IRREPLACEABILITY MATTERS: COUNTERFACTUAL INTERVENTION FOR VISUAL TOKEN PRUNING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) represent images as long visual-token sequences, making inference costly. Training-free pruning methods typically rank tokens by saliency or redundancy cues from a single forward pass, yet prominently scored tokens can remain compensable under tight budgets—importance alone does not ensure that visual evidence is hard to replace. We introduce **CiVi**, a plug-and-play inference-time selector that augments factual-view saliency and pairwise non-redundancy with a feature-space intervention-response proxy inspired by counterfactual reasoning. CiVi perturbs encoder features, compares clean and perturbed projected representations for ranking, and passes only retained clean tokens to the decoder. The intervention response provides a low-cost proxy for decoder-relevant sensitivity rather than an identified causal effect; decoder-level interventions and two-cue ablations support its decoder relevance and the complementarity of the three signals. Across LLaVA-, Qwen-, and video MLLM backbones, CiVi achieves a strong accuracy–efficiency trade-off, particularly under aggressive pruning. On LLaVA-1.5-7B, it preserves 95.7% of the unpruned nine-task average using only 64 of 576 visual tokens. Our code is available at `https://anonymous.4open.science/r/CiVi-1082`.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) achieve strong performance across diverse vision-language tasks by coupling vision encoders with autoregressive language decoders (Liu et al., 2023; 2024a; Bai et al., 2023; Li et al., 2023a). However, long visual token sequences substantially increase both memory consumption and decoding latency (Dosovitskiy, 2020; Vaswani et al., 2017; Li et al., 2025). High-resolution settings exacerbate this problem, yielding thousands of visual tokens per image (Liu et al., 2024b; Li et al., 2024a; Lin et al., 2024; Li et al., 2025). Inference-time Visual Token Pruning (VTP) therefore offers a practical strategy for improving MLLM efficiency (Chen et al., 2024; Zhang et al., 2025b; Yang et al., 2025; Shang et al., 2025; Xing et al., 2024).

VTP methods use instruction-aware relevance (Zhang et al., 2025b; Song et al., 2025), visual saliency (Chen et al., 2024; Zhang et al., 2025a; Arif et al., 2025), or token compression (Bolya et al., 2023; Shang et al., 2025; Li et al., 2024b; Yao et al., 2024), estimating importance from a single forward pass (Zou et al., 2026; Wen et al., 2025; Ma et al., 2026). These cues capture factual prominence (Zhang et al., 2025a; Yang et al., 2025; Zhang et al., 2025b), but not whether retained tokens can compensate for evidence lost after pruning. Under tight budgets, salient tokens may carry redundant evidence (Zou et al., 2026; Wen et al., 2025), while less salient tokens may be *irreplaceable* because they preserve unique information (Ma et al., 2026; Deng et al., 2026). This motivates selecting evidence that the retained set cannot replace, beyond factual importance.

The left panel of Fig. 1 illustrates this mismatch through a boundary-replacement diagnostic. Under 88.9% pruning, retained boundary tokens are swapped with discarded boundary tokens within each method's own ranking while the token budget is fixed. If the swapped retained evidence can be compensated by the remaining set, replacement should have only a mild effect; conversely, replacement is more likely to alter the prediction when the removed evidence is difficult to compensate. Replacement has little effect in the attention-based example, whereas swapping boundary tokens retained

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.

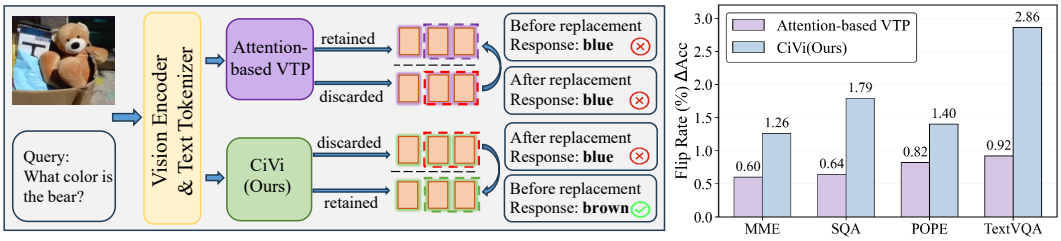


Figure 1: **Boundary-replacement toy experiment.** Left: Boundary-replacement setup. Under 88.9% pruning, retained boundary tokens are swapped with discarded boundary tokens within each method's own ranking, while keeping the token budget unchanged. Right: Flip Rate (ΔAcc). Higher values indicate that the retained tokens are harder to compensate for when replaced.

by our method is more likely to change a correct prediction into an incorrect one. This diagnostic provides descriptive evidence about the selected sets under the specified replacement protocol rather than token-level causal identification.

The right panel quantifies this diagnostic with the replacement-induced correct-to-wrong flip rate ΔAcc. A larger ΔAcc indicates that, under the same replacement protocol, the swapped retained evidence is harder for the remaining set to compensate. Its magnitude nevertheless depends on the replacement rule, decision margin, retained set, and surrounding context. Across four benchmarks, our method yields a larger ΔAcc than attention-based VTP under the same protocol. These observations motivate a relational view of irreplaceability: whether visual evidence is difficult to replace depends not only on an individual token but also on what information is already preserved by the retained set. Directly testing this relation through the decoder for every candidate is impractical because it requires many repeated decoder passes.

Guided by this operational view, we propose CiVi, a plug-and-play VTP method that uses controlled feature-space perturbations to obtain a complementary signal for harder-to-replace visual evidence. Between the vision encoder and decoder, CiVi compares clean and perturbed representations after the frozen projector and measures their intervention response for token scoring. It then combines factual-view saliency, the intervention-response proxy, and pairwise non-redundancy for top-K pruning. Perturbed and discrepancy-enhanced features are used only to compute scores; the frozen decoder receives the selected clean tokens in their original order. The intervention response is a low-cost proxy for decoder-relevant sensitivity rather than an identified causal effect. We therefore validate its decoder alignment separately by directly replacing individual projected tokens and measuring changes in the next-token distribution.

Our highlights are summarized as follows:

- We motivate high-ratio pruning from a relational perspective: the retained set should preserve evidence that is difficult for its remaining tokens to compensate, rather than only tokens prominent in a factual forward pass.
- We propose CiVi, an intervention-inspired selector that combines factual-view saliency, a post-projector intervention-response proxy, and pairwise non-redundancy while passing only retained clean tokens to the decoder.
- We directly validate the intervention-response proxy against expensive decoder interventions and evaluate CiVi across LLaVA-family, Qwen-family, high-resolution, and video settings, obtaining a competitive accuracy–efficiency trade-off without retraining.

2 Related Work

2.1 MLLMs and Efficiency Challenges

LLM advances have driven MLLM development (Touvron et al., 2023), including LLaVA (Liu et al., 2023; 2024b) and Qwen-VL (Bai et al., 2025b; 2023; 2025a). These models couple vision encoders with language decoders (Li et al., 2023a; Alayrac et al., 2022; Liu et al., 2023) to support diverse multimodal tasks (Zhang et al., 2024b; Li et al., 2024c; Ji et al., 2024; Zhang et al., 2024a). Long visual sequences increase memory and computation costs (Vaswani et al., 2017; Chen et al., 2024;

Li et al., 2025), particularly in high-resolution models such as LLaVA-NeXT (Liu et al., 2024b; Li et al., 2025; 2024a), which may generate thousands of tokens per image (Liu et al., 2024b; Li et al., 2024a; Lin et al., 2024). Reducing redundant visual tokens is therefore important for efficient inference (Chen et al., 2024; Yang et al., 2025; Zhang et al., 2025b; Shang et al., 2025). This makes token selection a practical way to balance inference efficiency and visual fidelity.

2.2 Visual Token Pruning Methods

VTP aims to shorten visual sequences while preserving decision-critical evidence. The central trade-off is between prioritizing individually important tokens and maintaining complementary information across the retained visual sequence. Existing methods use instruction-aware signals such as attention and question relevance (Chen et al., 2024; Zhang et al., 2025b), visual cues such as [CLS] attention, saliency, and locality (Chen et al., 2024; Zhang et al., 2025a; Yang et al., 2025; Arif et al., 2025), or token merging and compression (Bolya et al., 2023; Shang et al., 2025; Xing et al., 2024; Li et al., 2024b; Yao et al., 2024). Representative training-free methods include ToMe (Bolya et al., 2023), LLaVA-PruMerge (Shang et al., 2025), VisionZip (Yang et al., 2025), DART (Wen et al., 2025), DivPrune (Alvar et al., 2025), D^2Pruner (Zhang et al., 2026), and V^2Drop (Chen et al., 2026). These approaches explore redundancy, structural diversity, and decoder-layer variation; ZOO-Prune studies zeroth-order sensitivity (Kim et al., 2026). Compression approaches also include LLaMA-VID (Li et al., 2024b) and DeCo (Yao et al., 2024). Other work emphasizes holistic context retention (Zou et al., 2026) and saliency-coverage trade-offs (Deng et al., 2026).

Many methods derive selection cues from factual-view saliency (Zou et al., 2026; Yang et al., 2025; Ma et al., 2026), decoder-layer dynamics (Chen et al., 2026), or local redundancy (Bolya et al., 2023; Yang et al., 2025; Wen et al., 2025). These identify prominent or redundant tokens without directly measuring whether the remaining set can compensate for lost evidence. A token's usefulness therefore depends not only on its own score but also on its context. Under aggressive pruning, losing critical object evidence can also affect hallucination-sensitive predictions (Li et al., 2023c). CiVi complements saliency and pairwise non-redundancy with an intervention-response proxy from controlled feature-space perturbations. We validate its decoder relevance empirically and interpret it as an operational scoring signal rather than an identified causal effect.

2.3 Contrastive Decoding and Feature-Space Contrast

Contrastive decoding uses differences between output distributions to guide generation (Leng et al., 2024; Li et al., 2023b). VCD (Leng et al., 2024) contrasts clean and distorted visual inputs to mitigate hallucinations. Motivated by this comparison, CiVi scores differences between clean and perturbed projected features before decoding. Only retained clean tokens enter the decoder. Perturbed features serve only to assess token sensitivity, rather than to replace visual evidence or alter the decoder's output distribution. This avoids repeated decoder passes but does not directly capture nonlinear decoder responses, so we validate decoder alignment separately in Sec. 4.3. The connection is motivational rather than algorithmic.

3 Method

Building on the intervention-inspired view, we propose **CiVi**, a plug-and-play, training-free visual token pruning module inserted between the vision encoder and the LLM. Given encoder tokens $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{T}$, CiVi applies adaptive feature-space perturbations, constructs discrepancy-enhanced representations, and combines factual-view saliency, intervention response, and pairwise non-redundancy for top-K scoring. Scores from multiple perturbation samples are averaged, and only the top-K clean tokens are passed to the frozen LLM in their original sequence order. Figure 2 presents the overall pipeline.

3.1 Operational Target

Let $\mathcal{P}(\cdot)$ denote the frozen visual projector and $\mathbf{V} = \mathcal{P}(\mathbf{Z}) = \{\mathbf{v}_i\}_{i=1}^{T}$ the projected visual tokens. An informative but expensive reference for whether token i carries hard-to-compensate evidence is

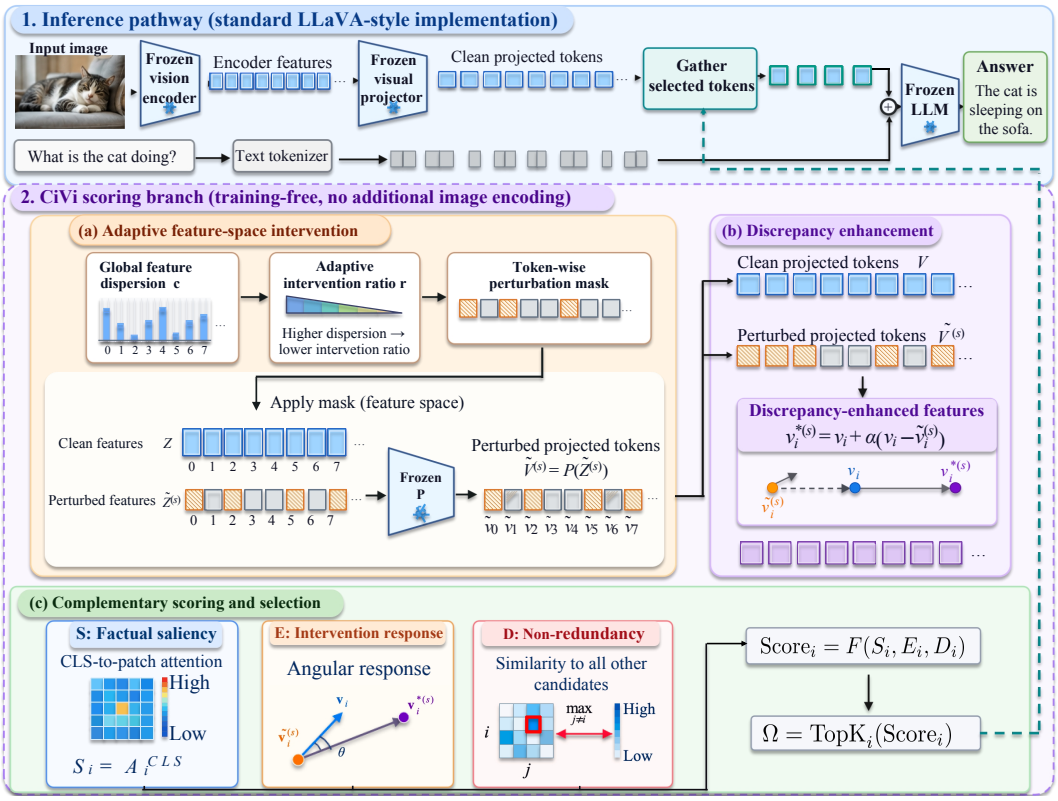


Figure 2: **Overview of CiVi.** CiVi applies adaptive feature-space perturbations, constructs discrepancy-enhanced representations from clean and perturbed projected tokens, and performs top-K pruning using factual-view saliency, intervention response, and pairwise non-redundancy. Only retained clean tokens are passed to the LLM.

the decoder-output change induced by replacing it while keeping the remaining context fixed:

$$\mathcal{I}_i = \mathbb{E}_{\xi \sim p(\xi|\mathbf{V})} \left[d\left(f(\mathbf{V}), f\left(\mathbf{V}^{(\mathbf{v}_i \leftarrow \xi)} \right) \right) \right], \tag{1}$$

where $f(\cdot)$ denotes the frozen decoder conditioned on the text query, $p(\xi \mid \mathbf{V})$ is a matched replacement distribution, and $d(\cdot, \cdot)$ measures output-distribution discrepancy. We use $\mathcal{I}_i$ as an operational reference rather than an identified causal effect. Evaluating it for every token requires repeated decoder passes; CiVi therefore approximates this signal before decoding through a low-cost feature-space intervention response, whose decoder alignment is evaluated in Sec. 4.3.

3.2 Adaptive Feature-Space Intervention

We first compute the token centroid $\bar{\mathbf{z}} = \frac{1}{T} \sum_{i=1}^{T} \mathbf{z}_i$ and estimate the global feature dispersion

$$c = \text{clip}\left(1 - \frac{1}{T} \sum_{i=1}^{T} \cos(\mathbf{z}_i, \bar{\mathbf{z}}), 0, 1 \right). \tag{2}$$

The adaptive intervention ratio is

$$r = \text{clip}\left(r_0(1.5 - c), 0.05, 0.95 \right), \tag{3}$$

where r_0 is the base mask ratio. We treat c as an inexpensive feature-space heuristic and compare it with fixed ratios in Sec. 4.5.

For intervention sample s, we draw $m_i^{(s)} \sim \text{Bernoulli}(r)$ and Gaussian noise $\boldsymbol{\epsilon}_i^{(s)}$. The perturbed feature is

$$\tilde{\mathbf{z}}_i^{(s)} = (1 - m_i^{(s)})\mathbf{z}_i + m_i^{(s)} \left(\sqrt{\bar{\alpha}_t}\, \mathbf{z}_i + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon}_i^{(s)} \right). \tag{4}$$

The intervention preserves token number and indices and requires no additional image encoding pass.

3.3 Intervention-Response Scoring

The frozen projector maps both views into the LLM-aligned space: $\mathbf{v}_i = \mathcal{P}(\mathbf{z}_i)$ and $\tilde{\mathbf{v}}_i^{(s)} = \mathcal{P}(\tilde{\mathbf{z}}_i^{(s)})$. We construct a discrepancy-enhanced representation

$$\mathbf{v}_i^{\star(s)} = \mathbf{v}_i + \alpha \left(\mathbf{v}_i - \tilde{\mathbf{v}}_i^{(s)} \right), \tag{5}$$

where α controls the enhancement strength.

CiVi combines three complementary cues. Factual-view saliency is computed once from the clean vision encoder:

$$S_i = A_i^{\text{cls}}, \tag{6}$$

where A_i^{cls} is the CLS-to-patch attention score. For LLaVA-1.5, $A_i^{\text{cls}} = H^{-1} \sum_h A_{h,\text{CLS},i}^{(-2)}$, the head-averaged CLS-to-patch attention from the penultimate encoder layer. Backbone-specific details are in Appendix B.4. For intervention sample s:

$$\begin{aligned} E_i^{(s)} &= 1 - \cos \left(\mathbf{v}_i^{\star(s)}, \tilde{\mathbf{v}}_i^{(s)} \right), \\ D_i^{(s)} &= 1 - \max_{j \neq i} \cos \left(\mathbf{v}_i^{\star(s)}, \mathbf{v}_j^{\star(s)} \right). \end{aligned} \tag{7}$$

S, E, and D respectively capture factual prominence, intervention response, and pairwise non-redundancy.

To reduce scale imbalance, we fuse the three cues in the log domain:

$$\text{Score}_i^{(s)} = \sum_{U \in \{S,E,D\}} \lambda_U^{(s)} \log \left(U_i^{(s)} + \varepsilon \right), \tag{8}$$

where $S_i^{(s)} \equiv S_i$ and

$$\lambda_U^{(s)} = \frac{1}{\max \left(\text{std}_i[\log(U_i^{(s)} + \varepsilon)], 0.1 \right)}. \tag{9}$$

The coefficient normalizes numerical dispersion across score channels; it is not an estimate of semantic reliability.

We average scores across N perturbation samples:

$$\overline{\text{Score}}_i = \frac{1}{N} \sum_{s=1}^{N} \text{Score}_i^{(s)}. \tag{10}$$

Given pruning ratio ρ, we set $K = \max(1, \text{round}(T(1-\rho)))$ and select

$$\Omega = \text{TopK} \left(\{\overline{\text{Score}}_i\}_{i=1}^{T}, K \right). \tag{11}$$

The selected indices are sorted by their original positions, and only the corresponding clean tokens $\{\mathbf{v}_i \mid i \in \Omega\}$ are passed to the LLM.

3.4 Practical Integration and Cost

CiVi requires no training or backbone modification and performs pruning once before decoding. After pruning, each decoder layer processes only K visual tokens instead of T. The additional pre-decoding cost is dominated by the $O(NT^2 d_p)$ pairwise non-redundancy computation, incurred once; the shortened visual sequence offsets this cost across subsequent decoder layers. Additional complexity analysis is provided in Appendix A.3.

Table 1: Main results on LLaVA-1.5-7B under three pruning ratios. Best results under each ratio are highlighted. "-" denotes missing or unreported values.

Method	MME	MMB	POPE	VQA_{Text}	GQA	SQA	VQA_{v2}	MM-Vet	OCR	Avg.
LLaVA-1.5-7B (No Pruning)	1862	64.7	85.9	58.2	61.9	69.5	78.4	31.1	297	100.0%
Pruning Ratio = 66.7% (Retain 192 Tokens)										
PDrop (CVPR'25)	1766	63.2	82.3	56.1	57.1	68.8	75.1	30.5	290	96.4%
VisionZip (CVPR'25)	1783	63.0	85.3	57.3	59.3	68.9	76.8	31.7	293	98.3%
SparseVLM (ICML'25)	1721	62.5	83.6	56.1	57.6	69.1	75.6	31.5	292	96.8%
LLaVA-PruMerge (ICCV'25)	1632	59.6	71.3	54.3	54.3	67.9	70.6	-	253	89.6%
HoloV (NeurIPS'25)	1820	**65.4**	85.6	57.4	59.0	**69.8**	76.7	28.9	**300**	98.3%
FiCoCo-V (AAAI'26)	1732	62.3	82.5	55.7	58.5	67.8	74.4	29.5	-	95.4%
CiVi (Ours)	**1823**	63.3	**87.3**	**57.7**	**60.1**	69.0	**77.1**	**32.7**	290	**99.3%**
Pruning Ratio = 77.8% (Retain 128 Tokens)										
PDrop (CVPR'25)	1644	61.1	82.3	55.1	56.0	68.3	72.9	30.5	**290**	94.5%
VisionZip (CVPR'25)	1768	63.4	84.7	56.8	57.6	68.8	75.6	**32.6**	**290**	97.8%
SparseVLM (ICML'25)	1696	60.0	80.5	54.9	56.0	67.1	73.8	30.0	280	93.8%
LLaVA-PruMerge (ICCV'25)	1554	58.1	67.2	54.3	53.3	67.1	68.8	-	248	87.3%
HoloV (NeurIPS'25)	**1802**	**63.9**	84.0	56.8	57.7	**69.8**	75.5	28.2	**290**	96.6%
FiCoCo-V (AAAI'26)	1711	61.1	82.2	55.6	57.6	68.3	73.1	27.9	-	94.0%
CiVi (Ours)	**1802**	63.1	**86.0**	**57.2**	**59.0**	69.5	**76.1**	31.5	**290**	**98.2%**
Pruning Ratio = 88.9% (Retain 64 Tokens)										
PDrop (CVPR'25)	1092	33.3	55.9	45.9	41.9	68.6	69.2	30.7	250	76.9%
VisionZip (CVPR'25)	1690	60.1	77.0	55.5	55.1	69.0	72.4	**31.7**	273	93.7%
SparseVLM (ICML'25)	1505	56.2	75.1	51.8	52.7	62.2	68.2	23.3	180	82.4%
LLaVA-PruMerge (ICCV'25)	1549	55.3	65.3	54.0	51.9	68.1	67.4	-	250	86.2%
HoloV (NeurIPS'25)	1715	**63.3**	80.3	55.4	55.3	69.5	72.8	26.9	270	93.1%
DART (EMNLP'25)	**1756**	61.2	79.1	54.4	56.5	68.9	69.9	28.9	270	93.1%
FiCoCo-V (AAAI'26)	1591	60.3	76.0	53.6	52.4	68.1	71.3	26.1	-	89.6%
D^2Pruner (AAAI'26)	1716	60.8	77.2	54.5	55.3	69.1	73.5	29.1	280	93.3%
V^2Drop (CVPR'26)	1641	55.2	75.1	51.8	53.5	68.9	72.9	29.2	280	90.7%
CiVi (Ours)	1708	62.8	**83.5**	**55.9**	**56.8**	**69.9**	**73.6**	29.7	**290**	**95.7%**

4 Experiments

4.1 Experimental Setup

Benchmarks. Following prior training-free pruning work, we evaluate on widely used MLLM benchmarks, including MME (Fu et al., 2025), MMBench (MMB) (Liu et al., 2024c), POPE (Li et al., 2023c), TextVQA (Singh et al., 2019), GQA (Hudson & Manning, 2019), ScienceQA (SQA) (Lu et al., 2022), VQA-v2 (Goyal et al., 2017), OCRBench (Liu et al., 2024d), and MM-Vet (Yu et al., 2024). For video understanding, we further evaluate on TGIF-QA (Jang et al., 2017), MSVD-QA (Xu et al., 2017), and MSRVTT-QA (Xu et al., 2016). These benchmarks cover broad visual understanding, hallucination sensitivity, text-rich reasoning, general compositional QA, and video question answering, which is important since CiVi is designed to preserve evidence that is difficult to compensate under aggressive compression rather than tokens that merely appear salient in the factual view.

Backbones and pruning settings. We evaluate CiVi on four representative MLLMs: LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025b), and Video-LLaVA-7B (Lin et al., 2024), covering standard image, high-resolution, cross-architecture, and video inputs. LLaVA-1.5-7B is tested at 66.7%, 77.8%, and 88.9% pruning, retaining 192/128/64 visual tokens. LLaVA-NeXT-7B is evaluated at 88.9% pruning, Qwen2.5-VL-7B at 80% and 90% pruning, and Video-LLaVA-7B retains 256 out of 2048 video tokens. Additional results on LLaVA-1.5-13B (Liu et al., 2024a) and Qwen3-VL-8B (Bai et al., 2025a) are reported in Appendix B.

Implementation and timing protocol. CiVi is inserted between the vision encoder and the decoder and performs one-shot top-K selection before decoding. All methods are compared under the same backbone and retained-token budget. Efficiency results report end-to-end costs, including scoring

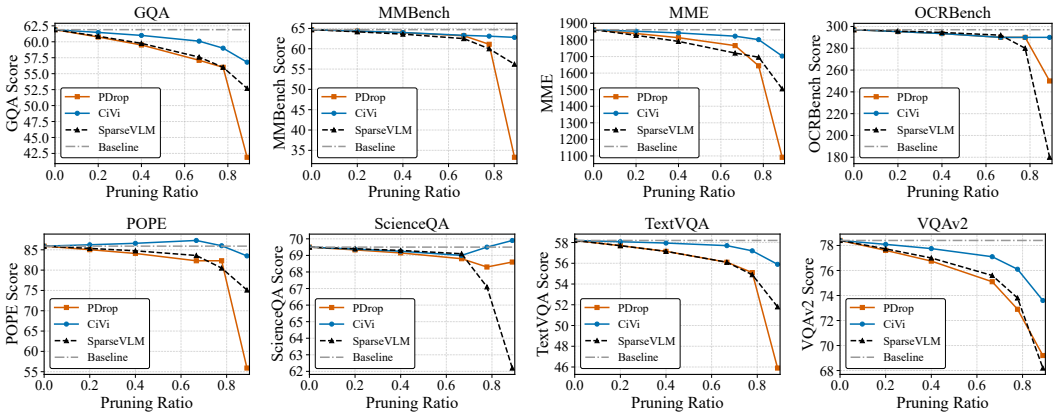


Figure 3: Performance trends on eight benchmarks under different pruning ratios for PDrop, SparseVLM, and CiVi, with the unpruned baseline shown as a reference.

and token-selection overhead, to reflect practical inference behavior. More benchmark settings, pruning ratios, implementation details, and timing protocols are provided in Appendix B.

4.2 Main Results

We compare CiVi with strong training-free acceleration baselines that appear in the following result tables, including HoloV (Zou et al., 2026), SparseVLM (Zhang et al., 2025b), LLaVA-PruMerge (Shang et al., 2025), VisionZip (Yang et al., 2025), PDrop (Xing et al., 2024), FiCoCo-V (Han et al., 2026), DART (Wen et al., 2025), D^2Pruner (Zhang et al., 2026), V^2Drop (Chen et al., 2026), FastV (Chen et al., 2024), FasterVLM (Zhang et al., 2025a), HiRED (Arif et al., 2025), and GlobalCom^2 (Liu et al., 2026), together with the vanilla no-pruning setting for efficiency reference. These baselines cover attention-based selection, duplication-aware and variation-aware pruning, debiased importance with structural diversity, zeroth-order sensitivity, token merging, and compression-allocation variants. Across these comparisons, we focus on whether each method preserves performance as the retained-token budget becomes increasingly constrained, which reflects robustness under high-ratio pruning.

4.2.1 Main Results on LLaVA-1.5-7B

Table 1 summarizes LLaVA-1.5-7B results under three pruning ratios. CiVi preserves 99.3%, 98.2%, and 95.7% of the unpruned nine-benchmark average at 66.7%, 77.8%, and 88.9% pruning, respectively. At 88.9%, the comparison further includes recent duplication-, diversity-, variation-, and sensitivity-based methods under matched protocol. Figure 3 shows per-benchmark degradation trends as pruning increases; CiVi degrades more gracefully than PDrop and SparseVLM, with clear advantages on benchmarks where answers depend on sparse or text-bearing evidence (MME, OCRBench, TextVQA, GQA).

4.2.2 Main Results on Higher-Resolution and Video Settings

We extend CiVi to LLaVA-NeXT-7B and Video-LLaVA-7B, covering higher-resolution inputs and video understanding, where longer visual sequences make sparse cues easier to discard under aggressive pruning. Table 2 reports Video-LLaVA-7B results, where all methods retain 256 of 2048 video tokens. Under this compression, CiVi outperforms all baselines and reaches 102.3% of the original average performance, suggesting that video inputs contain substantial redundancy and that intervention-inspired ranking removes redundant or distracting tokens while preserving decisive evidence. On LLaVA-NeXT-7B, visual sequences reach up to 2880 tokens; at 88.9% pruning, CiVi retains only

Table 2: Video understanding results on Video-LLaVA-7B at 87.5% pruning.

Method	TGIF	MSVD	MSRVTT	Avg (%)
Video-LLaVA-7B	46.9	68.2	56.3	100.0%
FastV (ECCV'24)	44.2	60.3	40.6	84.9%
PDrop (CVPR'25)	40.2	61.5	41.8	83.3%
VisionZip (CVPR'25)	44.2	65.1	54.2	95.3%
SparseVLM (ICML'25)	45.9	68.6	32.9	85.6%
CiVi (Ours)	**48.4**	**70.8**	**56.2**	**102.3%**

320 tokens and achieves 94.8% average retention across six benchmarks. As shown in Table 3, it attains the highest overall average and strong results across most benchmarks.

Table 3: Results on LLaVA-NeXT-7B at 88.9% pruning.

Method	MME	MMB	POPE	VQA_{Text}	GQA	SQA	Avg (%)
LLaVA-NeXT-7B (No Pruning)	1851	67.4	86.5	64.9	64.2	70.1	100.0%
HiRED (AAAI'25)	1692	**64.1**	83.4	**58.7**	59.3	66.7	93.4%
PDrop (CVPR'25)	1663	63.4	77.6	54.4	56.4	67.5	90.3%
VisionZip (CVPR'25)	1630	60.1	82.1	56.2	55.5	68.3	90.4%
SparseVLM (ICML'25)	1533	60.6	82.4	58.4	56.1	66.1	89.9%
FasterVLM (ICCV'25)	1701	61.6	83.6	56.5	56.9	66.5	91.7%
LLaVA-PruMerge (ICCV'25)	1534	61.3	60.8	50.6	53.6	66.4	83.4%
$GlobalCom^2$ (AAAI'26)	1698	61.8	83.8	57.2	57.1	67.4	92.3%
CiVi (Ours)	**1760**	63.4	**85.5**	56.8	**60.9**	**69.0**	**94.8%**

4.2.3 Main Results on Qwen2.5-VL-7B

To assess architectural robustness, we test CiVi on Qwen2.5-VL-7B, whose visual front-end and alignment pipeline differ from those of the LLaVA family. Across five benchmarks at 80% and 90% pruning, CiVi leads nearly all metrics, with a particularly large margin on POPE (Table 4). Appendix B extends the same-backbone comparison to DART, D^2Pruner, and VisionZip, revealing a more task-dependent pattern, and reports turn-wise multi-round dialogue, additional video reasoning, and four fixed-subset captioning evaluations. These results support transfer across token layouts while identifying settings where another selector performs better.

Table 4: Results on Qwen2.5-VL-7B at two pruning ratios.

Method	GQA	MME	POPE	MMB	SQA
Qwen2.5-VL-7B	60.5	2304	86.1	82.8	76.7
Pruning Ratio = 80%					
SparseVLM	54.7	2065	73.7	75.9	71.6
PDrop	54.9	2097	78.4	76.3	70.9
CiVi (Ours)	**56.6**	**2108**	**85.3**	**76.9**	**74.4**
Pruning Ratio = 90%					
SparseVLM	51.3	1849	71.9	71.7	68.2
PDrop	51.9	1867	75.0	71.9	69.6
CiVi (Ours)	**52.1**	**1872**	**80.3**	**72.1**	**71.2**

4.3 Decoder-Intervention Validation

We test whether CiVi's inexpensive proxy recovers the token ranking produced by explicit decoder interventions. On 900 POPE images, we construct a decoder-intervention reference by fixing the prompt and all other projected visual tokens, replacing each sampled token with its perturbed counterpart, and ranking tokens by the resulting Jensen–Shannon divergence between next-token distributions. NDCG@8 measures agreement on the eight most decoder-sensitive tokens; higher is better.

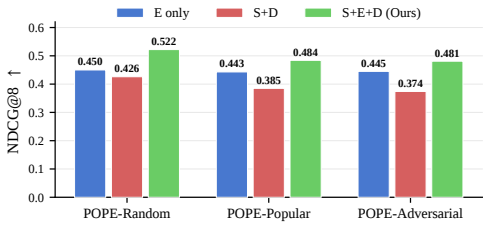


Figure 4: NDCG@8 agreement between proxy rankings and the explicit decoder-intervention ranking on POPE.

As shown in Fig. 4, the fused $S{+}E{+}D$ score achieves the highest NDCG@8 across all three subsets. Adding E to $S{+}D$ consistently improves agreement, indicating that the intervention response helps recover tokens whose explicit replacement most changes decoder output. Thus, the proxy approximates the decoder-intervention ranking.

4.4 Efficiency Analysis

Table 5 compares efficiency with three strong matched baselines on both LLaVA-1.5 and LLaVA-NeXT. At 90% pruning, all methods use the same 200-sample GQA subset, batch size 1, and three repeats on one NVIDIA H800. TTFT covers visual processing through the first decoder forward pass, Prefill includes token selection, and Prune isolates scoring and selection. CiVi achieves the lowest measured TTFT and Prefill among the evaluated pruners on both backbones, while D^2Pruner has the smallest selector overhead. Total generation time depends on output length and should not

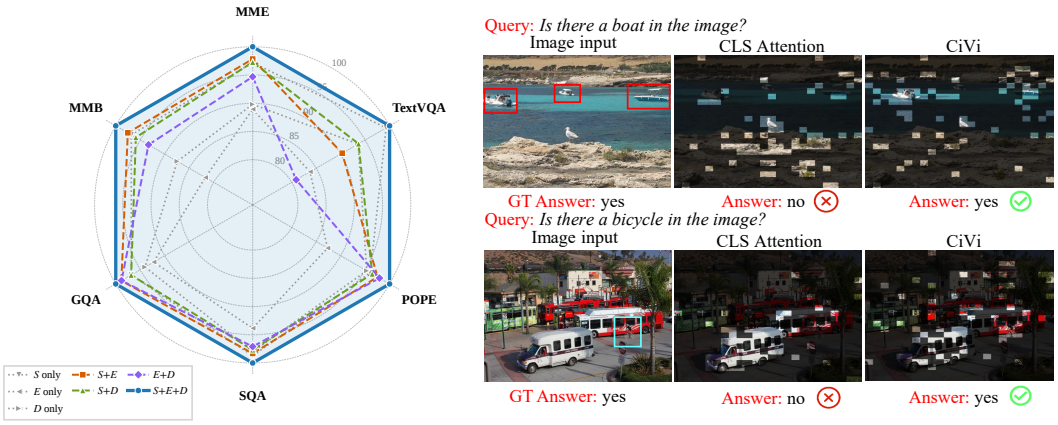
be interpreted as selector speed alone. Overall, the results demonstrate a competitive end-to-end accuracy–efficiency trade-off. Full timing definitions are provided in Appendix B.5.

Table 5: Matched inference efficiency at 90% pruning (ms).

LLaVA-1.5						LLaVA-NeXT					
Method	**TTFT↓**	**Prefill↓**	**Prune↓**	**Dec/tk↓**	**Total↓**	**Method**	**TTFT↓**	**Prefill↓**	**Prune↓**	**Dec/tk↓**	**Total↓**
No pruning	46.68 ms	37.65 ms	0.00 ms	16.72 ms	281.27 ms	No pruning	94.00 ms	79.74 ms	0.00 ms	**15.46 ms**	575.28 ms
VisionZip	31.24 ms	20.18 ms	1.07 ms	16.69 ms	247.92 ms	VisionZip	38.46 ms	24.75 ms	0.93 ms	16.95 ms	446.52 ms
D^2Pruner	32.68 ms	23.51 ms	**0.07 ms**	20.09 ms	309.55 ms	D^2Pruner	50.42 ms	36.32 ms	**0.25 ms**	19.06 ms	602.48 ms
DART	32.81 ms	23.55 ms	2.62 ms	17.92 ms	270.87 ms	DART	48.52 ms	34.23 ms	6.68 ms	18.15 ms	571.74 ms
CiVi	**29.45 ms**	**19.79 ms**	1.29 ms	**16.61 ms**	**245.44 ms**	**CiVi**	**36.74 ms**	**21.85 ms**	2.99 ms	16.56 ms	**359.97 ms**

4.5 Ablation Studies

Figure 5(a) compares single-, pairwise-, and three-cue variants across six benchmarks. The full $S{+}E{+}D$ combination performs best overall, while no individual cue consistently dominates across tasks. S provides the strongest standalone baseline; removing S particularly hurts TextVQA and MMB, whereas removing E causes the largest POPE drop. These task-dependent effects support combining all three cues. Additional ablations are provided in Appendix C.



(a) Cue ablation.

(b) Qualitative POPE comparison.

Figure 5: CiVi on LLaVA-1.5-7B at 88.9% pruning. (a) Cue ablation across six benchmarks, normalized to the $S{+}E{+}D$ variant. (b) Qualitative comparison with CLS-attention top-K selection, where CiVi better preserves sparse, decision-critical visual evidence under the token budget.

4.6 Visualization Analysis

Figure 5(b) shows qualitative POPE examples at 88.9% pruning. CLS-attention tends to retain globally prominent regions while missing small queried objects, whereas CiVi preserves tokens around the relevant evidence and produces correct predictions under the same budget. This behavior is consistent with the POPE gains in Table 1 and illustrates the benefit of prioritizing hard-to-compensate evidence over factual-view saliency alone. Additional multi-ratio comparisons and ablation visualizations are provided in Appendix D.

5 Conclusion and Limitations

We present CiVi, a training-free visual token pruning method that combines factual saliency, intervention-inspired response, and diversity to preserve hard-to-compensate visual evidence under aggressive pruning. Across diverse MLLM settings, CiVi achieves a competitive accuracy–efficiency trade-off. The main limitation is that CiVi estimates irreplaceability through a feature-space surrogate rather than exact decoder-level causal tests; extending counterfactual irreplaceability to broader compression targets, e.g., LLM weight or KV-cache compression, remains future work.

AI USE STATEMENT

We used generative AI tools only sparingly for language editing and formatting assistance during manuscript preparation. We did not use generative AI tools to formulate or refine research hypotheses, design the methodology or experiments, implement the method, process data, analyze or interpret results, or generate experimental figures or data. All AI-assisted text was reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work aims to improve the inference efficiency of multimodal large language models by reducing redundant visual tokens at inference time. More efficient inference can reduce computational cost, memory usage, and energy consumption, making MLLMs more practical for resource-constrained settings. Since the proposed method is applied to existing MLLMs, it may inherit risks of the underlying models, including hallucination, biased responses, and unsafe outputs. Our method does not introduce new datasets, new generative capabilities, private data, or human-subject data, and we evaluate it only in standard research settings. We use publicly available pretrained models, benchmarks, and baseline implementations in accordance with their original licenses and terms of use.

REPRODUCIBILITY STATEMENT

To support reproducibility, Sec. 3 specifies the proposed method and Sec. 4 describes the main experimental setup. Appendix B provides benchmark protocols, model backbones, pruning ratios, comparison methods, implementation hyperparameters, deterministic token-selection details, hardware, and the timing protocol. Additional ablations are reported in Appendix C. An anonymized implementation is available through the link in the abstract.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

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, 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. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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. *arXiv preprint arXiv:2502.13923*, 2025b.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *The Eleventh International Conference on Learning Representations*, 2023.

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, 2024.

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. Scope: Saliency-coverage oriented token pruning for efficient multimodel llms. *Advances in Neural Information Processing Systems*, 38:161527–161552, 2026.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1359–1367. IEEE, 2017.

Zhong Ji, Changxu Meng, Yan Zhang, Haoran Wang, Yanwei Pang, and Jungong Han. Eliminate before align: A remote sensing image-text retrieval framework with keyword explicit reasoning. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 1662–1671, 2024.

Youngeun Kim, Youjia Zhang, Huiling Liu, Aecheon Jung, Sunwoo Lee, and Sungeun Hong. Zoo-prune: 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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024b.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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, 2023c.

Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, Xunliang Cai, Charles Ling, and Boyu Wang. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*, 2024c.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 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.

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/`.

Xuyang Liu, Ziming Wang, Junjie Chen, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Siteng Huang, and Honggang Chen. Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7350–7358, 2026.

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, 2024c.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. OCRBench: On the hidden mystery of OCR in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024d.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022.

Qiankun Ma, Ziyao Zhang, Haofei Wang, Zhen Song, Jie Chen, and Hairong Zheng. Apet: Approximation-error guided token compression for efficient vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26306–26316, 2026.

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.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multimodal LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7614–7623, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9972–9991, 2025.

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. *arXiv preprint arXiv:2410.17247*, 2024.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016.

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, 2025.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. DeCo: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, pp. 57730–57754. PMLR, 2024.

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, 2026.

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

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.

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, 2025b.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *Advances in Neural Information Processing Systems*, 38:39800–39832, 2026.

CONTENTS OF TECHNICAL APPENDICES

A Additional Theoretical Details **16**

A.1 Feature-Space Surrogate Clarification 16

A.2 Stability of Multi-Sample Ranking 16

A.3 Complexity Discussion 16

B Experimental Details and Additional Results **17**

B.1 Benchmarks and Metrics 17

B.2 Existing Assets and Licenses 18

B.3 Backbones, Pruning Ratios, and Compared Methods 18

B.4 Implementation Details 18

B.5 Timing Protocol 18

B.6 Additional Results on LLaVA-1.5-13B 19

B.7 Additional Results on Qwen3-VL-8B 19

B.8 Additional Experimental Results 20

C Additional Ablation Studies **23**

D Additional Visualization Results **24**

D.1 Visualization of Ablation Variants 24

D.2 Comparison with Attention-based Selection 24

E Overall Summary and Limitations **26**

A Additional Theoretical Details

A.1 Feature-Space Surrogate Clarification

This subsection clarifies the feature-space surrogate used in Sec. 3. Directly evaluating token irreplaceability at the decoder-output level requires repeated forward passes over many candidate token interventions. To avoid this cost, CiVi estimates irreplaceability before decoding through a feature-space surrogate.

Specifically, the intervention is applied to encoder visual features before the visual projector, producing intervened features $\tilde{\mathbf{Z}}^{(s)}$. The factual and counterfactual tokens used for scoring are obtained after projection:

$$\mathbf{v}_i = \mathcal{P}(\mathbf{z}_i), \qquad \tilde{\mathbf{v}}_i^{(s)} = \mathcal{P}(\tilde{\mathbf{z}}_i^{(s)}), \tag{12}$$

where $\mathcal{P}(\cdot)$ denotes the visual projector. Since the intervention does not change token indices, $\mathbf{v}_i$ and $\tilde{\mathbf{v}}_i^{(s)}$ remain aligned and comparable in the LLM-aligned feature space.

The discrepancy-enhanced token $\mathbf{v}_i^{\star(s)}$ is used only as a scoring reference for estimating intervention sensitivity. After ranking, CiVi feeds the retained factual tokens, rather than the enhanced tokens, to the frozen decoder. This design preserves compatibility with the original decoder input distribution while avoiding repeated decoder-level interventions.

A.2 Stability of Multi-Sample Ranking

We provide a simple stability interpretation for multi-sample aggregation. Let $X_i^{(s)} = \mathrm{Score}_i^{(s)}$ denote the score of token i under intervention sample s. Assume different intervention samples are independently drawn, with $\mathbb{E}[X_i^{(s)}] = \mu_i$ and $\mathrm{Var}(X_i^{(s)}) \le \sigma_i^2$. The averaged score $\bar{X}_i = \frac{1}{N}\sum_{s=1}^{N} X_i^{(s)}$ then satisfies

$$\mathrm{Var}(\bar{X}_i) = \mathrm{Var}\left(\frac{1}{N}\sum_{s=1}^{N} X_i^{(s)}\right) \le \frac{\sigma_i^2}{N}. \tag{13}$$

Thus, increasing the number of intervention samples reduces the variance of the estimated token score.

We further consider the stability of pairwise ranking. For two tokens (i, j) with mean score gap $\Delta_{ij} = \mu_i - \mu_j > 0$, define $Y_{ij}^{(s)} = X_i^{(s)} - X_j^{(s)}$ and assume $\mathrm{Var}(Y_{ij}^{(s)}) \le \nu_{ij}$. Let $\bar{Y}_{ij} = \bar{X}_i - \bar{X}_j$. Since $\mathbb{E}[\bar{Y}_{ij}] = \Delta_{ij}$ and $\mathrm{Var}(\bar{Y}_{ij}) \le \nu_{ij}/N$, Chebyshev's inequality gives

$$\Pr(\bar{X}_i \le \bar{X}_j) = \Pr(\bar{Y}_{ij} \le 0) \le \frac{\nu_{ij}}{N\Delta_{ij}^2}. \tag{14}$$

This indicates that multi-sample aggregation reduces the probability of pairwise ranking inversion at an $\mathcal{O}(1/N)$ rate, explaining why averaged interventions yield more stable retained sets than a single perturbation.

A.3 Complexity Discussion

Let T be the original number of visual tokens and $T' = (1-\rho)T$ be the retained number after pruning. The pruning module is executed once before decoding. Its one-shot cost is

$$\mathcal{C}_{\mathrm{once}} = \mathcal{O}\big(NTd_p + NT^2 d_p + T\log K\big), \tag{15}$$

where N is the number of intervention samples, d_p is the projected feature dimension, and $K = T'$. The three terms correspond to intervention construction and projection, pairwise similarity computation for diversity, and top-K selection, respectively.

After pruning, the decoder processes only T' visual tokens instead of T. Ignoring text-only terms, the visual-token-dependent cost of a decoder layer can be abstractly written as

$$\mathcal{C}_{\mathrm{full}} = aT^2 d + bTd^2 + cTmd, \qquad \mathcal{C}_{\mathrm{prune}} = aT'^2 d + bT'd^2 + cT'md, \tag{16}$$

where d is the hidden dimension, m is the text length, and a, b, c absorb implementation-dependent constants. Since $\mathcal{C}_{\text{once}}$ is paid only once, while the saving $\mathcal{C}_{\text{full}} - \mathcal{C}_{\text{prune}}$ accumulates over L_{dec} decoder layers, CiVi is beneficial when

$$L_{\text{dec}} \left(\mathcal{C}_{\text{full}} - \mathcal{C}_{\text{prune}}\right) > \mathcal{C}_{\text{once}}. \tag{17}$$

This explains why the one-shot scoring overhead can be offset by accumulated decoder-side savings, especially under long visual sequences and high pruning ratios.

B Experimental Details and Additional Results

B.1 Benchmarks and Metrics

We evaluate CiVi on both image and video understanding benchmarks, following the default evaluation protocols used in prior MLLM pruning work (Chen et al., 2024; Zhang et al., 2025b). These benchmarks cover general perception, visual reasoning, object hallucination, text-rich understanding, OCR-sensitive reasoning, and video question answering. Unless otherwise specified, we report the official benchmark score or answer accuracy provided by each evaluation protocol.

MME. MME (Fu et al., 2025) evaluates MLLM perception and cognition, including object recognition, attribute understanding, commonsense reasoning, and numerical judgment. We report the official MME score.

MMBench. MMBench (Liu et al., 2024c) is a multi-choice benchmark for broad visual understanding and reasoning, covering recognition, localization, attribute comparison, relation reasoning, and visual commonsense. We report the official score.

POPE. POPE (Li et al., 2023c) evaluates object hallucination through binary object-existence questions. It is relevant to high-ratio pruning because removing sparse object evidence can directly induce false predictions. We report accuracy under the standard protocol.

TextVQA. TextVQA (Singh et al., 2019) evaluates question answering that requires reading and reasoning over scene text, testing whether pruning preserves fine-grained text-bearing regions. We report default benchmark accuracy.

GQA. GQA (Hudson & Manning, 2019) focuses on compositional visual reasoning over real-world scenes, including attributes, relations, spatial layout, and multi-hop questions. We report standard validation accuracy.

ScienceQA. ScienceQA (Lu et al., 2022) contains multimodal science questions with image context and multiple-choice answers, requiring visual, external, and commonsense reasoning. We report the ScienceQA-IMG accuracy.

VQA-v2. VQA-v2 (Goyal et al., 2017) evaluates open-ended visual question answering on natural images, covering recognition, counting, color, spatial relation, and scene understanding. We report official VQA accuracy.

OCRBench. OCRBench (Liu et al., 2024d) evaluates OCR-sensitive multimodal understanding, including text recognition, text-based reasoning, and document-like visual content. We report the official OCRBench score.

MM-Vet. MM-Vet (Yu et al., 2024) evaluates integrated multimodal capabilities such as recognition, knowledge, spatial reasoning, OCR, and grounded response generation. We report the official score when available.

TGIF-QA, MSVD-QA, and MSRVTT-QA. TGIF-QA (Jang et al., 2017), MSVD-QA (Xu et al., 2017), and MSRVTT-QA (Xu et al., 2016) evaluate video question answering on Video-LLaVA-7B (Lin et al., 2024), testing whether pruning preserves informative frame-level evidence under long visual sequences. We report answer accuracy under the default protocols.

B.2 Existing Assets and Licenses

We use publicly available pretrained MLLM backbones, evaluation benchmarks, and baseline implementations for research evaluation. All models, datasets, and codebases are credited through their original papers and repositories, and we follow their original licenses and terms of use. We do not introduce new datasets or collect any private data.

B.3 Backbones, Pruning Ratios, and Compared Methods

Models. We evaluate CiVi on six representative open-source MLLMs: LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-1.5-13B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and Video-LLaVA-7B (Lin et al., 2024). LLaVA-1.5-7B (Liu et al., 2024a) serves as the main image-understanding backbone, LLaVA-1.5-13B (Liu et al., 2024a) tests scalability within the same family, and LLaVA-NeXT-7B (Liu et al., 2024b) evaluates high-resolution inputs with up to 2880 visual tokens. Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B (Bai et al., 2025a) test cross-architecture transfer, while Video-LLaVA-7B (Lin et al., 2024) covers multi-frame video inputs. The LLaVA-1.5-13B (Liu et al., 2024a) and Qwen3-VL-8B (Bai et al., 2025a) results are reported as additional generalization studies in the appendix.

Pruning ratios. For LLaVA-1.5-7B (Liu et al., 2024a) and LLaVA-1.5-13B (Liu et al., 2024a), we evaluate 66.7%, 77.8%, and 88.9% pruning, retaining 192, 128, and 64 visual tokens in the 576-token LLaVA-1.5 setting. For LLaVA-NeXT-7B (Liu et al., 2024b), we use 88.9% pruning and retain 320 tokens from high-resolution inputs. For Qwen2.5-VL-7B (Bai et al., 2025b), we evaluate 80% and 90% pruning, while Qwen3-VL-8B (Bai et al., 2025a) follows the LLaVA-1.5 pruning ratios. For Video-LLaVA-7B (Lin et al., 2024), all methods retain 256 of 2048 visual tokens, corresponding to 87.5% pruning.

Comparison methods. We compare against representative training-free acceleration baselines in our main-result tables, including HoloV (Zou et al., 2026), SparseVLM (Zhang et al., 2025b), LLaVA-PruMerge (Shang et al., 2025), VisionZip (Yang et al., 2025), PDrop (Xing et al., 2024), FiCoCo-V (Han et al., 2026), FastV (Chen et al., 2024), FasterVLM (Zhang et al., 2025a), HiRED (Arif et al., 2025), and GlobalCom^2 (Liu et al., 2026), together with the Vanilla no-pruning upper bound for efficiency reference. These methods cover attention-based pruning, token merging, hierarchical retention, and compression-allocation strategies. Missing entries indicate unreported or non-comparable results in prior work.

B.4 Implementation Details

CiVi is applied once before decoding without additional training, backbone finetuning, or architecture-specific modification. For each input, we intervene on encoder visual features as described in Sec. 3, project factual and intervened features into the LLM-aligned space for scoring, and feed the retained factual tokens to the frozen decoder in their original spatial or temporal order. Unless otherwise stated, we use diffusion-style intervention (Ho et al., 2020), aggregate $N = 3$ intervention samples, set $\alpha_v = 0.5$, and determine the retained budget by $K = \max(1, \mathrm{round}(T(1-\rho)))$. During inference, perturbation masks and Gaussian noise are generated with deterministic image-dependent seeds derived from per-image feature statistics and the intervention-sample index, ensuring reproducible token selection under the same model and environment. For fair comparison, all methods use the same backbone, pruning ratio, and retained-token budget; baseline settings follow their original papers or official implementations. Most experiments are conducted on NVIDIA RTX 4090 GPUs, while LLaVA-1.5-13B (Liu et al., 2024a) and Qwen3-VL-8B (Bai et al., 2025a) use NVIDIA L20 GPUs. Sensitivity to intervention count and discrepancy scaling is reported in Appendix C.

B.5 Timing Protocol

For efficiency comparison, all methods in Table 5 are evaluated on a fixed 200-sample GQA subset using one NVIDIA H800, batch size 1, ten warm-up samples, and three timed repeats per sam-

ple. All pruners target 90% removal under the same decoding configuration. CUDA events with synchronization delimit the measured regions. *TTFT* spans visual processing through the first decoder forward pass; *Prefill* includes visual encoding, projection, method-specific token selection, and language-model prefill; *Prune* isolates the one-shot scoring and selection region; *Decode/token* normalizes autoregressive decode time by the number of decoded steps; and *Total* spans visual processing through generation completion. Because different methods can generate different output lengths, Total is an end-to-end measurement but not an isolated comparison of selector speed. CiVi's Prune measurement includes perturbation-based evidence scoring and its $\mathcal{O}(NT^2d_p)$ pairwise diversity computation. This cost grows from 1.29 ms on LLaVA-1.5 to 2.99 ms on LLaVA-NeXT's longer AnyRes sequences.

B.6 Additional Results on LLaVA-1.5-13B

To further examine scalability to a larger LLaVA-family backbone, we evaluate CiVi on LLaVA-1.5-13B (Liu et al., 2024a) under the same pruning ratios as the 7B setting. Table 6 reports results on six image understanding benchmarks. CiVi consistently achieves the best average performance across all pruning ratios, retaining 99.8%, 98.3%, and 94.4% of the original average performance at 66.7%, 77.8%, and 88.9% pruning, respectively. The advantage remains clear under aggressive pruning: at 88.9% pruning, CiVi outperforms both SparseVLM and VisionZip on five of six benchmarks and improves the average retention from 93.6% to 94.4%. These results indicate that the proposed intervention-grounded ranking remains effective when scaling from LLaVA-1.5-7B (Liu et al., 2024a) to the larger 13B backbone.

Table 6: Additional results on LLaVA-1.5-13B under three pruning ratios.

Method	MME	POPE	VQA_{Text}	GQA	SQA	MM-Vet	Avg (%)
LLaVA-1.5-13B	1818	85.9	61.3	63.2	72.8	35.3	100.0%
Pruning Ratio = 66.7%							
SparseVLM (ICML'25)	1768	82.2	45.4	58.7	73.1	35.5	93.5%
VisionZip (CVPR'25)	1754	85.1	59.5	59.1	73.5	37.5	98.9%
CiVi (Ours)	**1779**	**86.9**	**59.6**	**59.9**	**73.6**	**37.6**	**99.8%**
Pruning Ratio = 77.8%							
SparseVLM (ICML'25)	**1774**	81.1	49.9	57.9	69.9	34.1	92.3%
VisionZip (CVPR'25)	1743	85.2	58.7	57.9	**74.0**	**37.5**	98.3%
CiVi (Ours)	1748	**85.6**	**59.1**	**58.8**	**74.0**	36.4	**98.3%**
Pruning Ratio = 88.9%							
SparseVLM (ICML'25)	1402	65.0	47.1	50.6	69.0	30.0	81.6%
VisionZip (CVPR'25)	1676	76.0	57.4	56.2	74.4	**33.9**	93.6%
CiVi (Ours)	**1678**	**81.6**	**57.6**	**57.2**	**74.6**	32.5	**94.4%**

B.7 Additional Results on Qwen3-VL-8B

To further examine cross-architecture generalization on a stronger Qwen-family backbone, we evaluate CiVi on Qwen3-VL-8B (Bai et al., 2025a) under the same pruning ratios as the LLaVA-1.5 setting. Table 7 reports results on five image understanding benchmarks. CiVi consistently improves the average performance over SparseVLM across all pruning ratios, retaining 93.9%, 90.8%, and 89.1% of the original average performance at 66.7%, 77.8%, and 88.9% pruning, respectively. The advantage becomes more pronounced under aggressive pruning: at 88.9% pruning, CiVi outperforms SparseVLM on all five benchmarks and improves the average retention from 80.3% to 89.1%. These results indicate that the proposed intervention-grounded ranking generalizes well to the Qwen3-VL architecture.

Table 7: Additional results on Qwen3-VL-8B under three pruning ratios.

Method	MME	SQA	POPE	GQA	VQA_{Text}	Avg (%)
Qwen3-VL-8B	2319	93.8	86.6	61.6	82.1	100.0%
Pruning Ratio = 66.7%						
SparseVLM (ICML'25)	**2148**	87.4	87.4	53.8	50.8	87.2%
CiVi (Ours)	2035	**87.9**	**87.7**	**55.1**	**80.0**	**93.9%**
Pruning Ratio = 77.8%						
SparseVLM (ICML'25)	**1964**	85.3	83.8	52.9	72.4	89.3%
CiVi (Ours)	1899	**85.5**	**86.1**	**53.4**	**77.8**	**90.8%**
Pruning Ratio = 88.9%						
SparseVLM (ICML'25)	1699	81.0	75.6	47.1	64.1	80.3%
CiVi (Ours)	**1882**	**82.3**	**85.5**	**52.2**	**76.6**	**89.1%**

B.8 Additional Experimental Results

This subsection collects evaluations beyond the standard single-image QA matrix.

B.8.1 Matched Qwen2.5-VL Baselines

Table 8 compares methods on the same Qwen2.5-VL-7B-Instruct backbone. The leading method varies by task: CiVi is strongest on four of six reported metrics at the two tighter budgets, while VisionZip or D^2Pruner leads some ScienceQA/POPE results and DART leads MME perception at 77.8% and 88.9%.

Table 8: Matched Qwen2.5-VL-7B-Instruct comparison. MME-P/C are perception and cognition scores.

Method	POPE	GQA	SQA-I	MMB	MME-P	MME-C
No pruning	86.10	60.50	76.70	82.80	1711.59	592.64
Pruning Ratio = 66.7%						
DART	85.87	58.25	74.62	80.24	1675.16	582.50
D^2Pruner	85.92	55.53	**77.64**	77.41	1585.69	581.79
VisionZip	86.28	57.85	77.49	**81.70**	1669.23	**591.71**
CiVi	**86.63**	**58.85**	76.50	**81.70**	**1708.82**	589.64
Pruning Ratio = 77.8%						
DART	83.44	56.48	73.57	78.52	**1644.92**	560.00
D^2Pruner	84.08	52.32	76.55	74.74	1538.32	548.57
VisionZip	84.68	56.74	**76.80**	80.07	1629.38	544.29
CiVi	**85.59**	**57.14**	75.86	**80.90**	1641.83	**584.64**
Pruning Ratio = 88.9%						
DART	78.71	52.04	72.68	73.54	**1556.36**	517.14
D^2Pruner	78.04	46.04	73.57	61.51	1340.70	475.00
VisionZip	**82.00**	52.52	73.97	75.86	1524.34	421.07
CiVi	80.84	**54.15**	**74.67**	**76.00**	1528.80	**522.79**

B.8.2 Turn-wise Multi-round Dialogue

We evaluate an oracle-context protocol on Qwen2.5-VL-7B-Instruct: the complete dialogue history is included in every prompt and visual tokens are reselected at each turn. MultiVerse contains 2,536 turn-level samples from 647 single-image dialogues, and MMDU contains 1,645 samples from 110 multi-image dialogues. Table 9 supports turn-wise applicability under this protocol.

Table 9: Turn-level ROUGE-L under the oracle-context multi-round protocol.

Method	MultiVerse	MMDU
No pruning	0.2076	0.2426
Pruning Ratio = 66.7%		
DART	0.2038	0.2424
VisionZip	0.2055	**0.2432**
CiVi	**0.2058**	0.2428
Pruning Ratio = 77.8%		
DART	0.2010	0.2423
VisionZip	0.2027	**0.2426**
CiVi	**0.2040**	**0.2426**
Pruning Ratio = 88.9%		
DART	0.1987	0.2421
VisionZip	0.2013	0.2424
CiVi	**0.2015**	**0.2425**

B.8.3 Additional Video Reasoning

Table 10 adds MVBench and TempCompass together with fixed 500-sample NextQA-MC and PerceptionTest-val-MC evaluations. CiVi leads VisionZip on MVBench and TempCompass at every ratio, but the two methods trade places on the remaining tasks.

Table 10: Additional video reasoning results. NextQA and PerceptionTest use fixed 500-sample subsets.

Method	MVBench	TempCompass	NextQA	PerceptionTest
No pruning	43.15	50.00	59.2	60.0
Pruning Ratio = 66.7%				
VisionZip	43.30	49.51	57.4	**56.0**
CiVi	**43.42**	**50.56**	**57.6**	55.8
Pruning Ratio = 77.8%				
VisionZip	42.88	49.24	56.6	54.4
CiVi	**43.50**	**49.68**	**57.6**	**55.6**
Pruning Ratio = 88.9%				
VisionZip	42.15	48.04	**56.8**	**55.2**
CiVi	**42.92**	**48.23**	**56.8**	54.8

B.8.4 Image Captioning

We additionally measure CIDEr on fixed 500-sample lite splits of COCO, NoCaps, Flickr30k, and TextCaps (2,000 samples total) with LLaVA-1.5-7B. As Table 11 shows, CiVi is strongest on COCO and NoCaps at both ratios, while D^2Pruner or DART leads TextCaps.

Table 11: CIDEr on fixed 500-sample lite captioning splits.

Method	COCO	NoCaps	Flickr30k	TextCaps
No pruning	1.0472	0.9859	0.8506	1.0779
Pruning Ratio = 77.8%				
DART	1.0280	0.9301	**0.8512**	1.0770
D^2Pruner	0.9675	0.8984	0.7674	**1.1320**
VisionZip	0.9892	0.9171	0.7872	1.0212
CiVi	**1.0346**	**0.9710**	0.8052	1.0375
Pruning Ratio = 88.9%				
DART	0.9600	0.9139	0.7512	**1.0770**
D^2Pruner	0.7738	0.7121	0.5981	0.9942
VisionZip	0.9032	0.8138	0.7007	0.9322
CiVi	**0.9897**	**0.9375**	**0.7660**	0.9590

C Additional Ablation Studies

Table 12 reports additional ablations at 88.9% pruning. Using all three factors jointly gives the strongest overall performance. Intervention-aware scoring benefits from moderate multi-sample aggregation, with $N = 3$ providing the best overall trade-off between stability and one-shot scoring overhead. Increasing N to 5 gives comparable results on some metrics but does not consistently improve the overall performance, while further increasing the pre-decoding cost. The α_v block follows the sensitivity-analysis protocol used in Fig. 6; it shows that the method is robust over a broad range of α_v, with the strongest average relative performance around $\alpha_v \in [0.5, 1.0]$.

Table 12: Additional ablation results at 88.9% pruning.

Part	Setting	MME	VQA_{Text}	POPE	SQA	GQA	MMB
Factor	all	**1708**	**55.9**	**83.5**	**69.9**	**56.8**	**62.8**
	saliency only	1660	55.5	80.2	69.0	55.0	60.8
	evidence only	1519	46.8	72.9	65.6	52.4	51.3
	diversity only	1534	52.3	79.6	68.5	53.4	54.9
Intervention	no intervention	1678	55.8	80.7	68.7	55.1	61.5
	one intervention	1690	54.7	82.1	69.0	55.9	61.9
	three interventions	**1708**	**55.9**	**83.5**	**69.9**	**56.8**	**62.8**
	five interventions	**1708**	55.8	**83.5**	68.8	56.4	61.2
	seven interventions	1684	55.5	82.6	68.8	56.2	61.1
α_v	0	1669	55.4	82.7	69.4	55.8	61.3
	0.25	1674	55.6	82.9	69.4	56.1	61.5
	0.5	1703	**55.9**	**83.5**	**69.9**	**56.8**	**62.8**
	0.75	**1708**	55.5	83.2	69.7	56.1	62.3
	1.0	1703	55.2	82.7	69.5	56.2	62.1
	1.5	1699	55.5	82.5	69.1	56.0	61.7
Score fusion	Adaptive fusion (Ours)	**1708**	**55.9**	**83.5**	**69.9**	**56.8**	**62.8**
	Average fusion	1584	53.1	80.5	68.7	54.7	58.3
	Geometric fusion	1691	54.5	82.4	68.6	55.8	61.2
Perturbation type	Diffusion perturbation (Ours)	**1708**	**55.9**	**83.5**	**69.9**	**56.8**	**62.8**
	Gaussian perturbation	1705	55.6	82.4	69.2	56.0	61.1
	Mean replacement	1681	**55.9**	82.7	69.2	56.1	61.2

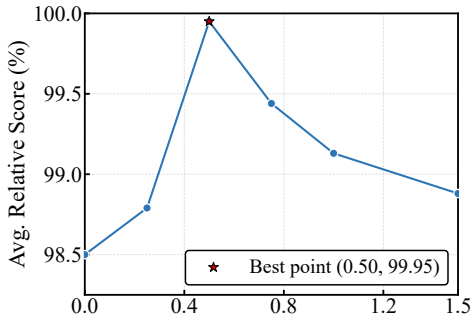


Figure 6: Sensitivity to α_v at 88.9% pruning. The average relative score peaks at $\alpha_v = 0.5$.

D Additional Visualization Results

This appendix provides additional qualitative visualizations complementing the POPE case study in Sec. 4.6. We first present a multi-ratio comparison between CiVi and an attention-based baseline, then show ablation visualizations.

D.1 Visualization of Ablation Variants

Figure 7 visualizes two design dimensions of CiVi: different fusion strategies for combining factual-view saliency, intervention response, and diversity, and different counterfactual intervention variants. Blacked-out regions denote discarded visual tokens, while visible regions correspond to retained ones. The qualitative results are consistent with the quantitative ablations in the main paper. In the first example, CiVi retains more tokens on the human body while maintaining relatively balanced spatial coverage over the whole image. In contrast, alternative fusion or perturbation variants tend to allocate tokens less stably, either missing parts of the foreground subject or concentrating on less informative regions. In the second example, CiVi better preserves evidence around both cats and keeps a broader scene-level view, which is important for recognizing the visual content under a tight token budget. These observations suggest that the default design yields more stable and semantically meaningful token retention patterns.

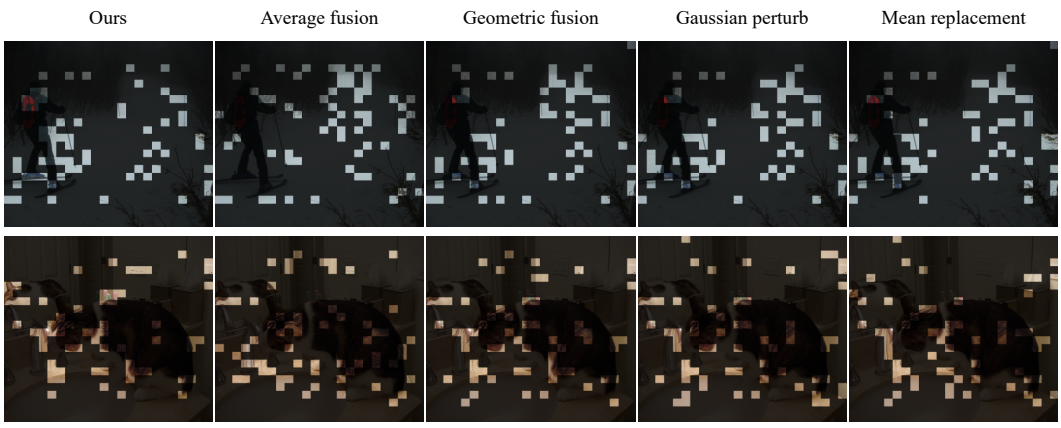


Figure 7: Qualitative visualization of the ablation settings in Sec. 4.5. The figure compares different fusion strategies and counterfactual intervention variants. Consistent with the quantitative results, the default design yields more stable retained-token patterns and better preserves semantically useful evidence under tight token budgets.

D.2 Comparison with Attention-based Selection

Figure 8 compares CiVi with a CLS-attention top-K baseline across representative examples and multiple pruning ratios. Blacked-out regions denote discarded visual tokens. As pruning becomes more aggressive, the attention baseline tends to over-concentrate on visually prominent yet compensable regions, whereas CiVi better preserves sparse but decision-critical evidence and maintains a retained set with broader coverage of sparse, decision-critical visual evidence.

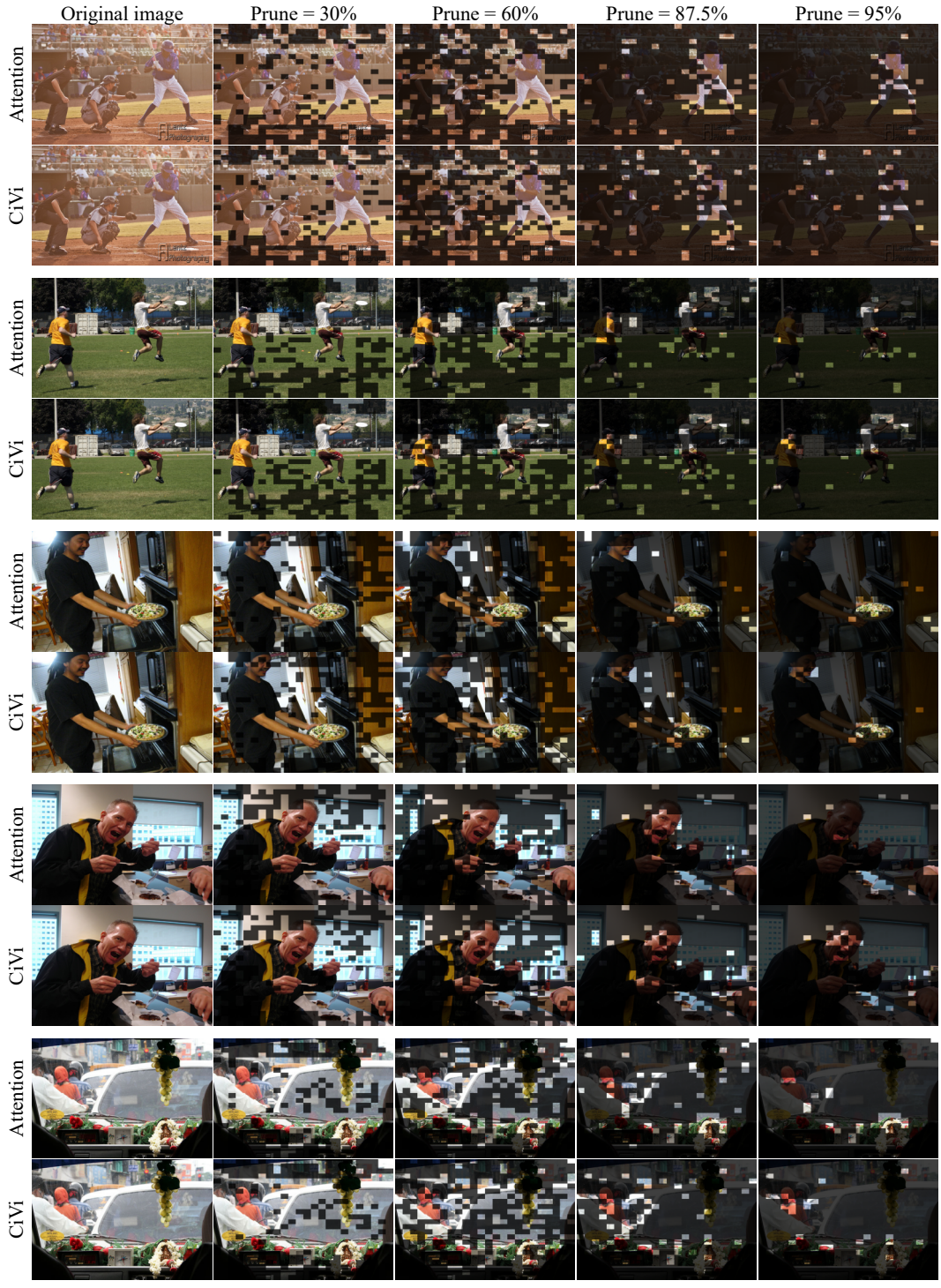


Figure 8: Qualitative comparison between a CLS-attention top-K baseline and CiVi across representative examples and pruning ratios. As the retained budget becomes tighter, the attention baseline increasingly focuses on visually prominent but compensable regions, whereas CiVi better preserves sparse visual evidence useful for the evaluated prediction.

E Overall Summary and Limitations

This paper revisits training-free VTP from the perspective of hard-to-compensate visual evidence and introduces CiVi, a plug-and-play method for retaining such evidence under tight token budgets. Instead of assuming that factual-view prominence alone reliably indicates irreplaceability, we use controlled feature-space perturbations to construct intervention-response and discrepancy-enhanced scoring references. CiVi then performs token selection by jointly considering factual-view saliency, intervention response, and diversity, while the decoder receives the retained clean tokens. Direct decoder-intervention measurements and experiments across multiple backbones, pruning ratios, and video settings support a strong accuracy–efficiency trade-off, especially under aggressive pruning.

Limitations and future work. Despite its effectiveness, CiVi also leaves several open directions. First, its feature-space score approximates an expensive decoder-intervention reference and does not identify a causal effect; direct measurements over larger candidate sets or full generations would require many repeated decoder passes. Second, the current formulation focuses on visual token pruning in MLLMs, while the broader principle of preserving hard-to-compensate information may also benefit KV-cache reduction or cross-modal memory compression. Finally, broader scenarios such as multi-image dialogue, longer video reasoning, and more complex multimodal tasks remain worth further exploration.