and F. All pretrained models, pruning methods, and benchmarks used in our experiments are publicly available and appropriately cited. Upon publication, we will publicly release the source code, configuration files, and evaluation scripts required to reproduce the reported results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:7400–7426, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13817–13827. IEEE, 2024.

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, 2024.

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

Tzu-Chun Chien, Chieh-Kai Lin, Shiang-Feng Tsai, Ruei-Chi Lai, Hung-Jen Chen, and Min Sun. Grounding-aware token pruning: Recovering from drastic performance drops in visual grounding caused by pruning, 2025. URL https://arxiv.org/abs/2506.21873.

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.

Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and clustering-based token reduction for faster visual language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14582–14592. IEEE, 2025.

Chenxi Du, Yongheng Deng, Jiani Liu, Yujia Zhang, Xi Chen, and Ju Ren. Coin: Coverage and informativeness-guided token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10492–10501, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Jiwan Kim, Kibum Kim, Wonjoong Kim, Byung-Kwan Lee, and Chanyoung Park. Why and when visual token pruning fails? a study on relevant visual information shift in mllms decoding, 2026a. URL https://arxiv.org/abs/2604.12358.

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, 2026b.

Ao Li, Yuxiang Duan, Jinghui Zhang, Congbo Ma, Yutong Xie, Gustavo Carneiro, Mohammad Yaqub, and Hu Wang. Transprune: Token transition pruning for efficient large vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39529–39538, 2026a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26679–26689. IEEE, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael

Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. URL https://arxiv.org/abs/2408.01800.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24972–24982. IEEE, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9382–9391, 2025.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *Advances in Neural Information Processing Systems*, 38:25438–25468, 2026.

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, 2025a.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume

267 of *Proceedings of Machine Learning Research*, pp. 74840–74857. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/zhang25s.html.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025c.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A Details of Causal Attention Profiling

A.1 Causal Attention Profiling

We provide additional implementation details for **Causal Attention Profiling**. For each task τ, we construct a profiling set $\mathcal{D}_\tau$ by randomly sampling 500 examples. Starting from the output side of the decoder, we progressively block textual-to-visual attention, evaluate the resulting task performance, and derive a marginal contribution curve. The **Causal Region Selection** algorithm then extracts a contiguous decoder region with a sustained positive contribution under intervention. The complete procedure is summarized in Alg. 1.

Algorithm 1 Causal Attention Profiling

Require: Model f_θ; profiling set $\mathcal{D}_\tau$; decoder depth L; intervention step s; early exclusion ratio ρ; excluded output blocks b_{out}
Ensure: Causal region $\mathcal{S}_\tau$

$\mathcal{K} \leftarrow \{s, 2s, \ldots, L-s\}$
Evaluate the unmodified model to obtain $P_\tau(0)$
for $k \in \mathcal{K}$ in ascending order **do**
    Block textual-to-visual attention in the last k decoder layers
    Evaluate the intervened model to obtain $P_\tau(k)$
    $C_\tau(k) \leftarrow P_\tau(k-s) - P_\tau(k)$
end for
$\ell_{\min} \leftarrow \max(s, \lfloor \rho L/s \rfloor\, s)$
$k_{\min} \leftarrow s(b_{\text{out}}+1)$
$k_{\max} \leftarrow L - \ell_{\min}$
$\mathcal{K}_{\text{c}} \leftarrow \{k \in \mathcal{K} : k_{\min} \le k \le k_{\max}\}$
if $\mathcal{K}_{\text{c}} = \emptyset$ **then**
    return $\emptyset$
end if
$\bar{C}_\tau \leftarrow \dfrac{1}{|\mathcal{K}_{\text{c}}|} \displaystyle\sum_{k \in \mathcal{K}_{\text{c}}} C_\tau(k)$
$[k_{\text{s}}, k_{\text{e}}] \leftarrow \text{CausalRegionSelection}\big(P_\tau, C_\tau, \mathcal{K}_{\text{c}}, \bar{C}_\tau, s\big)$
if no valid interval is returned **then**
    return $\emptyset$
end if
$\mathcal{S}_\tau \leftarrow \{L - k_{\text{e}}, \ldots, L - k_{\text{s}} + s - 1\}$
return $\mathcal{S}_\tau$

Profiling settings. Profiling is performed once for each model–task pair under a representative aggressive VisionZip Yang et al. (2025b) setting. As summarized in Table 6, we use VisionZip-32 for LLaVA-v1.5-7B, 20% retention for Qwen2.5-VL-7B, and VisionZip-640 for LLaVA-NeXT-7B. The resulting causal region is reused across pruning methods and token budgets without additional profiling.

Table 6: Pruning settings used for Causal Attention Profiling.

Model	Method	Profiling Budget
LLaVA-v1.5-7B	VisionZip	32 tokens
Qwen2.5-VL-7B	VisionZip	20% retention
LLaVA-NeXT-7B	VisionZip	640 tokens

A.2 Causal Region Selection

The marginal contribution curve $C_\tau(k)$ may contain isolated performance drops and temporary recoveries, making a single threshold insufficient to identify a stable causal region. We therefore apply the selection procedure in Alg. 2.

We first divide the candidate domain $\mathcal{K}_c$ into consecutive non-recovery sequences satisfying $C_\tau(k) \geq 0$. Within each sequence, blocks with $C_\tau(k) > \bar{C}_\tau$ are marked as *active*, and only sequences containing at least two active blocks are retained. Each resulting phase starts at the earliest intervention depth before its first active block where performance falls below the unmodified baseline. It ends when the consecutive positive contributions following its last active block cease. Adjacent phases are merged when the accumulated recovery between them is weak. We then select the phase spanning the most intervention blocks, breaking ties by the smallest k_s. If no supported phase is found, we select the eligible block with the largest marginal contribution.

Algorithm 2 Causal Region Selection

Require: Performance curve P_τ; marginal curve C_τ; candidate domain $\mathcal{K}_c$; mean $\bar{C}_\tau$; step size s
Ensure: Intervention interval $[k_s, k_e]$ or $\emptyset$

if $\bar{C}_\tau \leq 0$ **then**
  return $\emptyset$
end if
Form sequences $\{\mathcal{R}_i\}$ from consecutive depths satisfying $C_\tau(k) \geq 0$
  where adjacent depths differ by s
$\mathcal{H} \leftarrow \emptyset$
for each sequence $\mathcal{R}_i$ **do**
  $\mathcal{A}_i \leftarrow \{k \in \mathcal{R}_i : C_\tau(k) > \bar{C}_\tau\}$
  if $|\mathcal{A}_i| < 2$ **then**
    continue
  end if
  $k_{\text{first}} \leftarrow \min \mathcal{A}_i, \quad k_{\text{last}} \leftarrow \max \mathcal{A}_i$
  $\mathcal{B}_i \leftarrow \{k \in \mathcal{R}_i : k \leq k_{\text{first}},\ P_\tau(k) < P_\tau(0)\}$
  $k_s \leftarrow \min \mathcal{B}_i$ if $\mathcal{B}_i \neq \emptyset$, otherwise $k_s \leftarrow k_{\text{first}}$
  $k_e \leftarrow k_{\text{last}}$
  while $k_e + s \in \mathcal{R}_i$ and $C_\tau(k_e + s) > 0$ **do**
    $k_e \leftarrow k_e + s$
  end while
  Add $[k_s, k_e]$ to $\mathcal{H}$
end for
if $\mathcal{H} \neq \emptyset$ **then**
  Sort $\mathcal{H}$ by increasing k_s
  For adjacent phases $[a, b]$ and $[c, d]$, define
  $$\text{Rec}([a, b], [c, d]) = \sum_{\substack{k \in \mathcal{K}_c \\ b < k < c}} \max(-C_\tau(k), 0)$$
  while there exist adjacent $[a, b]$ and $[c, d]$ satisfying $\text{Rec}([a, b], [c, d]) \leq \bar{C}_\tau$ **do**
    Replace the leftmost such pair with $[a, d]$
  end while
  return the phase spanning the most intervention blocks
   breaking ties by the smallest k_s
end if
$k^* \leftarrow$ the smallest maximizer of $C_\tau(k)$ over $k \in \mathcal{K}_c$
return $[k^*, k^*]$

Search constraints. We exclude both ends of the decoder so that region selection focuses on task-specific visual dependence rather than broad effects from multimodal fusion or output processing. Shallow layers support early visual and textual integration, and disrupting their cross-modal attention can substantially reduce performance Yin et al. (2025); Zhang et al. (2025c). Causal tracing also identifies early self-attention and MLP blocks as important for visual information storage and transfer Basu et al. (2024). Higher layers further refine the integrated representation for answer generation Zhang et al. (2025c). Intervening in these boundary layers may therefore disrupt essential model functions and degrade performance, rather than reveal task-specific dependence on visual evidence.

Let ρ denote the excluded fraction of early layers and b_{out} the number of excluded output blocks. We define

$$\begin{aligned} \ell_{\min} &= \max\left(s, \left\lfloor \frac{\rho L}{s} \right\rfloor s\right), \\ k_{\min} &= s(b_{\text{out}} + 1), \qquad k_{\max} = L - \ell_{\min}, \\ \mathcal{K}_{\text{c}} &= \{k \in \mathcal{K} : k_{\min} \le k \le k_{\max}\}. \end{aligned} \tag{16}$$

We set the intervention step to $s = 2$ solely to obtain the marginal contribution curve more efficiently. We further set $\rho = 0.2$ and $b_{\text{out}} = 1$, excluding approximately the first 20% of decoder layers and the intervention block nearest the decoder output. The reference mean is computed within the same candidate domain:

$$\bar{C}_\tau = \frac{1}{|\mathcal{K}_{\text{c}}|} \sum_{k \in \mathcal{K}_{\text{c}}} C_\tau(k). \tag{17}$$

This prevents the excluded boundaries from affecting the selection threshold. All profiling and selection settings are fixed across models and tasks, as summarized in Table 7.

Table 7: Shared settings for Causal Attention Profiling and Region Selection.

Setting	**Value**
Intervention step s	2
Early layer exclusion ρ	20% (block aligned)
Excluded output blocks b_{out}	1
Minimum active blocks	2
Strongest block fallback	Enabled

Phase construction and merging. A non-recovery sequence consists of consecutive depths in $\mathcal{K}_{\text{c}}$ that satisfy $C_\tau(k) \ge 0$, where adjacent depths differ by s. Within each sequence, blocks satisfying $C_\tau(k) > \bar{C}_\tau$ are marked as active. We retain only sequences containing at least two active blocks, which prevents isolated fluctuations from defining a causal region.

A phase begins at the earliest depth within the sequence, no later than its first active block, where $P_\tau(k) < P_\tau(0)$. If no such depth exists, it begins at the first active block. The phase extends through the consecutive positive contributions following its last active block.

For adjacent phases $[a, b]$ and $[c, d]$, we define their accumulated recovery as

$$\text{Rec}([a, b], [c, d]) = \sum_{\substack{k \in \mathcal{K}_{\text{c}} \\ b < k < c}} \max(-C_\tau(k), 0)\,. \tag{18}$$

The phases are merged when $\text{Rec}([a, b], [c, d]) \le \bar{C}_\tau$. After repeatedly merging eligible adjacent phases, we select the phase spanning the most intervention blocks. If several phases span the same number of intervention blocks, we select the one with the smallest k_{s}.

Fallback. If $\bar{C}_\tau > 0$ but no phase contains sufficient active support, we select the eligible block with the largest positive marginal contribution. This preserves the strongest localized causal effect when no supported phase is found. If $\bar{C}_\tau \le 0$ or every eligible contribution is nonpositive, no causal region is returned. Once identified, $\mathcal{S}_\tau$ is reused across pruning methods and token budgets for the same model and task pair.

B Derivation of Minimum-KL Recalibration

We provide the derivation underlying the logit offset form used in Plateau Attention Recalibration. Consider a fixed textual query at a single decoder layer and attention head, and omit the corresponding indices for clarity. Let

$$p_j = \text{Softmax}_j(z_j) \tag{19}$$

denote the original attention distribution over visual and textual keys. We seek a new distribution $\mathbf{r}$ that reduces the total attention assigned to textual tokens while minimally perturbing the original distribution:

$$\min_{\mathbf{r}} D_{\mathrm{KL}}(\mathbf{r}\|\mathbf{p}) \quad \text{s.t.} \quad \sum_{j\in\mathcal{T}} r_j \le c, \qquad \sum_j r_j = 1, \qquad r_j \ge 0. \tag{20}$$

Here, $\mathcal{T}$ denotes the set of textual keys and c controls the maximum allowed textual attention mass.

When the textual mass constraint is active, the Lagrangian is

$$\mathcal{L}(\mathbf{r},\lambda,\nu) = \sum_j r_j \log\frac{r_j}{p_j} + \lambda\left(\sum_{j\in\mathcal{T}} r_j - c\right) + \nu\left(\sum_j r_j - 1\right), \tag{21}$$

where $\lambda \ge 0$ and ν are the Lagrange multipliers for the textual mass and normalization constraints, respectively. Setting the derivative with respect to r_j to zero gives

$$\log\frac{r_j}{p_j} + 1 + \nu + \lambda\mathbb{I}[j\in\mathcal{T}] = 0, \tag{22}$$

and therefore

$$r_j \propto p_j \exp(-\lambda\mathbb{I}[j\in\mathcal{T}])\,. \tag{23}$$

Since $p_j \propto \exp(z_j)$, the normalized solution can be written as

$$r_j^* = \mathrm{Softmax}_j\left(z_j - \lambda\mathbb{I}[j\in\mathcal{T}]\right). \tag{24}$$

Let

$$P_{\mathcal{T}} = \sum_{j\in\mathcal{T}} p_j, \qquad P_{\mathcal{V}} = 1 - P_{\mathcal{T}}. \tag{25}$$

When $0 < c < P_{\mathcal{T}}$, the constraint is active and

$$\sum_{j\in\mathcal{T}} r_j^* = \frac{e^{-\lambda}P_{\mathcal{T}}}{P_{\mathcal{V}} + e^{-\lambda}P_{\mathcal{T}}} = c, \tag{26}$$

which gives

$$\lambda = \log\left(\frac{P_{\mathcal{T}}(1-c)}{cP_{\mathcal{V}}}\right) \ge 0. \tag{27}$$

If $c \ge P_{\mathcal{T}}$, the constraint is inactive, yielding $\lambda = 0$ and $\mathbf{r}^* = \mathbf{p}$. Thus, the minimum-KL correction preserves the relative logits, and hence the relative attention ratios, within each modality while applying a shared negative offset to the textual logits.

This derivation determines the form of the minimum-KL correction, namely a shared offset applied to textual logits, but not the offset magnitude used by CAR. To avoid solving a separate constrained problem for every query and attention head, Plateau Attention Recalibration instantiates this offset as

$$\gamma_l = \beta\log\left(\frac{N_0}{K}\right) w_l^{(\tau)}, \tag{28}$$

where β is a scaling coefficient fixed for each model and task pair. Here, $w_l^{(\tau)}$ restricts recalibration to the causal region, while $\log(N_0/K)$ adapts its strength to pruning severity.

C Experimental Details

This appendix provides additional details on benchmark protocols, offline profiling, and CAR configurations for each task.

C.1 Benchmarks and Evaluation Protocols

We evaluate CAR on seven widely used multimodal benchmarks covering visual question answering, object hallucination, general multimodal understanding, scientific reasoning, and chart understanding. For each benchmark, we use the evaluation split and standard metric listed in Table 8.

GQA. GQA Hudson & Manning (2019) evaluates compositional visual question answering, with particular emphasis on object attributes, relations, and visual reasoning. We use the testdev-balanced split and report accuracy.

POPE. POPE Li et al. (2023b) evaluates object hallucination by asking whether specific objects are present in an image. We evaluate on the test split and report the F1 score.

MME. MME Fu et al. (2026) provides a comprehensive evaluation of multimodal perception and cognition. We evaluate on the test set and report the sum of the perception and cognition scores.

MMBench. MMBench Liu et al. (2024b) evaluates a broad range of multimodal perception and reasoning abilities through multiple-choice questions. We use the English test split and report accuracy.

MMBench-CN. MMBench-CN Liu et al. (2024b) is the Chinese counterpart of MMBench and evaluates multimodal understanding and reasoning in Chinese. We use the Chinese test split and report accuracy.

ScienceQA-Image. ScienceQA Lu et al. (2022) is a multimodal science question-answering benchmark covering diverse scientific topics and reasoning skills. We evaluate on the image subset of the test split, denoted by $\mathrm{SQA}^{\mathrm{I}}$, and report accuracy.

ChartQA. ChartQA Masry et al. (2022) evaluates visual and numerical reasoning over charts through question answering. We evaluate on the test split and report accuracy.

Table 8: Benchmarks and evaluation protocols.

Benchmark	Split/Subset	Metric
GQA	Testdev-balanced	Accuracy
POPE	Test	F1
MME	Test	Perception + Cognition
MMB	English test	Accuracy
$\mathrm{MMB}^{\mathrm{CN}}$	Chinese test	Accuracy
$\mathrm{SQA}^{\mathrm{I}}$	Test image subset	Accuracy
CQA	Test	Accuracy

C.2 Offline Profiling Setup and CAR Configurations

Profiling setup. For each task τ, we construct a profiling set $\mathcal{D}_\tau$ by randomly sampling 500 examples. We use the training splits of ChartQA, GQA, and $\mathrm{SQA}^{\mathrm{I}}$, and the development splits of MMBench and MMBench-CN. Since MME and POPE provide only official test splits, their profiling examples are sampled from these splits.

The profiling set represents the distribution of question types in each benchmark. It is used only to identify a shared causal region $\mathcal{S}_\tau$. Ground truth answers are used during offline profiling to measure task performance under causal interventions and compare candidate configurations.

Once selected, $\mathcal{S}_\tau$ is fixed for each model and task pair and applied uniformly to all evaluation examples. No example-specific region or coefficient is estimated, and no ground truth answer is accessed during inference. The profiling set is not used to update model parameters, select visual tokens, or condition individual predictions.

For MME and POPE, the main evaluation uses the complete official test split because no training or validation split is available. To examine potential overlap effects, we additionally evaluate both benchmarks after removing all profiling examples. As shown in Table 9, CAR retains comparable gains, indicating that the improvements are not driven by overlap with the profiling set.

Table 9: Evaluation after excluding the profiling examples. Disjoint evaluation removes all examples involved in offline profiling. MME reports its official aggregate score, while POPE reports F1. Δ denotes the difference between CAR and the pruning baseline.

Benchmark	Evaluation set	Baseline	+ CAR	Δ
MME	Full test	1586	1581	-5
MME	Disjoint	1577	**1591**	$+14$
POPE	Full test	68.76	**78.37**	$+9.61$
POPE	Disjoint	69.03	**78.27**	$+9.24$

Table 10 reports the profiling setting, causal region $\mathcal{S}_\tau$, and scaling coefficient β for each model and task pair. Both $\mathcal{S}_\tau$ and β are fixed and reused across pruning methods and token budgets. Thus, the listed profiling setting does not restrict CAR to a particular pruning configuration.

Table 10: Task-specific CAR configurations. Decoder layers are indexed from 0. Each causal region is identified once under the listed VisionZip profiling setting and reused across pruning methods and token budgets.

Model	**Benchmark**	**Profiling Setting**	$\mathcal{S}_\tau$	β
LLaVA-v1.5-7B	GQA	VisionZip-32	6–21	0.5
	POPE		8–19	2.0
	MME		6–15	0.5
	MMB		6–13	1.0
	$\mathrm{MMB}^{\mathrm{CN}}$		6–13	0.5
	$\mathrm{SQA}^{\mathrm{I}}$		6–11	1.0
	CQA		24–29	1.0
Qwen2.5-VL-7B	GQA	VisionZip-20%	8–25	0.5
	POPE		12–19	2.0
	MMB		4–17	1.0
	$\mathrm{SQA}^{\mathrm{I}}$		4–11	2.0
LLaVA-NeXT-7B	GQA	VisionZip-640	6–13	0.5
	POPE		6–7	1.5
	MMB		6–9	1.5
	$\mathrm{SQA}^{\mathrm{I}}$		6–11	0.5

C.3 Stability of Causal Attention Profiling

Causal Attention Profiling uses a randomly sampled profiling set of 500 examples for each task. We examine whether this sample size is sufficient to obtain stable causal regions by repeating the profiling procedure with different random subsets. Table 11 reports the regions obtained from six independent profiling runs on LLaVA-v1.5-7B. Despite variation at the region boundaries, the identified regions exhibit substantial overlap, with mean pairwise IoUs ranging from 0.793 to 0.867 across tasks.

To summarize the regions across runs, we define the consensus region as the set of layers selected by at least four of the six runs. The resulting regions are 6–21 for GQA, 8–19 for POPE, 6–13 for MMBench, and 6–11 for ScienceQA-Image, which coincide with the regions used in our main experiments. Moreover, every layer within these regions is supported by at least 5/6, 6/6, 4/6, and 6/6 runs, respectively. These results indicate that a profiling set of 500 examples is sufficient to recover a stable task-specific causal region, with most variation occurring only near the region boundaries.

Table 11: Stability of Causal Attention Profiling across different profiling subsets on LLaVA-v1.5-7B. Each run uses an independently sampled profiling set of 500 examples. The consensus region contains layers selected by at least four of the six runs. Support denotes the minimum selection frequency of any layer within the consensus region.

Task	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Consensus	Support	Pairwise IoU
GQA	8–21	6–21	6–21	6–25	6–21	6–19	**6–21**	5/6	0.853
POPE	8–23	8–19	8–19	8–23	8–21	8–19	**8–19**	6/6	0.855
MMB	6–13	6–13	6–13	6–11	6–13	6–11	**6–13**	4/6	0.867
SQA^I	6–11	6–13	6–13	6–11	6–15	6–11	**6–11**	6/6	0.793
Average pairwise IoU									**0.842**

D Ablation and Robustness Studies

D.1 Effect of Layer Selection

We compare three layer selection strategies while holding the recalibration rule, scaling coefficient, and region length fixed for each task. Random Region randomly selects a contiguous interval with the same number of layers as the region identified by Causal Attention Profiling. Attention Shift Region selects an equal-length contiguous interval exhibiting the largest average pruning-induced shift in visual attention. Causal Attention Profiling instead selects the interval according to the task-performance degradation caused by progressively blocking textual-to-visual attention. Thus, the variants differ only in how the decoder layers to be recalibrated are selected.

As shown in Table 12, Random Region and Attention Shift Region reduce AvgAcc relative to the pruned baseline, whereas recalibration over the causally profiled region improves AvgAcc by 2.38 percentage points. These results indicate that the magnitude of a pruning-induced attention shift is not, by itself, a reliable indicator of where recalibration should be applied.

Table 12: Comparison of layer selection strategies on LLaVA-v1.5-7B with VisionZip-32. All recalibration variants use the same recalibration rule, scaling coefficient, and region length and differ only in layer selection. AvgAcc is computed by normalizing each benchmark score by the corresponding full-token baseline and then averaging across benchmarks. ΔAcc denotes the change in AvgAcc relative to the pruned baseline.

Layer Selection	GQA	POPE	MME	MMB	MMB^{CN}	SQA^I	CQA	AvgAcc.	ΔAcc.
Full-token baseline	61.94	85.84	1866	65.77	57.99	69.61	18.24	100%	–
Pruned	51.73	68.76	1586	54.83	48.26	68.86	14.56	84.85%	–
Random Region	51.65	65.45	1536	55.03	47.92	68.82	6.96	77.90%	−6.95%
Attention Shift Region	51.38	41.18	1550	55.23	48.05	68.47	7.52	74.34%	−10.51%
Causal Attention Profiling	51.46	78.37	1581	55.23	48.32	69.01	15.52	**87.23%**	**+2.38%**

D.2 Effect of Boundary Exclusion

Table 13 examines the boundary constraints used in Causal Region Selection. The baseline excludes the first 20% of decoder layers and the final output-side intervention block. For ChartQA, allowing the output boundary expands the selected region from layers 24–29 to 24–31 and slightly reduces performance. For the other benchmarks, allowing early layers shifts the region boundary from layer 6 to layer 4 and consistently degrades performance. The largest reduction occurs on MME, where the score drops from 1581 to 1529. These results support excluding boundary layers whose effects are more likely to reflect early multimodal fusion or final output processing than the task-specific use of visual evidence.

D.3 Task Specificity of Causal Regions

Causal Attention Profiling is designed to identify a causal decoder region for each task rather than relying on a universal layer interval. To examine the importance of this task calibration, we use the region selected on GQA, $\mathcal{S}_{\mathrm{GQA}} = \{6, \ldots, 21\}$, for every other benchmark. This Transfer variant

Table 13: Ablation of boundary exclusion in Causal Region Selection. The baseline excludes the first 20% of decoder layers and the final output-side intervention block. Each ablation allows the indicated boundary to enter the search. Δ is measured relative to the baseline.

Benchmark	Baseline region	Baseline	Included boundary	Ablated region	Ablated (Δ)
ChartQA	24–29	**15.52**	Output	24–31	15.44 (-0.08)
MME	6–15	**1581**	Early	4–15	1529 (-52)
SQA^{I}	6–11	**69.01**	Early	4–11	67.63 (-1.38)
$\text{MMBench}^{\text{CN}}$	6–13	**48.32**	Early	4–13	48.26 (-0.06)
$\text{MMBench}^{\text{EN}}$	6–13	**55.23**	Early	4–13	54.90 (-0.33)

Table 14: Ablation of task-specific causal region selection on LLaVA-v1.5-7B with VisionZip at 32 retained tokens. Transfer applies the GQA region $\{6, \ldots, 21\}$ to every target benchmark, while Ours identifies a separate region for each task. AvgAcc. is computed over the six benchmarks, and ΔAcc. is measured relative to the pruning baseline. Bold indicates the best AvgAcc. and ΔAcc.

Method	**Region**	**POPE**	**MME**	**MMB**	$\textbf{MMB}^{\textbf{CN}}$	$\textbf{SQA}^{\textbf{I}}$	**CQA**	**AvgAcc.**	**ΔAcc.**
Baseline	None	68.76	1586	54.83	48.26	68.86	14.56	85.07%	–
Transfer	GQA: 6–21	77.23	1488	53.76	48.32	63.71	13.48	83.37%	-1.70%
Ours	Task-specific	78.37	1581	55.23	48.32	69.01	15.52	**87.92%**	**+2.85%**

replaces each task's profiled region with the fixed GQA region while keeping the remaining CAR settings unchanged. Ours instead uses the task-specific region $\mathcal{S}_\tau$ identified by Causal Attention Profiling.

As shown in Table 14, the GQA region does not transfer uniformly across tasks. Although Transfer substantially improves POPE and slightly improves $\text{MMBench}^{\text{CN}}$, it degrades MME, MMBench, SQA^{I}, and CQA relative to the pruning baseline, reducing AvgAcc. by 1.70 points. In contrast, task-specific selection improves AvgAcc. by 2.85 points and outperforms Transfer by 4.55 points. These results show that the causal region depends on the task and that identifying the appropriate decoder layers is essential to CAR.

D.4 Sensitivity to Recalibration Strength

We examine the sensitivity of CAR to the recalibration strength β on LLaVA-v1.5-7B with VisionZip. For each benchmark, we keep the causal region $\mathcal{S}_\tau$ fixed and vary only β, ensuring that the observed differences arise from recalibration strength rather than layer selection. We evaluate each setting with 32, 64, and 128 retained visual tokens and report the average across these budgets.

As shown in Table 15, an insufficient β may fail to adequately correct the attention imbalance, whereas an excessively large β can suppress textual attention too strongly and disrupt the original decoding behavior. The appropriate strength is task dependent. For example, increasing β from 1.5 to 2.0 improves the average POPE score, while a further increase to 2.5 reduces it. Overall, the selected values achieve the highest average performance on every benchmark, indicating that they provide an effective balance between insufficient correction and overcorrection.

D.5 Effect of Pruning-Severity Scaling

Plateau Attention Recalibration determines the layerwise strength by combining the causal gate $w_l^{(\tau)}$ with a function of the retained token budget. To study the effect of this function, we keep the causal region, β, and all other settings fixed and compare three variants:

$$\gamma_l = \beta g(K, N_0) w_l^{(\tau)}, \qquad g(K, N_0) = \begin{cases} 1, & \text{Constant,} \\ 1 - \dfrac{K}{N_0}, & \text{Linear,} \\ \log\left(\dfrac{N_0}{K}\right), & \text{Log.} \end{cases} \tag{29}$$

Table 15: Ablation of the recalibration strength β on LLaVA-v1.5-7B with VisionZip. The causal region is fixed for each benchmark, and only β is varied. † denotes the value used in the main experiments. Bold indicates the best average result across token budgets.

Benchmark	β	32	64	128	Avg.
GQA	$0.5^{\dagger}$	51.46	55.16	57.54	**54.72**
	1.0	50.53	54.82	57.45	54.27
POPE	1.5	78.70	79.50	83.72	80.64
	$2.0^{\dagger}$	78.37	81.41	84.05	**81.28**
	2.5	74.59	83.61	84.52	80.91
MME	$0.5^{\dagger}$	1581	1710	1769	**1686.67**
	1.0	1467	1631	1763	1620.33
MMB	0.5	55.37	60.34	61.95	59.22
	$1.0^{\dagger}$	55.23	60.54	62.01	**59.26**
	1.5	49.87	60.13	61.88	57.29
MMB^{CN}	$0.5^{\dagger}$	48.32	52.89	54.83	**52.01**
	1.0	47.52	52.22	54.97	51.57
SQA^{I}	0.5	68.72	69.11	68.52	68.78
	$1.0^{\dagger}$	69.01	69.51	68.47	**69.00**
	1.5	67.13	69.71	68.42	68.42
CQA	0.5	15.08	16.08	17.24	16.13
	$1.0^{\dagger}$	15.52	16.28	17.76	**16.52**
	1.5	14.84	16.96	17.76	**16.52**

The Constant variant applies the same strength at every token budget. The Linear variant increases the strength with the fraction of removed tokens but is bounded by β. Our Log variant instead scales with the compression ratio and continues to increase as pruning becomes more aggressive.

As shown in Table 16, all three variants improve AvgAcc. over the pruning baseline at every token budget. Constant improves AvgAcc. by 0.60, 0.32, and 0.75 points, while Linear improves it by 0.36, 0.25, and 0.68 points. Because these variants share the same causal region and differ only in their strength functions, their consistent gains provide complementary evidence that recalibrating the regions identified by Causal Attention Profiling is effective. This also highlights the importance of applying recalibration to the appropriate decoder layers.

The Log variant achieves the highest AvgAcc. at all three budgets, with gains of 0.84, 1.20, and 2.38 points. Its advantage becomes larger under more aggressive pruning, where a bounded or constant strength cannot fully reflect the increased loss of visual tokens. These results support the logarithmic scaling used in CAR.

E Additional Analysis of Modality Attention Imbalance

Figures 5 and 6 extend our analysis to three datasets on LLaVA-v1.5-7B and Qwen2.5-VL-7B. Across these settings, stronger pruning generally increases AVTA and decreases VTAR. The output distributions also tend to move away from the full-image condition and toward the no-image condition, showing that the imbalance and its associated prediction shift are not limited to the examples in the main paper.

LLaVA-NeXT exhibits a distinct attention response (Fig. 7): AVTA increases with pruning, whereas VTAR first rises at an intermediate token budget and then falls as pruning becomes more aggressive. This intermediate increase suggests stronger internal compensation than in LLaVA-v1.5-7B and Qwen2.5-VL-7B, although it does not prevent VTAR from falling below its full-token value under aggressive pruning.

Table 16: Ablation of pruning severity scaling on LLaVA-v1.5-7B with VisionZip. All variants use the same causal regions and differ only in the scaling function $g(K, N_0)$. AvgAcc. follows the normalized aggregate used in the main experiments. ΔAcc. is measured relative to the pruning baseline. Bold indicates the best result under each token budget.

Method	GQA	POPE	MME	MMB	MMB^{CN}	SQA^{I}	CQA	AvgAcc.	ΔAcc.
128 retained visual tokens									
Baseline	57.52	83.25	1767	61.61	54.77	68.67	16.96	94.90%	-
Constant	57.57	83.44	1772	61.88	54.83	68.57	17.56	95.50%	+0.60%
Linear	57.54	83.35	1769	61.88	54.77	68.57	17.32	95.26%	+0.36%
Log (Ours)	57.54	84.05	1769	62.01	54.83	68.47	17.76	**95.74%**	**+0.84%**
64 retained visual tokens									
Baseline	55.03	77.22	1718	60.07	52.95	68.86	15.88	91.36%	-
Constant	55.24	78.55	1693	60.47	52.89	69.11	16.04	91.68%	+0.32%
Linear	55.23	78.35	1698	60.47	52.75	69.06	16.00	91.61%	+0.25%
Log (Ours)	55.16	81.41	1710	60.54	52.89	69.51	16.28	**92.56%**	**+1.20%**
32 retained visual tokens									
Baseline	51.73	68.76	1586	54.83	48.26	68.86	14.56	84.85%	-
Constant	51.65	71.39	1579	55.03	47.92	68.62	15.16	85.60%	+0.75%
Linear	51.62	71.17	1578	55.03	47.99	68.62	15.12	85.53%	+0.68%
Log (Ours)	51.46	78.37	1581	55.23	48.32	69.01	15.52	**87.23%**	**+2.38%**

The relationship between accuracy and teacher-forced NTP loss also varies across model–dataset pairs. For LLaVA-v1.5-7B, both metrics decrease on GQA and MME, and increase together on SQA^{I}. For Qwen2.5-VL-7B, both decrease on MME, while the other evaluated pairs show opposite trends. LLaVA-NeXT also exhibits intervals of same-direction change in all three datasets. Since higher accuracy but lower NTP loss indicate better performance, such co-movement can give conflicting performance signals. These results show that teacher-forced NTP loss alone does not consistently track downstream accuracy after visual token pruning.

F Qualitative Examples

Figures 8 and 9 provide qualitative examples on natural-image VQA and chart understanding, respectively. In each row, the left column shows the pruned visual input and the locations of the retained tokens, the middle column presents the attention map and answer produced by pruning alone, and the right column shows the corresponding results after applying CAR. Across both types of tasks, the retained tokens continue to cover visual evidence relevant to the question, including target objects, chart legends, categories, and numerical values. Nevertheless, the pruned model fails to use this evidence effectively and produces an incorrect answer. Without changing the retained token set, CAR recalibrates decoder attention over the existing visual evidence and recovers the correct answer. These examples show that post-pruning errors can arise not only from removing relevant tokens, but also from insufficient use of the information that remains.

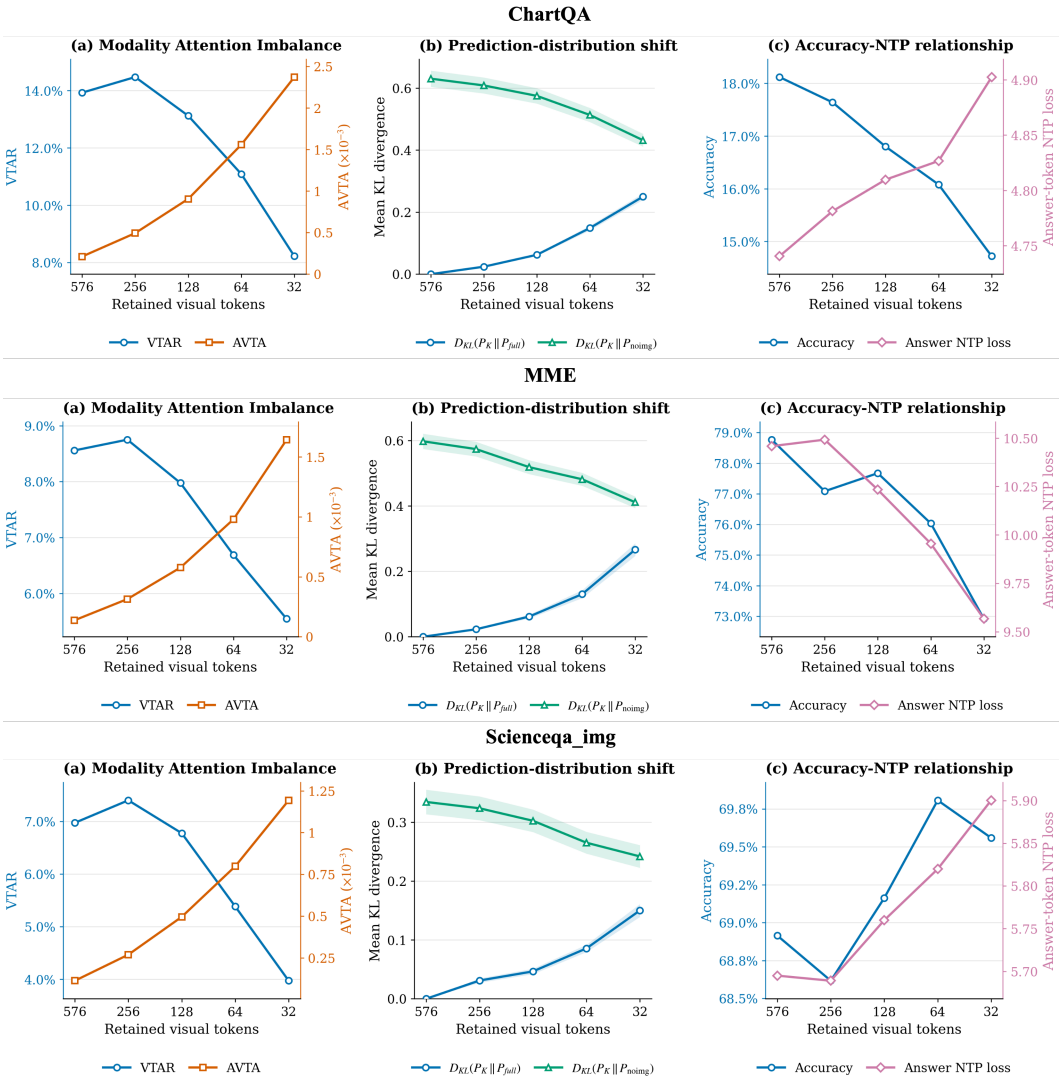


Figure 5: Additional analysis on LLaVA-v1.5-7B with VisionZip. Rows correspond to CQA, MME, and SQA^I from top to bottom. **(a)** Stronger pruning increases AVTA while decreasing VTAR. **(b)** The output distribution moves away from the full-image condition and toward the no-image condition. **(c)** The relationship between accuracy and teacher-forced NTP loss varies across datasets, with both same-direction and opposite-direction trends.

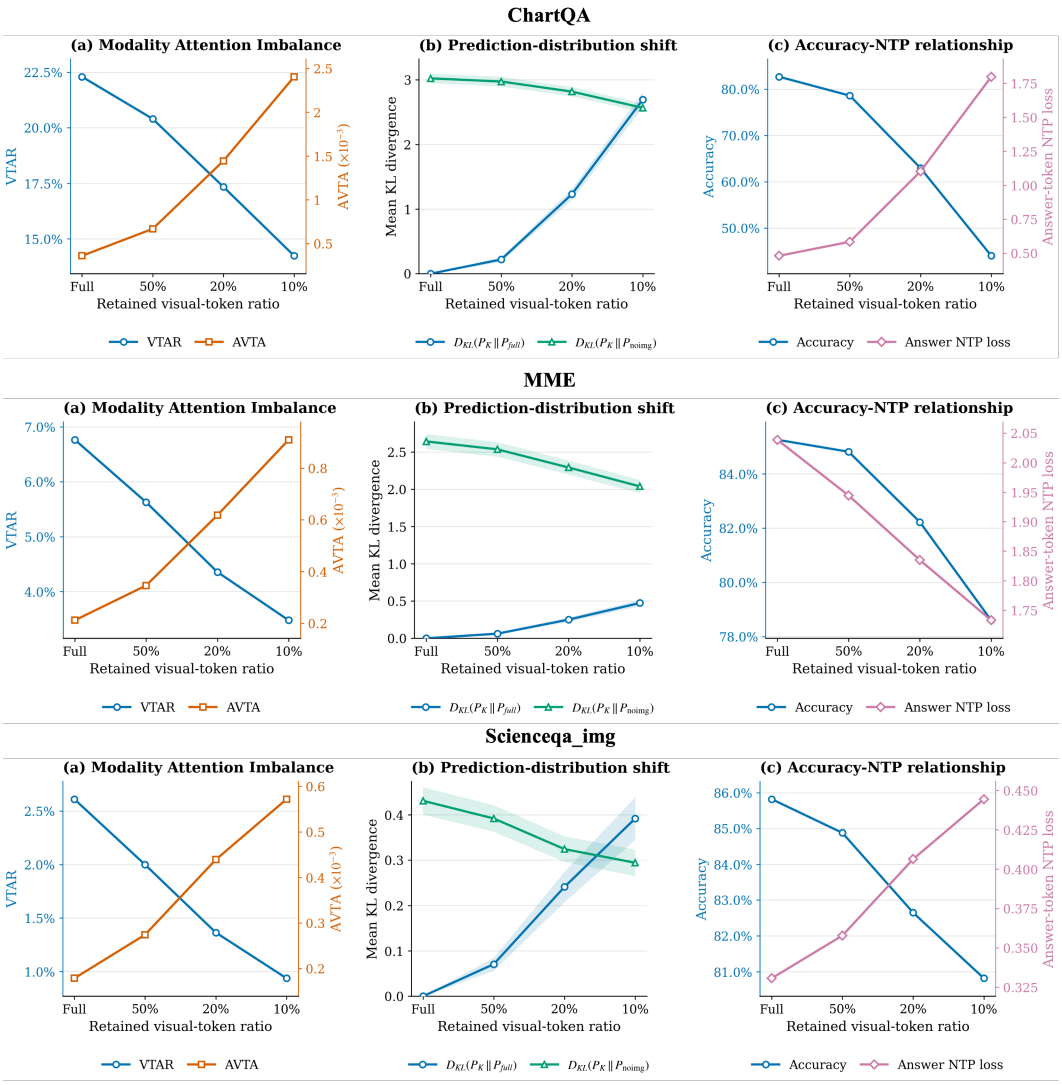


Figure 6: Additional analysis on Qwen2.5-VL-7B with VisionZip. Rows correspond to CQA, MME, and $\mathrm{SQA}^{\mathrm{I}}$ from top to bottom. **(a)** Stronger pruning consistently increases AVTA and decreases VTAR. **(b)** Predictions shift from the full-image condition toward the no-image condition. **(c)** Accuracy and teacher-forced NTP loss exhibit dataset-dependent relationships.

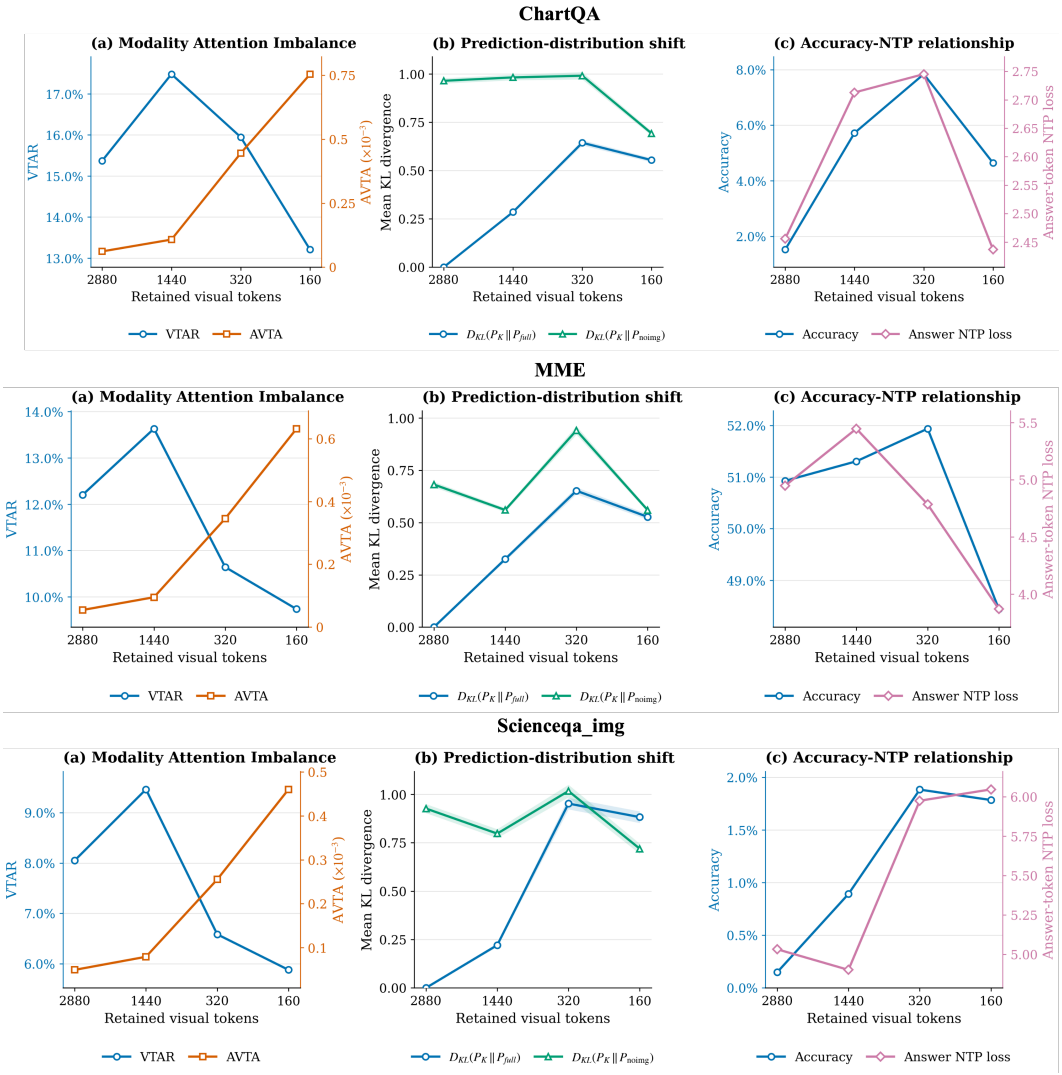


Figure 7: Additional analysis on LLaVA-NeXT-7B with VisionZip. Rows correspond to CQA, MME, and SQA^I from top to bottom. **(a)** AVTA increases with pruning, while VTAR peaks at an intermediate token budget before declining under stronger pruning. **(b)** KL divergence from the full-image and no-image output distributions. **(c)** Accuracy and teacher-forced NTP loss show same-direction changes over some pruning intervals across all three datasets.



Question:
Is there a car in the image?
Ground truth:
Yes

Pruning Only Answer:
No

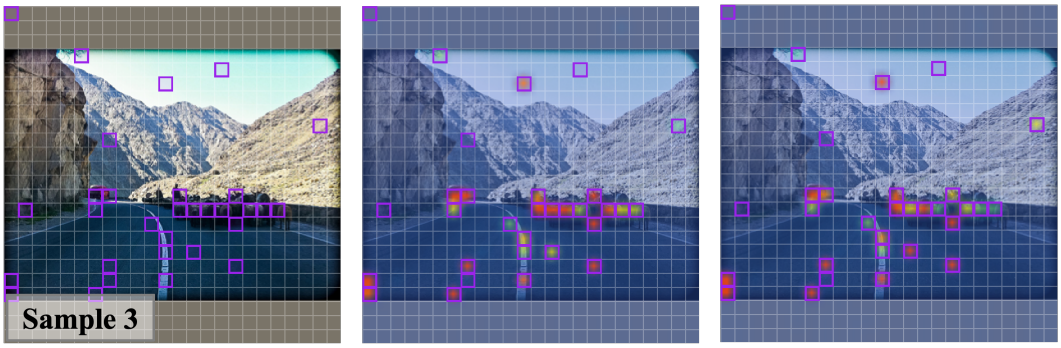
Pruning + CAR Answer:
Yes



Question:
Is there a spoon in the image?
Ground truth:
Yes

Pruning Only Answer:
No

Pruning + CAR Answer:
Yes



Question:
Is there a cow in the image?
Ground truth:
Yes

Pruning Only Answer:
No

Pruning + CAR Answer:
Yes

Figure 8: Qualitative examples on natural-image VQA. Purple boxes denote the visual tokens retained after pruning, while the heat maps visualize decoder attention over these tokens. Although task-relevant evidence remains, pruning alone produces incorrect answers. CAR recalibrates attention over the unchanged token set and recovers the correct answers.

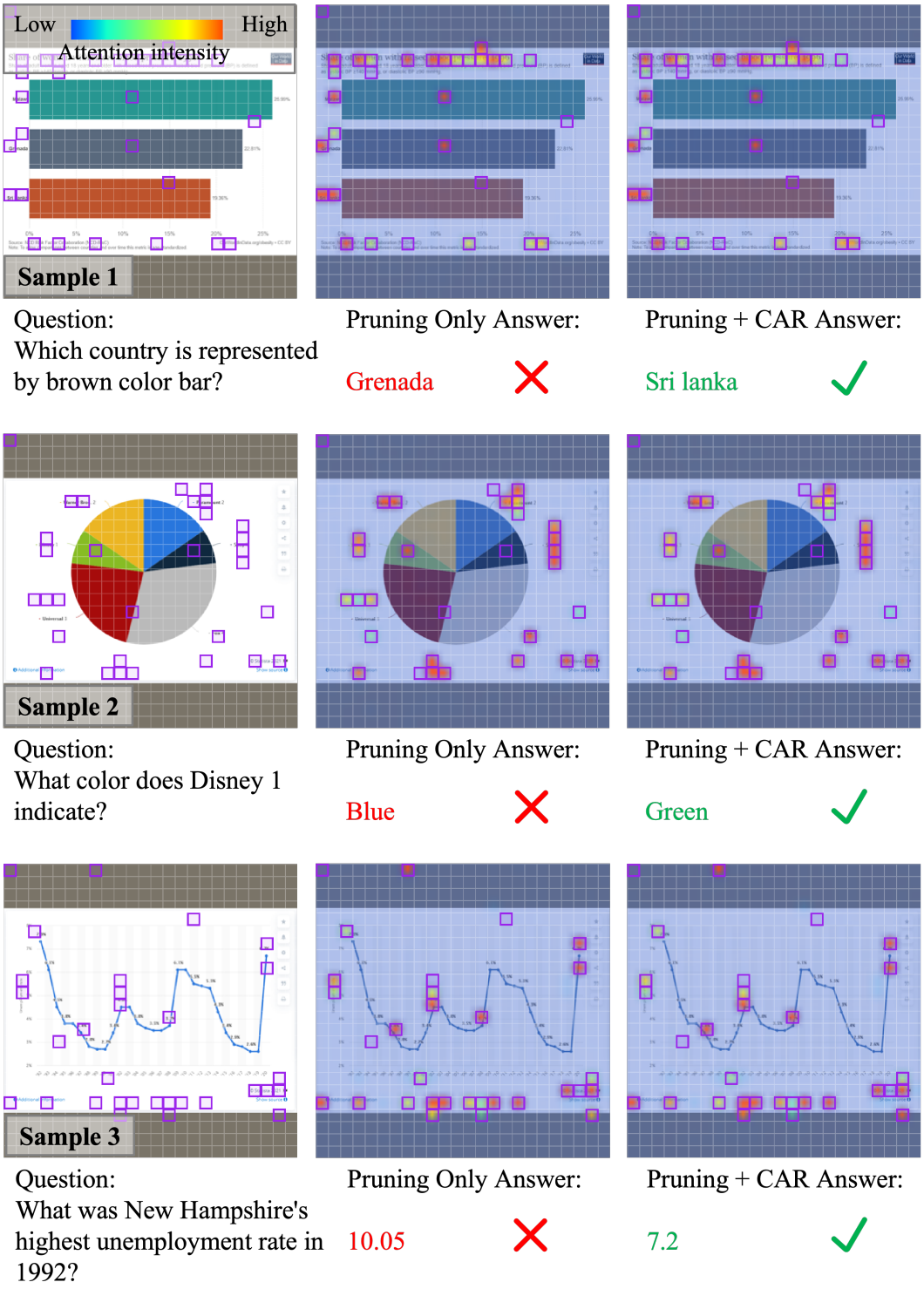


Figure 9: Qualitative examples on chart understanding. The retained visual tokens still cover the chart elements required to answer each question, yet pruning alone yields incorrect categorical, color, and numerical predictions. CAR reallocates attention over the same retained tokens and restores the correct answers.