STILL THERE, NO LONGER SEEN: EXPOSING COMPRESSION-INDUCED RISK IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual token compression reduces the inference cost of Large Vision-Language Models (LVLMs). However, aggregate robustness measures do not reveal whether a particular adversarial failure is induced by compression or inherited from the underlying model. We define a *compression-specific failure* (CSF) as an adversarial input that remains correct under full-token inference but fails after compression, casting compression-induced risk as a paired failure attribution problem. Within a controlled diagnostic cohort, counterfactuals show that retained-set allocation causally changes compressed correctness and reveal a negative association between recovery and representation drift in displaced evidence. Motivated by these findings, we propose **CIRA**, a **C**ompression-**I**nduced **R**isk **A**ttack for Large Vision-Language Models. Under a vision-encoder white-box setting, CIRA optimizes image perturbations through encoder-side objectives that manipulate token priorities across candidate compression budgets while preserving displaced evidence. CIRA uses no downstream questions or labels and requires no access to the language model, deployed compressor, or exact compression budget. Across 12 dataset–compressor settings evaluated at four budgets, CIRA achieves a mean CSFR of 20.35% while limiting full-token attack success to 6.92%, with similar behavior on additional LVLM families. A cross-view selection-stabilization defense substantially suppresses CIRA, although Adaptive CIRA partially restores its effectiveness. These results show that compression-specific failures persist under restricted access and support paired evaluation of full-token and compressed inference for attributing risk to visual-token compression. Code is provided in the anonymous repository.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) support broad visual understanding by mapping images to long visual-token sequences (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2024b). Visual-token compression reduces this inference cost through token selection, aggregation, and compact visual representations (Shang et al., 2025; Yang et al., 2025b; Li et al., 2025; Bulat et al., 2026). Decoder-side methods further reduce visual-token computation within the language model (Chen et al., 2024; Zhang et al., 2024; Xing et al., 2024). Meanwhile, adversarial attacks show that failures can be induced in the underlying LVLM through image-space or joint-modal perturbations (Zhang et al., 2022; Schlarmann & Hein, 2023; Yin et al., 2023). This raises the question: *what risks are induced by visual-token compression itself, beyond vulnerabilities already present in the underlying LVLM?*

When a full-token LVLM is used for robustness assessment while a compressed variant is deployed, failures confined to the compressed path are not captured by that assessment. We therefore evaluate each adversarial input with and without token compression. A *compression-specific failure* (CSF) occurs when full-token inference remains correct but compressed inference fails. This paired criterion distinguishes failures introduced by compression from those already present in the underlying LVLM.

Recent studies increasingly examine robustness under visual-token compression. Evidence shows that compression can reshape robustness in either direction, depending on how visual evidence is selected and retained (Wang et al., 2026a; Gu et al., 2026). Compression-aware attacks further demonstrate that adversarial behavior can depend on the compression path itself (Zhang et al., 2026c;a). These

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.

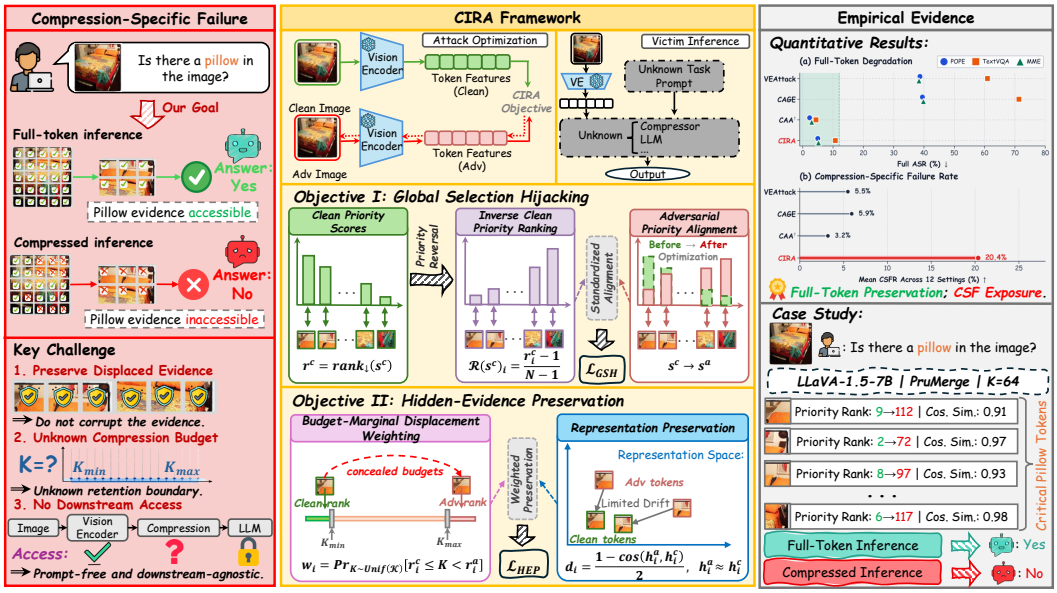


Figure 1: Overview of the compression-specific failure setting, the CIRA attack framework and the resulting behavior under visual-token compression.

developments shift the question from whether compression affects robustness to which adversarial failures can be attributed specifically to compression and whether they can be selectively induced under deployment uncertainty. Specifically, can a single perturbation optimized through the target vision encoder preserve full-token correctness while inducing compressed-path failures when the deployed compressor and budget are unknown during optimization?

To understand what governs this selectivity, we use controlled counterfactuals to isolate two factors. Retained-set interventions causally change compressed correctness by altering which visual evidence remains accessible after compression. Meanwhile, recovery is negatively associated with representation drift in displaced evidence. These results directly inform the attack design: reallocate token priority while preserving displaced evidence.

In this direction, we introduce **CIRA** (**C**ompression-**I**nduced **R**isk **A**ttack), an attack framework for inducing compression-specific failures. As illustrated in Figure 1, CIRA combines Global Selection Hijacking (GSH), which globally reallocates encoder-side proxy priorities, with Hidden-Evidence Preservation (HEP), which limits representation drift in displaced clean high-priority tokens. With white-box access restricted to the vision encoder, CIRA uses no downstream questions or labels and requires no access to the language model, deployed compressor, or exact compression budget. Across four compressors, three datasets, four budgets, and multiple LVLMs, CIRA induces compression-specific failures while keeping full-token degradation limited. We also introduce Translation-Consensus Selection (TCS), a cross-view selection-stabilization defense, and evaluate it against both standard and adaptive CIRA.

Our contributions are threefold:

❶ ***Compression Risk Attribution.*** We formulate compression-induced risk through a clean-conditioned paired CSF criterion and show that retained-set allocation causally affects compressed correctness while recovery decreases with representation drift.

❷ ***Compression-Induced Risk Attack.*** We introduce CIRA, a target-encoder-only attack that combines priority reallocation with evidence preservation and induces paired CSFs across compressor–budget settings without configuration-specific optimization.

❸ ***Comprehensive Evaluation and Analysis.*** We evaluate CIRA across four compressors, three datasets, four budgets, and multiple LVLM families, with mechanistic evidence of global priority reallocation and limited representation drift in displaced evidence. We also test TCS as a selection-stabilization defense against both standard and adaptive CIRA.

2 Related Work

Visual token compression. Token-selection methods exploit visual cues, diversity, and salience–coverage objectives (Zhang et al., 2025b; Alvar et al., 2025; Xu et al., 2026), while learned pruning addresses limitations of attention-based importance estimates (Takezoe et al., 2026). Hybrid methods combine pruning with clustering and merging (Endo et al., 2025; Dhouib et al., 2025; Yang et al., 2025a). Query-conditioned methods use textual instructions for token scoring or aggregation (Yu et al., 2026; Gao et al., 2026; Sun et al., 2026). Adaptive and progressive methods vary token reduction across inputs and layers (Ye et al., 2025; Chen et al., 2026; Li et al., 2026a), while video and multi-turn approaches address temporal redundancy and evolving context (Wang et al., 2025a; Shen et al., 2024; Li et al., 2026b; Wang et al., 2026c).

Adversarial robustness of vision-language models. Transfer-based attacks exploit set-level guidance, prompt-robust objectives, and iterative multimodal alignment (Lu et al., 2023; Luo et al., 2024; Liu et al., 2024a; Xie et al., 2025). Encoder-based and self-supervised attacks support cross-task and cross-model transfer (Zhang et al., 2025a; Hu et al., 2025; Zhang et al., 2026b). VEAttack disrupts visual representations (Mei et al., 2026b), while PA-Attack uses prototype-guided gray-box attacks (Mei et al., 2026a). Defenses include adversarial encoder fine-tuning (Schlarmann et al., 2024) and adversarial pre-training and instruction tuning (Wang et al., 2025c). Test-time defenses use prompt adaptation and augmented-view consistency (Sheng et al., 2025; Liu et al., 2025).

Robustness under visual token compression. Visual-token compression introduces distinct robustness risks and defense opportunities. Safety-Aware Pruning (SAP) (Wang et al., 2026a) mitigates pruning-induced vulnerabilities, while robustness-oriented pruning (Gu et al., 2026) removes visually misaligned tokens. On the attack side, CAGE (Zhang et al., 2026c) targets tokens expected to survive unknown compression settings. CAA (Zhang et al., 2026a) directly manipulates token-selection rankings, with white-box attacks tailored to known compression configurations and transfer attacks using surrogate models. Our focus is on selectively inducing compression-specific failures under encoder-only access, without knowledge of the deployed compressor or budget.

3 Diagnosing Compression-Specific Failures

In this section, we investigate why adversarial inputs can fail only after compression and use controlled diagnostics to inform the design of CIRA.

3.1 Problem Setup and Compression-Specific Failure

Let $\mathcal{C}$ denote a visual-token compressor and K its compression budget. For a dataset $\mathcal{D} = \{(x_i, q_i, y_i)\}_{i=1}^{M}$, let $f_\theta(x_i, q_i)$ and $f_\theta^{\mathcal{C},K}(x_i, q_i)$ denote full-token and compressed inference, respectively. Given a task-specific binary evaluator $\chi(\hat{y}, y) \in \{0, 1\}$, we define

$$c_i(x) = \chi(f_\theta(x, q_i), y_i)\,, \qquad c_{i,\mathcal{C},K}(x) = \chi\Big(f_\theta^{\mathcal{C},K}(x, q_i), y_i\Big)\,.$$

For a fixed $\mathcal{C}$, we write $c_{i,K}(x) \equiv c_{i,\mathcal{C},K}(x)$ and omit the compressor subscript below.

For each compression budget K, we restrict evaluation to samples for which both full-token and compressed inference produce correct answers before attack:

$$\mathcal{S}_K = \{i \mid c_i(x_i) = 1,\ c_{i,K}(x_i) = 1\}\,. \tag{1}$$

A *compression-specific failure* (CSF) occurs when an adversarial input remains correct under full-token inference but fails after compression:

$$\mathcal{F}_K^{\mathrm{CSF}} = \big\{i \in \mathcal{S}_K \;\big|\; c_i(x_i^{\mathrm{adv}}) = 1,\ c_{i,K}(x_i^{\mathrm{adv}}) = 0\big\}\,. \tag{2}$$

Aggregate accuracy can obscure instance-level prediction changes introduced by VLM acceleration (Sun et al., 2025). Our paired definition focuses on adversarial inputs that remain correct under full-token inference but fail after compression.

3.2 Retained-Set Allocation and Compressed Correctness

Diagnostic setting. We test whether retained-set allocation can change compressed correctness while the adversarial encoder state and compression capacity remain fixed. LLaVA–VisionZip exposes direct-token membership, providing a controlled interface for retained-set intervention. From each downstream-agnostic VEAttack trajectory, we retain the earliest checkpoint satisfying the CSF criterion, termed a *first-observed CSF*.

Counterfactual intervention. For each CSF, we exchange equal numbers of retained and omitted direct tokens, reconstruct compression, and rerun inference at the same encoder state and capacity. A post-hoc answer-aware oracle selects the exchanged token identities (**Appendix C**).

Recovery criteria. Let ρ denote the fraction of exchanged direct-token slots. We compare guided reallocation with a size-matched random exchange. Cumulative recovery credits correction at any tested $\rho' \leq \rho$, whereas exact recovery uses only ρ.

☆ Observation 1: Controlled retained-set reallocation changes compressed correctness. Figure 2 shows that guided recovery exceeds matched-random recovery for every tested budget–fraction pair with $\rho > 0$, with an average advantage of 22.5–24.1 pp across the 12 dataset–budget cells. With encoder state, capacity, and exchange size fixed, the contrast shows that robustness depends on token identity, not only on retention count.

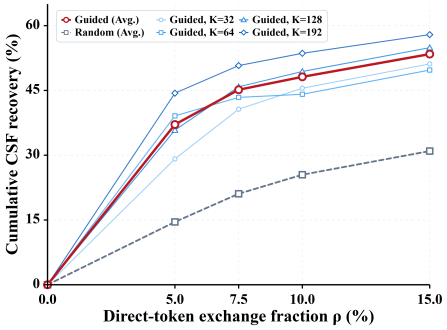


Figure 2: Cumulative CSF recovery under guided and matched-random retained-set exchanges across direct-token exchange fractions.

3.3 Retained-Set Allocation Components and Representation Drift

Having established retained-set allocation sensitivity, we next separate restoration from removal and ask when restored evidence remains useful. Observations 2–3 use $\rho^* = 7.5\%$ as a shared intermediate diagnostic point, where guided and random recovery are already clearly separated. Representation drift is the restored tokens' mean half-cosine distance, $[1 - \cos(\mathbf{h}^c, \mathbf{h}^a)]/2$, between their clean and adversarial representations; Table 1(b) reports this distance multiplied by 100.

☆ Observation 2: Recovery depends on what is restored and removed. Table 1(a) decomposes guided reallocation into restoring omitted tokens with high retrospective answer support and removing retained tokens with low support; larger scores indicate stronger support for the reference answer. Exact recovery rises from 17.4% under matched random exchange to 42.0% under guided reallocation. Both restoration and removal contribute to recovery, with their coordinated use providing the largest gain.

Table 1: Mechanistic diagnostics at $\rho^* = 7.5\%$: exact recovery under retained-set interventions (a) and evidence-restoration effects across representation-drift tertiles (b), with construction and aggregation details in **Appendix C**.

(a) Retained-Set Counterfactuals

Retained-Set Change	Exact Recovery (%)				Avg.
	$K = 192$	$K = 128$	$K = 64$	$K = 32$	
Random	17.9	17.2	18.4	16.1	**17.4**
+ Evidence restoration	31.1	33.5	34.0	34.9	**33.4**
+ Low-support removal	34.5	31.9	25.4	20.1	**28.0**
Guided reallocation	46.4	44.4	38.5	38.6	**42.0**

(b) Association with Representation Drift

Drift Tertile	Mean Drift (%)	Evidence Effect (pp)				Avg.
		$K = 192$	$K = 128$	$K = 64$	$K = 32$	
Lower	31.4	+16.6	+18.9	+18.2	+28.5	**+20.5**
Middle	39.5	+15.3	+12.7	+7.0	+20.8	**+14.0**
Upper	45.2	+5.4	+11.5	+18.0	+6.1	**+10.3**
Lower–Upper Gap	–	+11.2	+7.4	+0.2	+22.3	**+10.3**

☆ Observation 3: Restoration benefits are smaller under greater representation drift. Table 1(b) shows that evidence restoration is most effective for tokens with limited representation drift: the average effect falls from +20.5 pp in the lower-drift tertile to +10.3 pp in the upper. A within-cell continuous analysis shows the same negative association.

Design implication. Together, these diagnostics motivate CIRA's objective: make evidence inaccessible after compression while preserving its utility under full-token inference.

4 METHODOLOGY

4.1 ATTACK FORMULATION

We consider an untargeted, image-specific attack with white-box access only to the vision encoder. The downstream question, reference answer, language model, compressor, and exact deployed compression budget are unavailable during optimization; the attacker knows only the admissible range $\mathcal{K} = \{K \in \mathbb{N} : K_{\min} \le K \le K_{\max}\}$.

For an evaluation tuple (x, q, y), the ideal compression-specific outcome is to preserve full-token correctness while causing compressed inference to fail:

$$\max_{\|\delta\|_\infty \le \epsilon} \min_{K \in \mathcal{K}} \mathcal{L}\left(f_\theta^{\mathcal{C},K}(x+\delta, q), y\right) \quad \text{s.t.} \quad c(x+\delta) = 1, \tag{3}$$

where $c(\cdot)$ denotes full-token correctness and $\mathcal{L}$ increases with compressed prediction error. Equation 3 defines the desired compression-specific behavior but is not optimized directly. CIRA instead replaces it with compressor-independent encoder-side objectives and evaluates cross-configuration transfer empirically.

4.2 GLOBAL SELECTION HIJACKING

Let $s_i(x)$ denote the encoder-side proxy priority score of visual token i, where larger values indicate higher retention priority. We collect these scores in the priority-score vector $\mathbf{s}(x) = [s_1(x), \ldots, s_N(x)] \in \mathbb{R}^N$, whose descending order defines the priority ranking over tokens. We write $\mathbf{s}^c = \mathbf{s}(x)$ and $\mathbf{s}^a = \mathbf{s}(x+\delta)$ for the clean and adversarial priority-score vectors, respectively.

We first construct the clean rank, its inverse-priority encoding, and a stabilized standardization operator:

$$\mathbf{r}^c = \mathrm{rank}_\downarrow(\mathbf{s}^c), \qquad \mathcal{R}(\mathbf{s}^c)_i = \frac{r_i^c - 1}{N-1}, \qquad \mathcal{Z}_{\epsilon_s}(\mathbf{u}) = \frac{\mathbf{u} - \bar{u}\mathbf{1}}{\max\left\{N^{-1/2}\|\mathbf{u} - \bar{u}\mathbf{1}\|_2, \epsilon_s\right\}}. \tag{4}$$

Here, $\mathbf{r}^c$ ranks tokens from high to low clean priority, while $\mathcal{R}$ maps high-priority tokens near zero and low-priority tokens near one. $\mathcal{Z}_{\epsilon_s}$ standardizes its input vector and prevents a degenerate denominator when its variance vanishes.

To drive a global priority inversion, CIRA maximizes the standardized alignment between adversarial priorities and the inverse clean priority ranking:

$$\mathcal{L}_{\mathrm{GSH}}(\delta) = \mathrm{Align}_{\epsilon_s}(\mathbf{s}^a, \mathcal{R}(\mathbf{s}^c)) \equiv \frac{1}{N}\left\langle \mathcal{Z}_{\epsilon_s}(\mathbf{s}^a), \mathcal{Z}_{\epsilon_s}(\mathcal{R}(\mathbf{s}^c))\right\rangle. \tag{5}$$

This differentiable, scale-invariant objective encourages clean high-priority tokens to move downward while promoting clean low-priority tokens.

4.3 HIDDEN-EVIDENCE PRESERVATION

Priority reallocation may also perturb the representations of displaced evidence, undermining full-token correctness. HEP therefore focuses preservation on clean high-priority tokens that cross candidate retention boundaries. CIRA computes the adversarial descending ranks $\mathbf{r}^a = \mathrm{rank}_\downarrow(\mathbf{s}^a)$ and defines $\mathcal{H}_K(\delta) = \{i : r_i^c \le K < r_i^a\}$, the clean proxy Top-$K$ tokens whose adversarial ranks move beyond the retention boundary at compression budget K. Under an unknown compression budget, their displacement weights are

$$w_i = \Pr_{K \sim \mathrm{Unif}(\mathcal{K})}[i \in \mathcal{H}_K(\delta)] = \frac{[\min(K_{\max}, r_i^a - 1) - \max(K_{\min}, r_i^c) + 1]_+}{K_{\max} - K_{\min} + 1}. \tag{6}$$

Here, $[z]_+ = \max(z, 0)$, and w_i is the fraction of admissible budgets under which token i becomes hidden.

To limit representation drift in these tokens, let $\mathbf{h}_i^c$ and $\mathbf{h}_i^a$ denote their clean and adversarial visual-token features. We normalize their directions and measure the resulting representation drift by

$$\widetilde{\mathbf{h}}_i^c = \frac{\mathbf{h}_i^c}{\|\mathbf{h}_i^c\|_2}, \qquad \widetilde{\mathbf{h}}_i^a = \frac{\mathbf{h}_i^a}{\|\mathbf{h}_i^a\|_2}, \qquad d_i = \frac{1}{2}\left(1 - (\widetilde{\mathbf{h}}_i^c)^\top \widetilde{\mathbf{h}}_i^a\right), \tag{7}$$

where the factor $1/2$ normalizes cosine distance to $[0, 1]$.

Hidden-Evidence Preservation aggregates these token-level distances using the budget-marginal weights:

$$\mathcal{L}_{\mathrm{HEP}}(\delta) = -\sum_{i=1}^{N} \widetilde{w}_i d_i, \qquad \widetilde{w}_i = \frac{\mathrm{sg}(w_i)}{\max\left\{\sum_{j=1}^{N} \mathrm{sg}(w_j), \epsilon_h\right\}}. \tag{8}$$

The stop-gradient freezes rank-derived weights within an update, while ϵ_h stabilizes normalization; weights are recomputed at the next update so evidence displaced across more of $\mathcal{K}$ receives greater protection.

4.4 Joint Optimization

CIRA combines the two objectives as

$$\max_{\delta} \quad \mathcal{L}_{\mathrm{CIRA}}(\delta) = \mathcal{L}_{\mathrm{GSH}}(\delta) + \lambda \cdot \mathcal{L}_{\mathrm{HEP}}(\delta), \qquad \text{s.t.} \quad \|\delta\|_\infty \le \epsilon. \tag{9}$$

Here, λ controls the preservation strength. We maximize equation 9 using projected sign-gradient ascent over the valid-image domain. The complete optimization procedure is given in **Appendix A**.

5 Experiments

5.1 Experimental Setup

Models and Benchmarks. We evaluate LLaVA-v1.5-7B (Liu et al., 2024b) on 1,000 randomly sampled image–question pairs from each of POPE (Li et al., 2023b), TextVQA (Singh et al., 2019), and MME (Fu et al., 2025), covering object hallucination, scene-text understanding, and general visual perception and reasoning, respectively. We additionally evaluate Qwen3-VL-8B-Instruct (Bai et al., 2025) and InternVL3.5-8B (Wang et al., 2025b).

Compression Settings. We evaluate VisionZip (Yang et al., 2025b), VisPruner (Zhang et al., 2025b), PruMerge (Shang et al., 2025), and FastV (Chen et al., 2024) at $\mathcal{K}_{\mathrm{eval}} = \{32, 64, 128, 192\}$. For each input, one adversarial image is reused across all compressor–budget settings.

Baselines. We compare CIRA with the downstream-agnostic VEAttack (Mei et al., 2026b) and CAGE (Zhang et al., 2026c); CAA (Zhang et al., 2026a) is reported as a stronger-access reference.

Evaluation Metrics. On the clean-eligible set $\mathcal{S}_K$ defined in Eq. 1, we report the Compression-Specific Failure Rate (CSFR) and full-token attack success rate:

$$\mathrm{CSFR}_K = \frac{1}{|\mathcal{S}_K|} \sum_{i \in \mathcal{S}_K} c_i(x_i^{\mathrm{adv}}) \left[1 - c_{i,K}(x_i^{\mathrm{adv}})\right], \qquad \mathrm{ASR} = \frac{1}{|\mathcal{S}_{\mathrm{full}}|} \sum_{i \in \mathcal{S}_{\mathrm{full}}} \left[1 - c_i(x_i^{\mathrm{adv}})\right]. \tag{10}$$

Here, $\mathcal{S}_{\mathrm{full}} = \{i : c_i(x_i) = 1\}$, and Avg. CSFR is the arithmetic mean of CSFR_K over $K \in \mathcal{K}_{\mathrm{eval}}$ for a fixed dataset–compressor setting. Broader summaries weight each reported dataset–compressor setting equally. CSFR counts post-attack failures confined to the compressed path, normalized over inputs answered correctly by both clean inference paths; ASR is normalized over inputs answered correctly by clean full-token inference. Clean and post-attack accuracy results are reported in **Appendix G**.

Implementation Settings. We run all experiments on a single NVIDIA GeForce RTX 4090 GPU. All downstream-agnostic attacks use $\epsilon = 4/255$ and 100 optimization steps; CIRA uses projected sign-gradient ascent with $\lambda = 0.8$. Further protocol details and sensitivity analyses appear in **Appendix B** and **Appendix D**.

Table 2: CSFR and Full ASR on LLaVA-v1.5-7B. For downstream-agnostic attacks, **red** and **blue** cells mark highest CSFR and lowest Full ASR, and underlining marks second best; CAA† uses the downstream question and language model during optimization and is excluded from the markings.

Metric	POPE				TextVQA				MME			
	VEAttack	CAGE	CAA†	CIRA	VEAttack	CAGE	CAA†	CIRA	VEAttack	CAGE	CAA†	CIRA
Full-token Inference (Compressor-independent)												
Full ASR ↓	38.65	39.48	2.72	**4.96**	60.95	71.35	3.82	**10.61**	38.26	39.77	3.41	**5.18**
VisionZip												
CSFR@192 ↑	4.05	3.19	1.72	**7.61**	4.87	2.92	2.14	**11.35**	2.83	4.04	1.62	**7.40**
CSFR@128 ↑	4.98	5.72	2.87	**14.05**	5.32	5.32	2.05	**18.21**	3.33	5.55	1.80	**10.26**
CSFR@64 ↑	5.05	9.71	3.46	**22.47**	6.28	6.28	2.39	**27.41**	3.64	5.97	2.92	**19.07**
CSFR@32 ↑	6.52	9.48	4.57	**26.96**	9.76	5.24	5.01	**31.70**	4.56	5.19	2.86	**22.17**
Avg. CSFR ↑	5.15	7.03	3.16	**17.77**	6.56	4.94	2.90	**22.17**	3.59	5.19	2.30	**14.73**
VisPruner												
CSFR@192 ↑	3.42	3.79	1.23	**5.87**	2.92	3.50	1.75	**10.50**	2.56	3.50	1.22	**7.14**
CSFR@128 ↑	4.60	4.35	2.97	**9.07**	3.80	3.40	2.18	**13.18**	2.88	4.79	2.33	**11.23**
CSFR@64 ↑	6.27	7.45	4.68	**15.29**	5.89	5.47	4.08	**19.13**	3.33	5.36	4.47	**18.70**
CSFR@32 ↑	7.59	10.46	8.29	**22.21**	8.07	7.33	4.63	**25.57**	4.13	5.66	6.73	**21.87**
Avg. CSFR ↑	5.47	6.51	4.29	**13.11**	5.17	4.93	3.16	**17.09**	3.22	4.83	3.69	**14.73**
PruMerge												
CSFR@192 ↑	5.33	3.65	2.66	**22.86**	6.02	4.86	3.44	**28.66**	2.64	4.69	1.75	**20.64**
CSFR@128 ↑	5.39	5.54	2.92	**29.59**	6.51	5.35	4.86	**36.23**	3.62	5.28	3.02	**22.93**
CSFR@64 ↑	6.96	9.38	3.63	**32.53**	8.29	6.64	4.49	**43.91**	3.50	5.78	3.04	**29.98**
CSFR@32 ↑	9.40	11.76	3.44	**32.29**	9.18	6.70	5.20	**48.68**	4.06	5.78	2.81	**35.16**
Avg. CSFR ↑	6.77	7.58	3.16	**29.32**	7.50	5.89	4.50	**39.37**	3.45	5.38	2.66	**27.18**
FastV												
CSFR@192 ↑	4.61	4.74	2.30	**7.04**	4.29	4.09	1.43	**9.61**	2.77	2.90	1.38	**5.81**
CSFR@128 ↑	6.85	6.59	2.90	**11.46**	7.10	5.38	2.15	**13.55**	3.17	3.89	1.44	**7.64**
CSFR@64 ↑	9.90	10.47	3.44	**22.24**	9.11	7.67	4.08	**23.74**	3.65	5.18	2.89	**15.07**
CSFR@32 ↑	12.32	10.40	4.48	**25.44**	7.80	8.38	3.76	**32.95**	4.24	5.55	6.20	**20.39**
Avg. CSFR ↑	8.42	8.05	3.28	**16.55**	7.08	6.38	2.85	**19.96**	3.46	4.38	2.98	**12.23**

5.2 Main Results

In this section, we evaluate CIRA's compression selectivity and cross-configuration transfer, and compare it with stronger-access attacks.

Compression-specific selectivity. Across the four budgets and 12 dataset–compressor settings in Table 2, CIRA achieves a mean CSFR of 20.35%, compared with 5.49% for VEAttack and 5.92% for CAGE. Its Full ASR is 6.92%, substantially below 45.95% and 50.20%, respectively. CIRA therefore induces more compression-specific failures while causing much less full-token degradation than either downstream-agnostic baseline.

Cross-configuration transfer. The same adversarial image is reused without re-optimization across selection-, pruning-, and merging-based compression rules and all evaluated budgets. CIRA remains effective across these configurations, reaching 31.96% mean CSFR on PruMerge compared with 6.28% for CAGE. Averaged equally over all 12 dataset–compressor settings, its CSFR increases from 12.04% at $K = 192$ to 28.78% at $K = 32$. Compression-selective behavior also extends to Qwen3-VL and InternVL (Table 10); qualitative cross-setting examples appear in **Appendix I**.

Comparison with stronger-access CAA. CAA† optimizes with access to the downstream question and language model, whereas CIRA uses only the vision encoder. Despite its lower Full ASR (3.32% for CAA† versus 6.92% for CIRA), CAA† achieves only 3.24% mean CSFR, well below CIRA's 20.35%. The contrast shows that preserving the full-token prediction is not sufficient to induce compression-specific failures; CIRA's priority-reallocation objective addresses this distinct requirement under narrower access.

➪ **Takeaway.** CIRA induces compression-specific failures across heterogeneous compression rules and budgets while limiting full-token degradation.

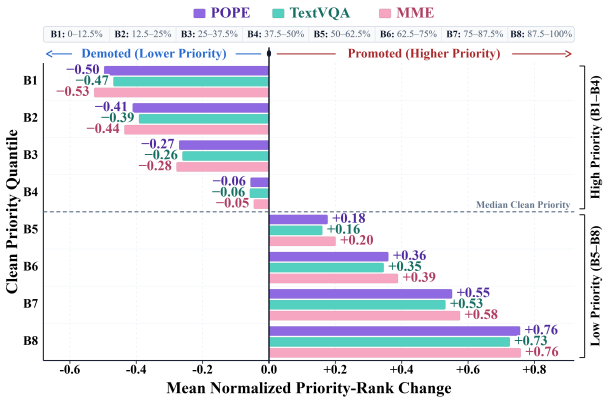


Figure 3: CIRA-induced priority reallocation across the clean visual-token ranking at $K = 128$.

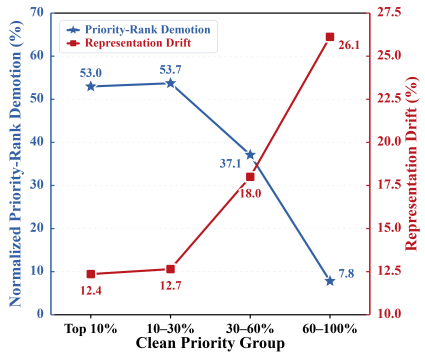


Figure 4: Priority reallocation and representation drift across token groups.

5.3 Mechanistic Analysis

To connect CIRA's behavior to its objectives, we test whether it globally reallocates token priority while limiting representation drift in displaced evidence.

Priority reallocation. Figure 3 shows a near-monotonic priority reallocation across all three benchmarks: the highest-priority octile is demoted by 0.47–0.53 normalized-rank units, whereas the lowest-priority octile is promoted by 0.73–0.76, with the sign changing near the median. This cross-quantile pattern is not confined to a single retention boundary.

Priority reallocation versus representation drift. Figure 4 shows that rank displacement and representation drift behave differently across clean-priority groups. Rank demotion peaks at 53.0% and 53.7% in the top 10% and 10–30% groups, where representation drift is lowest at 12.4% and 12.7%. The lowest-priority 60–100% group shows the reverse pattern, with 7.8% rank demotion and 26.1% drift. This separation indicates that CIRA reallocates clean high-priority tokens while limiting their representation drift, consistent with the HEP objective.

Component ablation. On VisionZip, Table 3 isolates the two objectives. Removing GSH reduces CSFR averaged across the three benchmarks and four budgets from 18.22% to 3.68%. Removing HEP leaves this mean nearly unchanged (17.04%) but raises mean Full ASR from 6.92% to 23.10%; replacing it with global representation preservation yields 15.71% mean CSFR and 8.66% Full ASR. Global preservation yields lower CSFR and higher Full ASR than HEP, supporting preservation focused on displaced evidence rather than a uniform constraint.

➪ **Takeaway.** GSH drives compression-specific failure induction, while HEP limits representation drift in displaced evidence and substantially reduces full-token degradation.

6 Selection Stabilization Defense

To test whether stabilizing token selection can suppress CIRA, TCS exploits cross-view priority stability: clean high-priority evidence tends to remain stable under small translations, whereas attack-induced replacements are more view-sensitive. At the LLaVA–VisionZip token-priority interface, TCS uses four translated views $\mathcal{V} = \{T_{0,0}, T_{d,0}, T_{0,d}, T_{d,d}\}$ with $d = 7$ pixels. For view v, $\mathbf{s}^{(v)}$ is its priority-score vector, $\mathcal{A}_v$ maps the translated score grid back to the reference coordinates, and rank 1 denotes the highest priority. TCS selects the K tokens with highest aligned rank-quantile consensus:

$$q_i^{(v)} = 1 - \frac{r_i\big(\mathcal{A}_v(\mathbf{s}^{(v)})\big) - 1}{N - 1}, \quad \bar{q}_i = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} q_i^{(v)}, \quad \mathcal{I}_K^{\mathrm{TCS}} = \mathrm{TopK}(\bar{\mathbf{q}}, K). \tag{11}$$

Table 3: CIRA ablation on VisionZip; arrows show changes from full CIRA.

Metric	Full CIRA	Selection w/o Sel.	Preservation w/o Pres.	Preservation Global Pres.
POPE				
Full ASR ↓	**4.96**	2.01↓2.95	15.72↑10.76	4.73↓0.23
CSFR@192 ↑	**7.61**	1.60↓6.01	9.94↑2.33	4.79↓2.82
CSFR@128 ↑	**14.05**	1.99↓12.06	15.42↑1.37	9.95↓4.10
CSFR@64 ↑	**22.47**	4.39↓18.08	22.07↓0.40	19.95↓2.52
CSFR@32 ↑	**26.96**	5.63↓21.33	26.81↓0.15	26.37↓0.59
Avg. CSFR ↑	**17.77**	3.40↓14.37	18.56↑0.79	15.26↓2.51
TextVQA				
Full ASR ↓	**10.61**	6.91↓3.70	36.00↑25.39	15.82↑5.20
CSFR@192 ↑	**11.35**	3.90↓7.45	9.94↓1.41	8.58↓2.77
CSFR@128 ↑	**18.21**	2.05↓16.16	15.20↓3.02	12.53↓5.69
CSFR@64 ↑	**27.41**	5.43↓21.98	23.70↓3.72	23.04↓4.37
CSFR@32 ↑	**31.70**	6.44↓25.26	26.73↓4.97	36.28↑4.57
Avg. CSFR ↑	**22.17**	4.46↓17.71	18.89↓3.28	20.11↓2.06
MME				
Full ASR ↓	**5.18**	2.40↓2.77	17.57↑12.40	5.44↑0.26
CSFR@192 ↑	**7.40**	1.62↓5.79	7.55↑0.14	5.80↓1.61
CSFR@128 ↑	**10.26**	1.93↓8.33	11.33↑1.06	7.73↓2.53
CSFR@64 ↑	**19.07**	4.09↓14.98	17.23↓1.84	14.45↓4.62
CSFR@32 ↑	**22.17**	5.08↓17.09	18.57↓3.60	19.05↓3.12
Avg. CSFR ↑	**14.73**	3.18↓11.55	13.67↓1.06	11.76↓2.97

Table 4: TCS evaluation on VisionZip; arrows denote changes from matched no-defense baselines.

Metric	CIRA None	CIRA TCS	CAGE None	CAGE TCS	Adaptive CIRA None	Adaptive CIRA TCS
POPE						
Full ASR	4.96		39.48		3.78	
CSFR@192 ↓	7.61	2.60↓5.01	3.19	4.34↑1.15	6.99	6.57↓0.42
CSFR@128 ↓	14.05	3.40↓10.65	5.72	4.79↓0.93	10.07	10.71↑0.64
CSFR@64 ↓	22.47	2.84↓19.63	9.71	5.68↓4.03	20.08	13.65↓6.43
CSFR@32 ↓	26.96	4.21↓22.75	9.48	5.52↓3.96	28.00	21.92↓6.08
Avg. CSFR ↓	17.77	3.26↓14.51	7.03	5.08↓1.95	16.29	13.21↓3.08
TextVQA						
Full ASR	10.61		71.35		11.45	
CSFR@192 ↓	11.35	2.39↓8.96	2.92	1.59↓1.33	8.38	8.17↓0.21
CSFR@128 ↓	18.21	2.07↓16.14	5.32	2.48↓2.84	11.09	9.30↓1.79
CSFR@64 ↓	27.41	4.50↓22.91	6.28	3.43↓2.85	26.09	14.13↓11.95
CSFR@32 ↓	31.70	3.09↓28.61	5.24	4.99↓0.25	38.90	25.65↓13.25
Avg. CSFR ↓	22.17	3.01↓19.16	4.94	3.12↓1.82	21.11	14.31↓6.80
MME						
Full ASR	5.18		39.77		6.57	
CSFR@192 ↓	7.40	2.72↓4.68	4.04	2.31↓1.73	3.64	5.03↑1.39
CSFR@128 ↓	10.26	2.64↓7.62	5.55	2.50↓3.05	9.53	7.22↓2.31
CSFR@64 ↓	19.07	4.53↓14.54	5.97	4.09↓1.88	18.54	11.68↓6.86
CSFR@32 ↓	22.17	5.75↓16.42	5.19	4.67↓0.52	23.17	18.35↓4.82
Avg. CSFR ↓	14.73	3.91↓10.82	5.19	3.39↓1.80	13.72	10.57↓3.15

Rank quantiles make priorities comparable across views despite differences in score scale. The selected indices are applied to the unshifted-view features, leaving aggregation and full-token inference unchanged. Construction and mechanism details appear in **Appendix F**; complementary utility results are reported in **Appendix G**.

Defense effectiveness. Within the matched evaluation blocks in Table 4, TCS reduces standard CIRA CSFR averaged across the three datasets and four budgets from 18.22% to 3.39%, an 81.4% relative reduction, compared with a 32.5% reduction for CAGE. Suppression strengthens as the compression budget decreases, with the largest reductions under tighter compression.

Adaptive stress test. Because TCS is deterministic and public, we also evaluate an adaptive attacker that optimizes the CIRA objectives over the same four views while sharing one image-space perturbation:

$$\mathcal{L}_{\text{adapt}}(\delta) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left[\mathcal{L}_{\text{GSH}}^{(v)}(\delta) + \lambda \mathcal{L}_{\text{HEP}}^{(v)}(\delta) \right]. \tag{12}$$

Adaptive CIRA uses the same access assumptions and perturbation budget as standard CIRA. Under TCS, its mean CSFR rises from 3.39% to 12.70%, reaching 74.5% of Adaptive CIRA's matched undefended value (17.04%). TCS therefore retains a smaller but nonzero effect against the adaptive attack. Additional optimization and mechanism diagnostics appear in **Appendix F.3**.

7 Conclusion

Visual-token compression changes not only inference cost but also which visual evidence remains available after compression. By pairing full-token and compressed inference on the same adversarial input, we attribute failures specifically to the compression path. CIRA induces such failures across unknown compressor and budget settings by reallocating token priorities while limiting full-token degradation. Controlled diagnostics further show that retained-set allocation affects compressed correctness and that recovery is negatively associated with representation drift in displaced evidence. Selection stabilization substantially suppresses CIRA, although adaptive optimization partially restores its effectiveness. Together, these results support treating the compression boundary as a security-relevant component of LVLM deployment and motivate paired robustness evaluation for compressed LVLMs.

AI USE STATEMENT

Generative AI tools were used in a limited supporting role for this work, including language editing and polishing, literature retrieval and discovery, research ideation and technical execution support, and drafting and revising parts of the manuscript. In particular, these tools were used to improve the clarity and fluency of writing, help identify related literature, provide technical suggestions for coding and experimental workflows, and assist in refining sections of the paper during revision. All AI-assisted content was independently checked, validated, and, where necessary, modified by the authors. The authors retain full responsibility for the scientific content, methodological choices, experimental evidence, interpretations, and final form of the manuscript.

REPRODUCIBILITY STATEMENT

We provide comprehensive details to facilitate the reproduction and verification of CIRA. The threat model, compression-specific failure formulation, and attack objectives are described in Sections 3 and 4, including the Global Selection Hijacking (GSH) and Hidden-Evidence Preservation (HEP) objectives and their joint optimization. The optimization procedure and experimental details are provided in Appendices A and B, respectively. Main results, mechanistic analyses, and defense evaluations are reported in Sections 5.2, 5.3, and 6. Additional analyses, including retained-set diagnostics, sensitivity studies, cross-model evaluation, task-utility results, limitations, and qualitative cases, are provided in Appendices C, D, E, G, H, and I. Code and scripts for reproducing the reported experiments are provided in the anonymous repository.

ETHICS STATEMENT

This work investigates the adversarial robustness of LVLMs under visual-token compression in a controlled research setting. Our experiments use publicly available benchmark datasets (e.g., POPE, TextVQA, and MME) and publicly released pretrained models (e.g., LLaVA-v1.5-7B, Qwen3-VL-8B-Instruct, and InternVL3.5-8B), involving no human subjects or personally identifiable information. We recognize the dual-use nature of adversarial robustness research. CIRA is developed to identify compression-specific vulnerabilities and support robustness evaluation and defense development. The proposed attack is evaluated under a restricted threat model with bounded image-space perturbations and vision-encoder-only access. We transparently report the attack assumptions and evaluation protocols to facilitate reproducibility and responsible security research.

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.

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*, 2025.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. Compress & cache: Vision token compression for efficient generation and retrieval. *Advances in Neural Information Processing Systems*, 38:31943–31968, 2026.

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.

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

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.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835. IEEE, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

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

Shishen Gu, Jiequan Cui, Wenbo Hu, Zenglin Shi, Zhenzhen Hu, and Richang Hong. Visual token compression enhances robustness of mllms. *arXiv preprint arXiv:2607.22716*, 2026.

Kai Hu, Weichen Yu, Li Zhang, Alexander Robey, Andy Zou, Chengming Xu, Haoqi Hu, and Matt Fredrikson. Transferable adversarial attacks on black-box vision-language models. *arXiv preprint arXiv:2505.01050*, 2025.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong Yu, Laizhong Cui, and Qi Tian. N\" uwa: Mending the spatial integrity torn by vlm token pruning. *arXiv preprint arXiv:2602.02951*, 2026.

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.

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.

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

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.

Zhenyu Li, Zuchao Li, Ping Wang, Lefei Zhang, and Haojun Ai. Vista-llm: Decoupled query-guided visual token pruning for efficient long-video large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13171–13187, 2026b.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's box: Towards building universal attackers against real-world large vision-language models. *Advances in Neural Information Processing Systems*, 37:52127–52158, 2024a.

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, 2024b.

Jiaxiang Liu, Jiawei Du, Xiao Liu, Prayag Tiwari, and Mingkun Xu. Self-calibrated consistency can fight back for adversarial robustness in vision-language models. *arXiv preprint arXiv:2510.22785*, 2025.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 102–111. IEEE, 2023.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv preprint arXiv:2403.09766*, 2024.

Jianxin Ma, Shibo Jin, and Lujuan Dang. Visual token compression via run-length pruning in multimodal large language models. *Pattern Recognition Letters*, 2026.

Hefei Mei, Zirui Wang, Chang Xu, Jianyuan Guo, and Minjing Dong. Pa-attack: Guiding gray-box attacks on lvlm vision encoders with prototypes and attention. *arXiv preprint arXiv:2602.19418*, 2026a.

Hefei Mei, Zirui Wang, Shen You, Minjing Dong, and Chang Xu. Veattack: Downstream-agnostic vision encoder attack against large vision language models. In *International Conference on Learning Representations*, volume 2026, pp. 18135–18161, 2026b.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21527–21536, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3679–3687. IEEE, 2023.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

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

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International conference on learning representations*, volume 2024, pp. 30853–30885, 2024.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29958–29967, 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Guohao Sun, Yufei Wang, Sizhuo Ma, Yuege Xie, Yuting Cheng, Zhiqiang Tao, and Jian Wang. If-prune: Information-flow guided token pruning for efficient vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3522–3531, 2026.

Yizheng Sun, Hao Li, Chang Xu, Hongpeng Zhou, Chenghua Lin, Riza Theresa Batista-Navarro, and Jingyuan Sun. Does acceleration cause hidden instability in vision language models? uncovering instance-level divergence through a large-scale empirical study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8453–8467, 2025.

Rinyoichi Takezoe, Yaqian Li, Zi-Hao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. Learnpruner: Rethinking attention-based token pruning in vision language models. In *International Conference on Learning Representations*, volume 2026, pp. 66381–66400, 2026.

Han Wang, Yuxiang Nie, Yongjie Ye, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. Dynamic-vlm: Simple dynamic visual token compression for videollm. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20812–20823. IEEE, 2025a.

Shuailong Wang, Xinyu Lyu, Shengming Yuan, Jingkuan Song, Heng Tao Shen, and Lianli Gao. Understanding and mitigating token-pruning-induced vulnerabilities in VLMs. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=D3OHVbePvz.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Yahong Wang, Juncheng Wu, Zhangkai Ni, Longzhen Yang, Yihang Liu, Chengmei Yang, Ying Wen, Lianghua He, Xianfeng Tang, Hui Liu, et al. When token pruning is worse than random: Understanding visual token information in vllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31910–31919, 2026b.

Yi Wang, Haofei Zhang, Qihan Huang, Anda Cao, Gongfan Fang, Wei Wang, Xuan Jin, Jie Song, Mingli Song, and Xinchao Wang. Rethinking token reduction for large vision-language models. *arXiv preprint arXiv:2603.21701*, 2026c.

Zeyu Wang, Cihang Xie, Brian Bartoldson, and Bhavya Kailkhura. Double visual defense: Adversarial pre-training and instruction tuning for improving vision-language model robustness. *arXiv preprint arXiv:2501.09446*, 2025c.

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14679–14689. IEEE, 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.

Tong Xu, Hailong Shi, and Xingyu Gao. Score: Salience-coverage reduction for vision token pruning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24686–24695, 2026.

Cheng-Yu Yang, Shao-Yuan Lo, and Yu-Lun Liu. Reroute, don't remove: Recoverable visual token routing for vision-language models. *arXiv preprint arXiv:2606.12412*, 2026.

Longrong Yang, Dong Shen, Chaoxiang Cai, Kaibing Chen, Fan Yang, Tingting Gao, Di Zhang, and Xi Li. Libra-merging: Importance-redundancy and pruning-merging trade-off for acceleration plug-in in large vision-language model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9402–9412. IEEE, 2025a.

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

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.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956, 2023.

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. Visiontrim: Unified vision token compression for training-free mllm acceleration. *arXiv preprint arXiv:2601.22674*, 2026.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 5005–5013, 2022.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19900–19909. IEEE, 2025a.

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

Xiaomei Zhang, Zhaoxi Zhang, Leo Yu Zhang, Yanjun Zhang, Guanhong Tao, and Shirui Pan. Less is more–until it breaks: Security pitfalls of vision token compression in large vision-language models. *arXiv preprint arXiv:2601.12042*, 2026a.

Xinwei Zhang, Li Bai, Tianwei Zhang, Youqian Zhang, Qingqing Ye, Yingnan Zhao, Ruochen Du, and Haibo Hu. Grounding-driven attack: Improving encoder-based adversarial transferability against large vision-language models. *arXiv preprint arXiv:2602.09431*, 2026b.

Xinwei Zhang, Hangcheng Liu, Li Bai, Hao Wang, Qingqing Ye, Tianwei Zhang, and Haibo Hu. On the adversarial robustness of large vision-language models under visual token compression. *arXiv preprint arXiv:2601.21531*, 2026c.

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

APPENDIX CONTENTS

Appendix A. CIRA Optimization Procedure 15
Appendix B. Detailed Experimental Setup 15
Appendix C. Controlled Retained-Set Allocation Diagnostics 17
Appendix D. Sensitivity Analyses 20
Appendix E. Cross-Model Scope 21
Appendix F. Selection Stabilization Defense 23
Appendix G. Complementary Task-Utility Results 24
Appendix H. Limitation Discussion and Future Work 26
Appendix I. Qualitative Case Studies 27

A CIRA OPTIMIZATION PROCEDURE

Algorithm 1 gives the image-space optimization induced by CIRA's two encoder-side objectives. Clean features and priority scores are cached once; adversarial scores, ranks, and displacement weights are refreshed at every step. The priority rule $\mathcal{P}$ maps encoder attention to the token-score vector used by GSH.

Algorithm 1 Projected sign-gradient optimization of CIRA.

Require: Image x; encoder $\mathcal{E}$; priority rule $\mathcal{P}$
Require: Candidate interval $[K_{\min}, K_{\max}]$; $\epsilon, \epsilon_s, \epsilon_h$
Require: Step size α; steps T; preservation weight λ
Ensure: Adversarial image x^{adv}

$\mathcal{K} \leftarrow \{K_{\min}, \ldots, K_{\max}\}$
$(\mathbf{H}^c, \mathbf{s}^c) \leftarrow (\mathcal{E}(x), \mathcal{P}(x))$
$\mathbf{r}^c \leftarrow \mathrm{Rank}_{\downarrow}(\mathbf{s}^c)$
$\mathbf{u}^c \leftarrow \mathcal{R}(\mathbf{s}^c)$
Initialize $\delta^{(0)} \leftarrow \mathbf{0}$
for $\tau = 0, \ldots, T-1$ **do**
  $x^{(\tau)} \leftarrow x + \delta^{(\tau)}$
  $(\mathbf{H}^a, \mathbf{s}^a) \leftarrow (\mathcal{E}(x^{(\tau)}), \mathcal{P}(x^{(\tau)}))$
  $\mathcal{L}_{\mathrm{GSH}} \leftarrow \mathrm{Align}_{\epsilon_s}(\mathbf{s}^a, \mathbf{u}^c)$
  $\mathbf{r}^a \leftarrow \mathrm{Rank}_{\downarrow}(\mathbf{s}^a)$
  $w_i \leftarrow \mathrm{sg}\big(\Pr_{K \sim \mathrm{Unif}(\mathcal{K})}[r_i^c \le K < r_i^a]\big)$
  $d_i \leftarrow \frac{1}{2}(1 - \cos(\mathbf{h}_i^c, \mathbf{h}_i^a))$
  $\mathcal{L}_{\mathrm{HEP}} \leftarrow -\frac{\sum_i w_i d_i}{\max(\sum_i w_i, \epsilon_h)}$
  $\mathcal{L}_{\mathrm{CIRA}} \leftarrow \mathcal{L}_{\mathrm{GSH}} + \lambda \mathcal{L}_{\mathrm{HEP}}$
  $g \leftarrow \mathrm{sign}(\nabla_\delta \mathcal{L}_{\mathrm{CIRA}})$
  $\delta^{(\tau+1)} \leftarrow \Pi_{\Delta_\epsilon(x)}(\delta^{(\tau)} + \alpha g)$
end for
return $x^{\mathrm{adv}} \leftarrow x + \delta^{(T)}$

Sorting remains outside of the gradient path, $\mathrm{sg}(\cdot)$ blocks gradients through the discrete weights, and $\Delta_\epsilon(x) = \{\delta : \|\delta\|_\infty \le \epsilon, x + \delta \in [0,1]^d\}$ is the feasible perturbation set. The update therefore uses only the vision encoder, priority-scoring rule, and candidate compression-budget interval.

B DETAILED EXPERIMENTAL SETUP

B.1 MODELS

LLaVA-v1.5-7B. Our primary model is LLaVA-v1.5-7B (Liu et al., 2024b) with a CLIP ViT-L/14-336 vision encoder (Radford et al., 2021). The released model feeds 576 penultimate-layer patch tokens to its multimodal projector. We use LLaVA for the full benchmark–compressor matrix and for the mechanism, ablation, sensitivity, and defense studies.

Additional LVLM families. We additionally evaluate Qwen3-VL-8B-Instruct (Bai et al., 2025) and InternVL3.5-8B (Wang et al., 2025b). Qwen3-VL uses its native dynamic resolution and spatial merger, so its visual-token count varies by image. InternVL uses one 448×448 tile, producing 256 tokens after spatial downsampling. Accordingly, we use family-specific evaluation budgets of $\{96, 64, 32\}$ for Qwen3-VL and $\{128, 64, 32\}$ for InternVL. CIRA is optimized separately for each model using its native vision encoder.

B.2 Benchmarks

We use fixed, randomly sampled 1,000-pair subsets from POPE (Li et al., 2023b), TextVQA (Singh et al., 2019), and MME (Fu et al., 2025). POPE evaluates object hallucination through balanced object-presence questions; TextVQA tests reading of text in natural images and provides ten reference answers per question; MME covers perception and cognition with binary questions. Within each model–benchmark setting, all attacks are evaluated on the same image–question pairs using a common prompt, decoding protocol, and answer evaluator. POPE and MME outputs are lowercased, stripped of punctuation, and matched by the first normalized yes/no token. TextVQA uses its EvalAI-style normalizer and accepts a match to any reference answer.

These task-specific rules provide the per-example binary correctness required by the paired CSF definition. For TextVQA, we use a match to any normalized reference rather than the benchmark-level soft agreement score; MME is evaluated per question rather than by its aggregate category score. The same evaluators are used for clean, full-token, and compressed inference.

B.3 Compression Mechanisms

VisionZip. VisionZip (Yang et al., 2025b) retains high-attention tokens and merges the remainder around uniformly sampled contextual tokens. Compression budgets $K = 32, 64, 128, 192$ use (dominant, contextual) counts $(27, 5)$, $(54, 10)$, $(108, 20)$, and $(162, 30)$, respectively.

VisPruner. VisPruner (Zhang et al., 2025b) combines attention-based importance with feature diversity. Half of each budget is assigned to important tokens and the remainder to diverse tokens.

PruMerge. PruMerge (Shang et al., 2025) selects attention-ranked representatives and merges nearby tokens by feature similarity. We use $K-1$ representatives together with one attention-weighted residual aggregate, yielding exactly K output tokens.

FastV. FastV (Chen et al., 2024) ranks visual tokens using language-model attention and removes low-ranked tokens after layer 2.

These configurations are used for the primary LLaVA evaluation. For Qwen3-VL and InternVL, the corresponding compression rules are instantiated on their native visual-token interfaces while preserving each method's selection and aggregation principle. Throughout the evaluation, compression budget K denotes the number of post-compression visual tokens passed to subsequent computation, whether obtained through selection, merging, or both.

B.4 Implementation

Inference. All experiments run on one NVIDIA GeForce RTX 4090 GPU. Decoding is deterministic (`do_sample=False`) with at most 64 new tokens, using each model's native image processor and conversation template.

Priority-score instantiation. CIRA derives token priorities from model-native late-layer visual attention. Let $\mathcal{L}_{\text{score}}$ denote the encoder layers used to compute token priority scores. We use the final, third-to-last, and fifth-to-last encoder blocks, corresponding to relative layer indices $(-1, -3, -5)$. For the CLIP encoder in the primary setting, the score from Section 4.2 is

$$s_i(x) = \frac{1}{|\mathcal{L}_{\text{score}}|} \sum_{\ell \in \mathcal{L}_{\text{score}}} \sum_{h=1}^{H} \alpha_{0,i}^{\ell,h}(x), \qquad i = 1, \ldots, N, \tag{13}$$

where $\alpha_{0,i}^{\ell,h}$ is attention from the class token to patch i at head h and visual layer ℓ, and H is the number of attention heads.

InternVL derives token priorities from class-to-patch visual self-attention and aggregates them over its spatial-downsampling groups. For Qwen3-VL, which lacks a class-token routing interface, we use the mean incoming attention over visual queries and aggregate scores within its native spatial-merger groups. For both models, priorities are averaged over the same three relative encoder layers to produce one score per downstream visual token.

Preservation features. For LLaVA, HEP measures tokenwise cosine distance between final CLIP patch features after post-layer normalization. For Qwen3-VL, it averages the tokenwise distances of the merged main visual features and the DeepStack features. For InternVL, it uses the visual tokens returned by the model's feature-extraction module after pixel shuffle and MLP projection.

Optimization. All downstream-agnostic attacks use an ℓ_∞ perturbation budget of $4/255$ and 100 optimization steps. CIRA starts from the clean image and uses projected sign-gradient ascent with step size $1/255$, $\lambda = 0.8$, and candidate budget range $[K_{\min}, K_{\max}] = [32, 192]$. The remaining hyperparameters of VEAttack (Mei et al., 2026b) and CAGE (Zhang et al., 2026c) follow their released settings; CAGE uses its released budget interval $[16, 192]$. CAA† (Zhang et al., 2026a) optimizes a question-conditioned objective at language-model layer 2 with $\epsilon = 4/255$ and 100 steps. One adversarial image is generated per image–question pair and then evaluated across compressors and budgets without re-optimization.

Evaluation protocol. Each downstream-agnostic method generates one adversarial image per clean input. After optimization, the image is fixed and evaluated across all corresponding questions, compressors, and budgets. Clean eligibility is determined separately for each compressor–budget setting using $\mathcal{S}_K$ in equation 1, while Full ASR is computed over $\mathcal{S}_{\text{full}}$ defined in Section 5.1.

C Controlled Retained-Set Allocation Diagnostics

The diagnostic in Section 3 isolates retained-set allocation through counterfactual exchanges at fixed first-observed CSF states. Answer-aware scores are used only for post-hoc exchange selection after the failure state is fixed and are not part of CIRA optimization.

C.1 Diagnostic Cohort and Fixed Failure State

Diagnostic setting. We use LLaVA–VisionZip on POPE, TextVQA, and MME at $K \in \{32, 64, 128, 192\}$. VisionZip retains dominant patch tokens individually and aggregates the remainder into contextual tokens. We call the individually retained patches *direct tokens*; their capacity D_K is smaller than the total compressed budget K.

Common clean cohort. Let j index an image–question observation, comprising image x_j, its associated question, and reference answer; index i is reserved for visual tokens. We retain only observations that are correct under full-token inference and under every evaluated compressed setting:

$$c_j(x_j) = 1, \qquad c_{j,K}(x_j) = 1, \quad \forall K \in \{192, 128, 64, 32\}, \tag{14}$$

where c_j and $c_{j,K}$ denote full-token and compressed correctness for observation j. Every trajectory therefore begins from a common state without pre-existing compression errors in any evaluated path.

Attack trajectory. We generate clean-initialized feature-objective VEAttack (Mei et al., 2026b) trajectories with $\epsilon = 4/255$, step size $1/255$, and at most 100 steps, evaluating predictions every 10 steps. For each observation and setting, we select the earliest evaluated checkpoint satisfying

$$c_j(x_j^{\text{adv}}) = 1, \qquad c_{j,K}(x_j^{\text{adv}}) = 0. \tag{15}$$

This *first-observed CSF* supplies the fixed state for all subsequent counterfactuals.

Balanced diagnostic cohort. Within each comparison, eligible images are ordered using a fixed sampling order and truncated to the smallest available image count across cells. All eligible questions associated with the retained images are then included.

C.2 Controlled Counterfactual Reallocation

At the fixed adversarial state, exchange identities are selected using retrospective answer support computed from the adversarial representations. For adversarial token i, we apply a multiplicative gate $\widetilde{h}_i = g_i h_i^{\mathrm{adv}}$ and define

$$a_i = - \left. \frac{\partial \mathcal{L}_{\mathrm{NLL}}(y \mid \widetilde{H}^{\mathrm{adv}}, q)}{\partial g_i} \right|_{g_i=1}. \tag{16}$$

Larger a_i indicates that token i provides stronger support for the reference answer.

Let R_K be the direct-token set and O_K its complement. We rank O_K in descending and R_K in ascending order of a_i. For exchange size r,

$$R_{K,r}^{\mathrm{guide}} = (R_K \setminus L_r) \cup H_r, \tag{17}$$

where H_r contains the r highest-support omitted tokens and L_r the r lowest-support retained tokens.

We evaluate the nominal schedule $\rho \in \{0, 5\%, 7.5\%, 10\%, 15\%\}$, with

$$r_K(\rho) = \mathrm{round}(\rho D_K), \tag{18}$$

where D_K is the number of direct slots. For $\rho = (0, 5\%, 7.5\%, 10\%, 15\%)$, this gives $r_{32} = (0, 1, 2, 3, 4)$, $r_{64} = (0, 3, 4, 5, 8)$, $r_{128} = (0, 5, 8, 11, 16)$, and $r_{192} = (0, 8, 12, 16, 24)$.

The matched-random control independently permutes the retained and omitted pools. Each observation uses 40 nested paths, with larger exchanges extending smaller ones. VisionZip merging and aggregation are recomputed after every exchange.

C.3 Retained-Set Allocation Sensitivity and Cumulative Recovery

Let $Y_{jKz,p}(\rho) \in \{0, 1\}$ indicate correctness for observation j under condition z, path p, and ratio ρ. Guided reallocation has one path; the random control has $P = 40$ nested paths. Because recovery need not persist under a larger exchange, curves report cumulative recovery:

$$C_{jKz,p}(\rho) = \max_{\rho' \le \rho} Y_{jKz,p}(\rho'). \tag{19}$$

Thus, $C_{jKz,p}(\rho)$ asks whether a CSF is corrected at any schedule point up to the displayed nominal ratio.

Estimand and aggregation. Random paths are averaged within each observation before cell aggregation:

$$\overline{C}_{jK,\mathrm{rand}}(\rho) = \frac{1}{P} \sum_{p=1}^{P} C_{jK,\mathrm{rand},p}(\rho), \tag{20}$$

and we set $\overline{C}_{jK,\mathrm{guide}}(\rho) = C_{jK,\mathrm{guide},1}(\rho)$. Let $\mathcal{G}$ denote the 12 dataset–budget cells and $\mathcal{J}_{dK}$ their analyzed image–question observations. The reported equal-cell average is

$$\widehat{\mu}_z(\rho) = \frac{1}{|\mathcal{G}|} \sum_{(d,K) \in \mathcal{G}} \left[\frac{1}{|\mathcal{J}_{dK}|} \sum_{j \in \mathcal{J}_{dK}} \overline{C}_{jK,z}(\rho) \right]. \tag{21}$$

The inner mean averages over image–question observations within each cell, while the outer mean assigns equal weight to every dataset–budget cell. The paired contrast is $\widehat{\Delta}(\rho) = \widehat{\mu}_{\mathrm{guide}}(\rho) - \widehat{\mu}_{\mathrm{rand}}(\rho)$.

The average guided–random advantage ranges from 22.5 to 24.1 pp across all nonzero exchange ratios.

Table 5: Cumulative guided (G) and matched-random (R) recovery (%) by compression budget and exchange ratio, with $\Delta = G - R$ reported in percentage points and Avg. denoting the equal-cell average.

K	$\rho = 5\%$			$\rho = 7.5\%$			$\rho = 10\%$			$\rho = 15\%$		
	G	**R**	**Δ**	**G**	**R**	**Δ**	**G**	**R**	**Δ**	**G**	**R**	**Δ**
32	29.1	12.3	+16.8	40.6	19.3	+21.4	45.5	23.6	+21.9	51.1	27.5	+23.6
64	39.1	16.5	+22.6	43.4	21.8	+21.6	44.1	25.9	+18.3	49.7	32.5	+17.2
128	35.8	13.8	+22.0	45.8	20.9	+24.9	49.4	25.4	+24.0	54.9	30.8	+24.1
192	44.4	15.6	+28.8	50.8	22.3	+28.4	53.6	27.0	+26.6	57.9	33.0	+24.9
Avg.	**37.1**	**14.5**	**+22.6**	**45.2**	**21.1**	**+24.1**	**48.2**	**25.5**	**+22.7**	**53.4**	**31.0**	**+22.5**

C.4 Factorial Decomposition of Recovery

At $\rho^* = 7.5\%$, we separate which omitted tokens enter from which retained tokens leave. Incoming tokens are high-support evidence or random omissions; outgoing tokens are low-support retained tokens or random ones:

	Random out	Low-support out
Random in	RR	RL
Evidence in	ER	EL

Let $Q_{jK,z} \in [0, 1]$ denote exact recovery under arm z, averaged over random paths where applicable. We compute the two main effects and their interaction per observation before equal-cell aggregation:

$$e_{jK} = \frac{(Q_{jK,ER} - Q_{jK,RR}) + (Q_{jK,EL} - Q_{jK,RL})}{2}, \tag{22}$$

$$u_{jK} = \frac{(Q_{jK,RL} - Q_{jK,RR}) + (Q_{jK,EL} - Q_{jK,ER})}{2}, \tag{23}$$

$$\eta_{jK} = Q_{jK,EL} - Q_{jK,ER} - Q_{jK,RL} + Q_{jK,RR}. \tag{24}$$

The four arm rates and observation-level effects are aggregated with equation 21; Table 6 reports the resulting point estimates by budget.

Table 6: Exact recovery and factorial effects at $\rho^* = 7.5\%$ by compression budget, with Avg. denoting the equal-cell average.

K	**Exact Recovery (%)**				**Factorial Effect (pp)**		
	RR	**ER**	**RL**	**EL**	**Evidence Restoration**	**Low-Support Removal**	**Interaction**
32	16.1	34.9	20.1	38.6	+18.6	+3.8	−0.3
64	18.4	34.0	25.4	38.5	+14.4	+5.7	−2.4
128	17.2	33.5	31.9	44.4	+14.4	+12.8	−3.9
192	17.9	31.1	34.5	46.4	+12.6	+15.9	−1.3
Avg.	**17.4**	**33.4**	**28.0**	**42.0**	**+15.0**	**+9.6**	−2.0

Evidence restoration is positive at every budget, while the removal effect increases with K. Interaction estimates are negative and smaller in magnitude than either main effect at every budget.

C.5 Representation-Drift Moderation

For a restored evidence token i, let h_i^{clean} and h_i^{adv} be its clean and adversarial encoder representations. We define

$$d_i^{\text{diag}} = \frac{1 - \cos\left(h_i^{\text{clean}}, h_i^{\text{adv}}\right)}{2}. \tag{25}$$

Let E_{jK} denote the high-support omitted tokens restored for observation j at budget K under the $\rho^* = 7.5\%$ intervention. We define their mean representation drift as

$$D_{jK} = \frac{1}{|E_{jK}|} \sum_{i \in E_{jK}} d_i^{\text{diag}}. \tag{26}$$

Clean representations are used only in this post-hoc diagnostic. Let $\overline{D}_{dK}$ and $\overline{e}_{dK}$ be within-cell means. We estimate the common within-cell association between drift and the evidence-restoration effect e_{jK} from equation 24:

$$\widehat{\beta} = \frac{\sum_{d,K} \sum_{j \in \mathcal{J}_{dK}} (D_{jK} - \overline{D}_{dK})(e_{jK} - \overline{e}_{dK})}{\sum_{d,K} \sum_{j \in \mathcal{J}_{dK}} (D_{jK} - \overline{D}_{dK})^2}. \tag{27}$$

Thus, $\widehat{\beta}$ uses only within-cell variation. Drift tertiles provide a grouped summary, and the continuous slope summarizes the corresponding within-cell association.

Table 7: Evidence-restoration effects across within-cell representation-drift tertiles and compression budgets, with mean drift defined as $100\times$ half-cosine distance and Gap as the lower-minus-upper effect.

K	Lower		Middle		Upper		Gap (pp)
	Drift (%)	Effect (pp)	Drift (%)	Effect (pp)	Drift (%)	Effect (pp)	
32	26.1	+28.5	38.0	+20.8	45.2	+6.1	+22.3
64	31.0	+18.2	39.3	+7.0	45.5	+18.0	+0.2
128	34.0	+18.9	40.4	+12.7	45.1	+11.5	+7.4
192	34.4	+16.6	40.2	+15.3	45.1	+5.4	+11.2
Avg.	**31.4**	**+20.5**	**39.5**	**+14.0**	**45.2**	**+10.3**	**+10.3**

The estimated within-cell association is -8.1 pp per 0.1 increase in D_{jK}. Budget-specific tertiles are not uniformly monotone and are therefore interpreted descriptively. The retained-set intervention establishes that changing token allocation can causally alter correctness within this fixed cohort, whereas the drift–recovery analysis supports an association rather than causal mediation.

D Sensitivity Analyses

We test whether CIRA's selective operating point depends on individual design or optimization choices. Unless stated otherwise, each analysis varies one choice on LLaVA-v1.5-7B, POPE, and VisionZip while retaining the evaluation definitions in Section 5.1.

D.1 Preservation Weight

The relative loss weight λ in equation 9 controls the tradeoff between failure induction and full-token preservation. We summarize this operating tradeoff by Selective Gap (Avg. CSFR minus Full ASR), used only as a configuration score because the two metrics have different conditioning sets. Table 8 shows that increasing λ reduces Full ASR while retaining substantial CSFR. Selective Gap remains stable for $\lambda \in [0.6, 1.0]$, with $\lambda = 0.8$ attaining the highest observed value. We therefore use $\lambda = 0.8$ throughout the evaluation.

Table 8: Sensitivity of Full ASR, CSFR, and Selective Gap to the preservation weight λ on POPE. Best values in each row are shown in **bold**.

Metric	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$
Full ASR↓	9.57	6.03	5.79	4.96	**4.61**
CSFR@192↑	8.59	9.08	**9.20**	7.61	6.50
CSFR@128↑	**15.67**	15.30	15.55	14.05	13.68
CSFR@64↑	22.07	22.07	**22.47**	**22.47**	21.94
CSFR@32↑	**27.56**	27.26	25.93	26.96	27.41
Avg. CSFR↑	**18.47**	18.43	18.29	17.77	17.38
Selective Gap↑	8.90	12.40	12.50	**12.81**	12.77

D.2 Candidate Compression-Budget Interval

The admissible interval $[K_{\min}, K_{\max}]$ determines the budget-marginal displacement weights in equation 6. Table 9 varies this interval while keeping the evaluation budgets fixed at $K \in$

$\{192, 128, 64, 32\}$. Selectivity varies modestly across the six candidate intervals. We use $[32, 192]$ as the default because it matches the evaluation range; its Selective Gap is within 0.28 pp of the best observed value.

Table 9: Sensitivity of Full ASR, CSFR, and Selective Gap to the candidate compression-budget interval on POPE. Best values in each row are shown in **bold**.

Metric	Default	Vary $K_{\min}$		Vary $K_{\max}$		
	$[32, 192]$	$[16, 192]$	$[64, 192]$	$[32, 64]$	$[32, 128]$	$[32, 384]$
Full ASR↓	4.96	**4.62**	5.56	5.68	4.97	5.33
CSFR@192↑	7.61	7.38	6.40	**8.86**	7.63	6.52
CSFR@128↑	14.05	**14.86**	12.86	13.23	14.73	12.86
CSFR@64↑	22.47	21.14	21.68	**23.01**	19.95	22.61
CSFR@32↑	26.96	27.43	28.17	**29.20**	26.55	25.22
Avg. CSFR↑	17.77	17.70	17.28	**18.57**	17.21	16.80
Selective Gap↑	12.81	**13.09**	11.71	12.89	12.24	11.48

D.3 Perturbation Budget and Optimization Steps

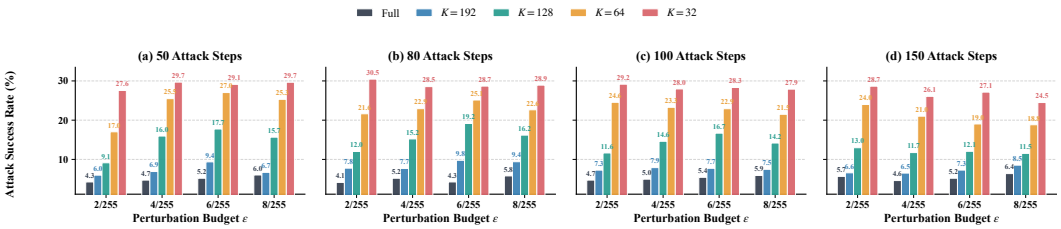


Figure 5: Sensitivity of Full ASR and CSFR to the perturbation budget and number of optimization steps on POPE.

Figure 5 varies $\epsilon \in \{2, 4, 6, 8\}/255$ and the number of optimization steps in $\{50, 80, 100, 150\}$; the step size is set to $\epsilon/4$. Across all 16 configurations, Full ASR remains between 4.14% and 6.38%, whereas Avg. CSFR ranges from 14.93% to 20.82%. Thus, substantial compression-specific failure rates persist while full-token degradation remains limited, and the default $\epsilon = 4/255$, 100-step setting lies within a stable selective region rather than at an isolated optimum.

D.4 Scoring-Layer Configuration Study

The primary score in equation 13 averages class-to-patch attention over the scoring-layer configuration $\mathcal{L}_{\text{score}}$. We compare 14 single-layer, contiguous, and spaced multi-layer configurations using 100-example design subsets from POPE, TextVQA, and MME, evaluated with VisionZip, PruMerge, and VisPruner. We select the configuration with the highest mean Selective Gap across these nine dataset–compressor environments and use it throughout the reported evaluation.

As shown in Figure 6, the selected spaced late-layer triplet $\{-1, -3, -5\}$ achieves 23.45% Avg. CSFR with 5.59% Full ASR, corresponding to a 17.86 pp Selective Gap. Its exclusion rate for clean high-priority tokens also varies less across observer layers than that of the single-layer $\{-2\}$ configuration (range 0.055 versus 0.242). The negative association between cross-layer variation and selectivity ($\rho_s = -0.70$) indicates that configurations with more stable exclusion behavior across depth tend to exhibit higher selectivity. This association is descriptive rather than causal.

E Cross-Model Scope

The main experiments establish CIRA across multiple compressors and budgets on LLaVA. We next test whether paired selectivity persists when the vision encoder, multimodal interface, and language

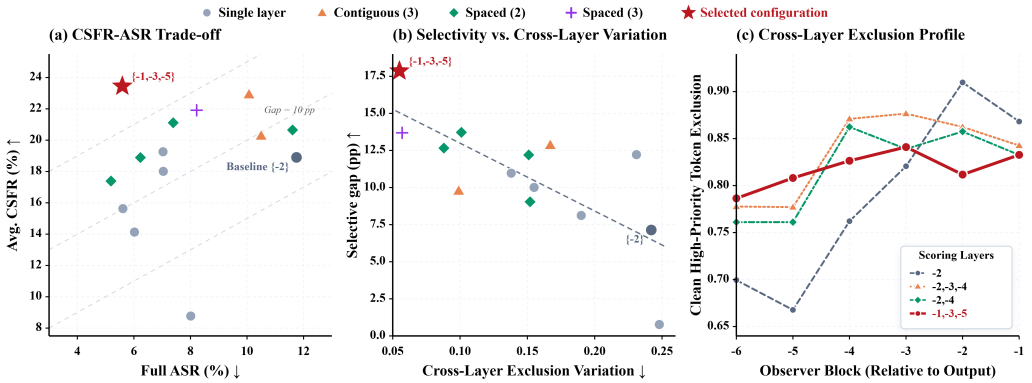


Figure 6: Selectivity and cross-layer exclusion stability across scoring-layer configurations.

model change together, using Qwen3-VL-8B-Instruct (Bai et al., 2025) and InternVL3.5-8B (Wang et al., 2025b).

Here K denotes the number of compressed visual tokens passed to subsequent computation in each model's native interface. For compact reporting, we use $K_1/64/32$, where $K_1 = 96$ for Qwen3-VL and $K_1 = 128$ for InternVL. Avg. CSFR is the arithmetic mean over the three budgets within each model family. Table 10 reports the corresponding CSFR and Full ASR results, while clean accuracy is provided in Table 11.

Table 10: CSFR and Full ASR on Qwen3-VL-8B-Instruct and InternVL3.5-8B. Compressor entries report CSFR at $K_1/64/32$ followed by their average. Bold and underlined values denote the best and second-best downstream-agnostic results, respectively; CAA† is shown as a stronger-access reference and excluded from these rankings.

Dataset	Attack	Full ASR (%)↓	VisionZip $K_1/64/32/\text{Avg.}\uparrow$	VisPruner $K_1/64/32/\text{Avg.}\uparrow$	PruMerge $K_1/64/32/\text{Avg.}\uparrow$	FastV $K_1/64/32/\text{Avg.}\uparrow$
Qwen3-VL-8B-Instruct ($K_1 = 96$)						
POPE	VEAttack	45.93	4.63 / 5.83 / 10.06 / 6.84	2.15 / 2.78 / 4.67 / 3.20	3.91 / 5.01 / 7.49 / 5.47	2.32 / 3.95 / 4.60 / 3.62
	CAGE	46.47	7.13 / 8.13 / 10.19 / 8.48	4.65 / 5.68 / 8.32 / 6.22	**7.23 / 8.23 / 8.95 / 8.14**	4.65 / 6.25 / 11.49 / 7.46
	CAA†	4.40	4.03 / 5.82 / 11.31 / 7.05	2.61 / 4.86 / 8.70 / 5.39	2.47 / 3.33 / 7.06 / 4.29	8.44 / 11.61 / 14.80 / 11.61
	CIRA	**10.66**	**7.59 / 9.82 / 14.14 / 10.52**	**4.98 / 6.56 / 11.22 / 7.59**	3.65 / 4.39 / 7.94 / 5.33	**8.56 / 10.08 / 13.07 / 10.57**
TextVQA	VEAttack	80.14	10.11 / 10.03 / 3.58 / 7.91	10.14 / 7.75 / 1.98 / 6.62	10.22 / 7.87 / 3.34 / 7.14	3.54 / 3.58 / 3.65 / 3.59
	CAGE	66.03	23.96 / 22.56 / 15.22 / 20.58	19.59 / 16.71 / 10.62 / 15.64	21.54 / 18.12 / 16.56 / 18.74	12.38 / 11.00 / 12.41 / 11.93
	CAA†	8.24	46.37 / 42.86 / 35.22 / 41.48	29.28 / 31.23 / 32.84 / 31.12	31.69 / 30.19 / 31.91 / 31.27	43.81 / 42.20 / 38.69 / 41.57
	CIRA	**16.37**	**51.65 / 52.13 / 45.37 / 49.72**	**30.63 / 28.81 / 34.57 / 31.34**	**34.88 / 32.13 / 35.20 / 34.07**	**38.31 / 36.57 / 34.31 / 36.40**
MME	VEAttack	44.11	7.14 / 6.59 / 9.29 / 7.67	5.95 / 6.51 / 6.29 / 6.25	7.60 / 8.31 / 9.69 / 8.54	2.58 / 3.84 / 6.58 / 4.33
	CAGE	41.60	**11.67** / 7.78 / 10.19 / **9.88**	**11.41 / 7.83** / 9.37 / **9.54**	**12.47 / 11.87 / 15.09 / 13.14**	4.42 / 6.64 / 8.53 / 6.53
	CAA†	4.76	8.27 / 9.96 / 12.35 / 10.19	7.37 / 7.72 / 10.34 / 8.48	9.71 / 9.72 / 12.08 / 10.51	9.16 / 11.71 / 15.66 / 12.18
	CIRA	**12.43**	8.34 / **7.79 / 12.79** / 9.64	8.82 / 7.37 / **9.91** / 8.70	7.14 / 5.26 / 7.05 / 6.49	**5.28 / 6.77 / 10.21 / 7.42**
InternVL3.5-8B ($K_1 = 128$)						
POPE	VEAttack	51.46	6.01 / 8.17 / 10.35 / 8.18	**4.20 / 7.44** / 7.91 / **6.52**	**5.28 / 7.56 / 9.03 / 7.29**	7.13 / 11.29 / 12.99 / 10.47
	CAGE	45.01	**7.67 / 13.25 / 17.45 / 12.79**	3.82 / 5.61 / 8.19 / 5.87	3.48 / 5.33 / 8.49 / 5.77	**9.68 / 13.31 / 16.35 / 13.11**
	CAA†	9.85	4.35 / 7.50 / 8.78 / 6.88	1.90 / 4.14 / 6.66 / 4.23	0.77 / 2.21 / 2.65 / 1.88	7.26 / 20.03 / 28.32 / 18.54
	CIRA	**11.68**	4.87 / 7.90 / 11.65 / 8.14	3.04 / 4.01 / **8.42** / 5.16	2.83 / 3.26 / 4.75 / 3.61	3.82 / 8.60 / 12.26 / 8.23
TextVQA	VEAttack	56.92	7.50 / 12.77 / 14.81 / 11.69	3.83 / 7.17 / 11.67 / 7.56	7.52 / 10.50 / 11.86 / 9.96	5.24 / 7.37 / 12.68 / 8.43
	CAGE	55.24	11.00 / 16.23 / 16.24 / 14.49	6.20 / 10.99 / 11.41 / 9.53	6.64 / 10.37 / 13.50 / 10.17	4.94 / 9.89 / 14.83 / 9.89
	CAA†	5.03	10.58 / 22.73 / 23.58 / 18.96	13.64 / 15.66 / 23.68 / 17.66	5.25 / 9.51 / 10.53 / 8.43	10.18 / 33.63 / 45.69 / 29.84
	CIRA	**12.03**	**33.17 / 36.80 / 36.93 / 35.63**	**10.55 / 17.00 / 21.58 / 16.38**	**17.35 / 21.20 / 15.79 / 18.11**	**10.03 / 25.36 / 38.28 / 24.56**
MME	VEAttack	40.63	2.57 / 4.44 / 4.91 / 3.98	2.27 / 4.05 / 4.44 / 3.58	3.28 / 3.93 / 5.71 / 4.31	1.72 / 3.27 / 5.35 / 3.45
	CAGE	37.25	**7.13** / 9.14 / 9.16 / 8.48	**4.07** / 4.93 / 6.32 / 5.10	4.38 / **5.73 / 7.82** / 5.98	**4.70 / 7.02** / 8.43 / 6.72
	CAA†	5.53	2.92 / 3.58 / 8.11 / 4.87	2.41 / 6.19 / 8.31 / 5.64	2.81 / 2.81 / 5.04 / 3.55	3.44 / 13.68 / 18.88 / 12.00
	CIRA	**5.76**	6.88 / **10.63 / 12.50 / 10.00**	3.98 / **7.45 / 8.72 / 6.72**	**5.37** / 5.61 / 7.17 / **6.05**	4.01 / 6.54 / **11.38 / 7.31**

Across both additional model families, CIRA continues to induce compression-specific failures while keeping Full ASR substantially below those of VEAttack and CAGE. The pattern is strongest on TextVQA, where CIRA achieves the highest Avg. CSFR among downstream-agnostic attacks for every compressor on both model families. Results on POPE and MME are more heterogeneous across compressors, but compression-specific failure induction remains observable under target-encoder-only access. Together, these results show that CIRA's compression-selective behavior extends beyond LLaVA to distinct vision encoders and native visual-token interfaces.

F SELECTION STABILIZATION DEFENSE

Translation-Consensus Selection (TCS) stabilizes priority rankings by aggregating aligned scores across spatially translated views.

F.1 TRANSLATION-CONSENSUS SELECTION

For pixel displacement d, define

$$\mathcal{V} = \{T_{0,0}, T_{d,0}, T_{0,d}, T_{d,d}\}. \tag{28}$$

We set $d = 7$ pixels and construct each view by reflection-padding and cropping to the original size. The four views form one batched vision-encoder input.

Score alignment and consensus. For view $v \in \mathcal{V}$, let $\mathbf{s}^{(v)} \in \mathbb{R}^N$ be its encoder-side priority-score vector. Operator $\mathcal{A}_v$ inverse-aligns the score grid to $T_{0,0}$ by bilinear sampling with reflection padding. With patch size $p = 14$, the offset is $d/p = 0.5$ patch:

$$\widetilde{\mathbf{s}}^{(v)} = \mathcal{A}_v\left(\mathbf{s}^{(v)}\right). \tag{29}$$

Equation 11 converts the aligned scores to descending rank quantiles, so each view contributes a priority ranking rather than a score scale.

Selection interface. At compression budget K, TCS replaces the original priority ranking with the cross-view consensus ranking. The unshifted view supplies token features and the compressor's key-similarity metric, while translated views contribute aligned priority scores. Token counts, aggregation, and language-model input length remain unchanged.

Matched evaluation and cost. We evaluate None and TCS on matched adversarial images, questions, references, eligibility sets, and compression budgets. Full-token inference is unchanged, so Full ASR is shared within each matched pair in Table 4. The four views are processed by the vision encoder in one batch, with no additional language-model inference.

F.2 CROSS-VIEW SUPPORT MECHANISM

Cross-view support characterizes the contrast between stable clean evidence and view-specific adversarial replacements. At each compression budget K, the canonical clean Top-K set serves as the reference, while CIRA replacements are tokens that enter the canonical adversarial Top-K set from outside this reference. A candidate's *view support* is the number of aligned views in which it remains within the Top-K set.

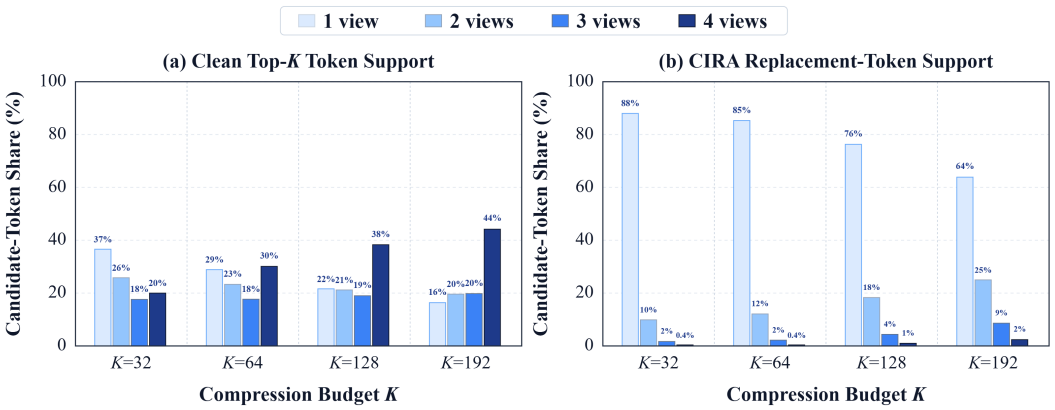


Figure 7: Cross-view support distributions of clean Top-K tokens and CIRA replacement tokens across compression budgets.

Figure 7 shows that CIRA replacements are predominantly view-specific. As K increases from 32 to 192, the one-view share decreases from 88.0% to 63.9%, while fewer than 2.5% are supported by all four views. Clean Top-K tokens show the opposite pattern: their four-view share increases from 20.0% to 44.2%. Averaging aligned rank quantiles therefore downweights isolated replacement spikes while favoring evidence supported across translations.

F.3 Adaptive Evaluation

Standard CIRA is optimized on the unshifted view, with TCS applied only at evaluation. Adaptive CIRA instead optimizes equation 12 over all four public transformations, using the view-specific encoder objectives as differentiable surrogates for rank conversion and Top-K selection. For each view, we cache clean features and ranks and compute GSH and HEP with view-specific hidden-evidence weights over $[K_{\min}, K_{\max}]$. The averaged gradient updates one shared perturbation using the original $\epsilon = 4/255$, step size $1/255$, 100 steps, and $\lambda = 0.8$. Adaptive CIRA retains the same target-encoder access as standard CIRA.

Selection and rank response. For each condition, clean Top-K retention measures the fraction of tokens in the clean Top-K set that remain in the adversarial Top-K set under the corresponding ranking rule. Signed normalized rank change is $(r_i^a - r_i^c)/(N-1)$, where r_i^c and r_i^a denote clean and adversarial ranks under the same ranking rule, and positive values indicate demotion. Retention is averaged per image.

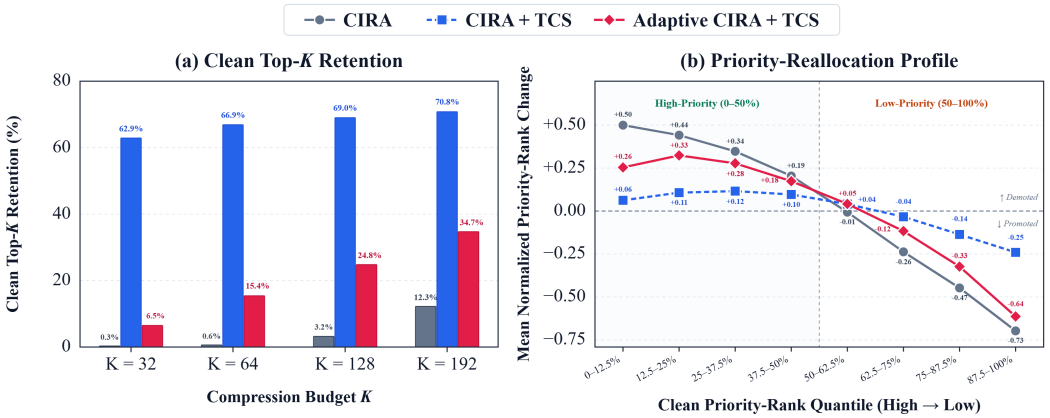


Figure 8: Clean Top-K retention and priority-reallocation profiles under CIRA, CIRA + TCS, and Adaptive CIRA + TCS.

Across the four compression budgets, Figure 8(a) shows that TCS raises clean Top-K retention under CIRA from 0.3–12.3% to 62.9–70.8%. Adaptive CIRA reduces this retention under TCS to 6.5–34.7%. Figure 8(b) shows the corresponding priority reallocation: TCS attenuates both the demotion of clean high-priority tokens and the promotion of initially low-priority tokens, whereas Adaptive CIRA restores much of this signed reallocation.

Together, the cross-view support patterns and the selection responses under TCS show that TCS suppresses view-fragile priority reallocation. Corresponding task-utility results are reported in **Appendix G**.

G Complementary Task-Utility Results

CSFR is the primary clean-conditioned metric for compression-specific failure. We complement it with accuracy-based results that characterize clean utility under compression, post-attack performance, and task-utility recovery under TCS.

Table 11: Clean full-token and compressed accuracy (%) across model families, datasets, compressors, and family-specific compression budgets. The final value in each compressed cell reports the average across the listed budgets.

Dataset	Full ACC	VisionZip	VisPruner	PruMerge	FastV
LLaVA-v1.5-7B ($K = 192/128/64/32/\text{Avg.}$)					
POPE	84.6	84.3/83.5/79.6/73.0/80.1	84.1/82.9/80.5/75.3/80.7	75.6/74.1/72.0/70.1/73.0	81.1/80.0/74.7/68.0/76.0
TextVQA	60.3	54.3/51.6/49.5/45.1/50.1	58.4/58.0/56.8/51.9/56.3	51.9/51.6/51.7/49.4/51.2	51.6/49.4/44.7/37.8/45.9
MME	79.2	77.2/76.3/73.5/69.0/74.0	76.9/76.5/74.0/71.7/74.8	72.9/70.9/71.1/70.8/71.4	76.1/73.9/70.8/67.4/72.1
Qwen3-VL-8B-Instruct ($K = 96/64/32/\text{Avg.}$)					
POPE	86.3	86.1/84.1/80.7/83.6	86.6/85.1/82.3/84.7	86.4/86.1/83.8/85.4	85.1/82.4/74.4/80.6
TextVQA	88.6	46.7/41.1/34.1/40.6	45.8/42.6/41.5/43.3	33.9/31.8/31.2/32.3	52.6/40.6/28.1/40.4
MME	90.0	88.2/88.6/83.4/86.7	88.4/88.1/83.5/86.7	86.7/86.7/83.3/85.6	88.3/85.7/79.4/84.5
InternVL3.5-8B ($K = 128/64/32/\text{Avg.}$)					
POPE	82.2	81.4/80.8/77.1/79.8	80.4/80.5/78.3/79.7	79.0/79.0/78.8/78.9	81.0/79.5/74.9/78.5
TextVQA	71.5	64.7/48.6/37.5/50.3	57.7/47.0/40.4/48.4	46.8/37.7/32.8/39.1	68.5/57.6/44.3/56.8
MME	88.6	87.5/83.8/78.2/83.2	84.8/82.2/77.0/81.3	84.0/80.8/78.5/81.1	88.2/85.2/77.8/83.7

G.1 Clean Utility under Compression

Table 11 shows that high-budget settings preserve most POPE and MME accuracy across model families, while TextVQA generally exhibits larger losses under compression. Related analyses also find task-dependent visual-token requirements, with OCR tasks relying on visual information deeper into the decoder (Wang et al., 2026b). These values provide the clean reference for the post-attack comparisons below.

G.2 Post-Attack Task Utility

Unlike CSFR, adversarial accuracy is computed over the complete evaluation set and therefore reflects both pre-existing compression errors and attack-induced failures. It provides a complementary view of overall task degradation under CIRA.

Table 12: Full-token clean-to-adversarial accuracy and compressed adversarial accuracy (%) under CIRA across model families, datasets, compressors, and family-specific budgets. The final value in each compressed cell reports the budget average.

Dataset	Full ACC	VisionZip	VisPruner	PruMerge	FastV
LLaVA-v1.5-7B (compressed Adv. ACC: $K = 192/128/64/32/\text{Avg.}$)					
POPE	84.6 → 82.6	78.4/74.0/64.6/58.0/68.8	80.0/77.7/71.7/64.1/73.4	63.3/55.9/51.2/50.3/55.2	76.8/71.4/61.0/53.4/65.7
TextVQA	60.3 → 57.1	46.7/40.0/32.2/26.1/36.3	51.3/49.5/44.6/38.6/46.0	36.9/32.2/27.4/25.0/30.4	45.5/41.9/33.5/25.7/36.7
MME	79.2 → 77.2	73.2/70.7/65.5/59.5/67.2	73.2/70.9/64.1/60.5/67.2	62.2/58.9/55.4/51.5/57.0	73.2/71.6/63.7/58.8/66.8
Qwen3-VL-8B-Instruct (compressed Adv. ACC: $K = 96/64/32/\text{Avg.}$)					
POPE	86.3 → 81.3	75.9/74.2/71.2/73.8	78.1/76.7/73.5/76.1	79.8/79.8/76.0/78.5	76.0/73.9/67.7/72.5
TextVQA	88.6 → 75.7	21.1/20.5/21.4/21.0	31.9/31.9/31.0/31.6	22.2/22.4/22.0/22.2	30.7/25.6/20.6/25.6
MME	90.0 → 81.5	74.9/76.1/69.8/73.6	73.9/76.5/72.5/74.3	74.3/77.1/74.6/75.3	79.9/77.5/70.9/76.1
InternVL3.5-8B (compressed Adv. ACC: $K = 128/64/32/\text{Avg.}$)					
POPE	82.2 → 76.1	75.5/75.4/71.8/74.2	75.8/75.9/73.4/75.0	75.7/75.0/75.4/75.4	77.3/73.1/70.2/73.5
TextVQA	71.5 → 65.9	42.3/32.3/27.1/33.9	53.9/43.6/36.1/44.5	39.5/31.8/29.5/33.6	58.4/42.7/29.6/43.6
MME	88.6 → 85.3	81.4/75.7/71.5/76.2	81.3/77.8/74.9/78.0	80.2/78.3/74.7/77.7	83.8/79.8/72.9/78.8

Table 12 shows that CIRA reduces budget-averaged compressed accuracy by 3.3–20.8 pp across the evaluated model–dataset–compressor settings. After averaging compressors within each model–dataset pair and weighting the nine pairs equally, the mean compressed accuracy drop is 9.6 pp, compared with 5.4 pp under full-token inference. This aggregate view complements the paired CSFR analysis by quantifying end-task degradation under compressed inference.

G.3 Task-Utility Recovery under TCS

Table 13 reports matched task accuracy without and with TCS. Each comparison fixes the input image, question, and compression budget, with TCS applied only to compressed inference.

Across the three benchmarks, TCS changes clean Avg. ACC by at most 0.52 pp in magnitude, while recovering 6.77–13.20 pp under CIRA, with larger gains at tighter budgets. Under Adaptive

Table 13: Matched compressed accuracy (%) without and with TCS on LLaVA-v1.5-7B with VisionZip under clean, CIRA, and Adaptive CIRA conditions across datasets and compression budgets. Entries show None → TCS, with annotations reporting the corresponding change.

Evaluation	$K = 192$	$K = 128$	$K = 64$	$K = 32$	Avg.
POPE					
Clean	84.30 → 83.10 ↓1.20	83.50 → 82.50 ↓1.00	79.60 → 78.40 ↓1.20	73.00 → 74.30 ↑1.30	80.10 → 79.58 ↓0.52
CIRA	78.40 → 81.30 ↑2.90	74.00 → 80.90 ↑6.90	64.60 → 78.30 ↑13.70	58.00 → 75.40 ↑17.40	68.75 → 78.98 ↑10.23
Adaptive CIRA	78.80 → 78.60 ↓0.20	76.40 → 76.00 ↓0.40	66.80 → 71.60 ↑4.80	56.30 → 62.20 ↑5.90	69.58 → 72.10 ↑2.52
TextVQA					
Clean	54.30 → 53.20 ↓1.10	51.60 → 51.30 ↓0.30	49.50 → 49.40 ↓0.10	45.10 → 45.40 ↑0.30	50.13 → 49.83 ↓0.30
CIRA	46.70 → 52.40 ↑5.70	40.00 → 50.90 ↑10.90	32.20 → 49.00 ↑16.80	26.10 → 45.50 ↑19.40	36.25 → 49.45 ↑13.20
Adaptive CIRA	48.30 → 49.00 ↑0.70	45.10 → 46.90 ↑1.80	36.00 → 42.10 ↑6.10	26.70 → 33.00 ↑6.30	39.03 → 42.75 ↑3.72
MME					
Clean	77.20 → 76.50 ↓0.70	76.30 → 76.00 ↓0.30	73.50 → 74.30 ↑0.80	69.00 → 70.60 ↑1.60	74.00 → 74.35 ↑0.35
CIRA	73.20 → 75.00 ↑1.80	70.70 → 75.80 ↑5.10	65.50 → 74.00 ↑8.50	59.50 → 71.20 ↑11.70	67.23 → 74.00 ↑6.77
Adaptive CIRA	74.50 → 73.60 ↓0.90	70.90 → 72.00 ↑1.10	62.90 → 68.40 ↑5.50	58.50 → 62.00 ↑3.50	66.70 → 69.00 ↑2.30

CIRA, TCS recovers less utility, consistent with the attack partially restoring the priority reallocation suppressed by TCS.

H Limitations and Future Work

Temporal and contextual dependencies. The present evaluation is restricted to single-image inference. In video, multi-image, and multi-turn settings, evidence retention also depends on temporal redundancy and evolving context (Shen et al., 2024; Li et al., 2026b; Wang et al., 2026c). Run-Length Pruning, for instance, combines temporal redundancy removal with token distillation (Ma et al., 2026). Whether encoder-only priority manipulation remains compression-selective when evidence is distributed across frames or dialogue turns remains an open question.

Compression mechanisms beyond token selection. Our diagnostics examine retained-set allocation and representation drift, but do not exhaust the mechanisms underlying compression-induced errors. Adaptive pruning changes allocation across inputs or layers (Ye et al., 2025; Chen et al., 2026; Li et al., 2026a); learned summarization transforms the token representation (Bulat et al., 2026); and recoverable routing permits deferred tokens to re-enter subsequent selection stages (Yang et al., 2026). These mechanisms complicate a description based on a single token ranking. Moreover, spatial disruption (Huang et al., 2026) and positional or attentional distortion (Cho et al., 2026) are not separately identified by our retained-set interventions. Paired CSF evaluation remains applicable, but attributing failures within these compression mechanisms requires additional diagnostics.

Beyond task correctness. CSF is defined through task-level correctness rather than response safety or targeted attack success. Accordingly, preserving full-token correctness does not establish safety preservation, and compressed-path errors do not necessarily constitute safety-alignment failures. Targeted manipulation and multimodal jailbreaks (Zhang et al., 2025a; Qi et al., 2024; Shayegani et al., 2024) provide distinct settings for paired evaluation, requiring outcome criteria tailored to the corresponding security objective.

I QUALITATIVE CASE STUDIES

POPE — **LLaVA-v1.5-7B**

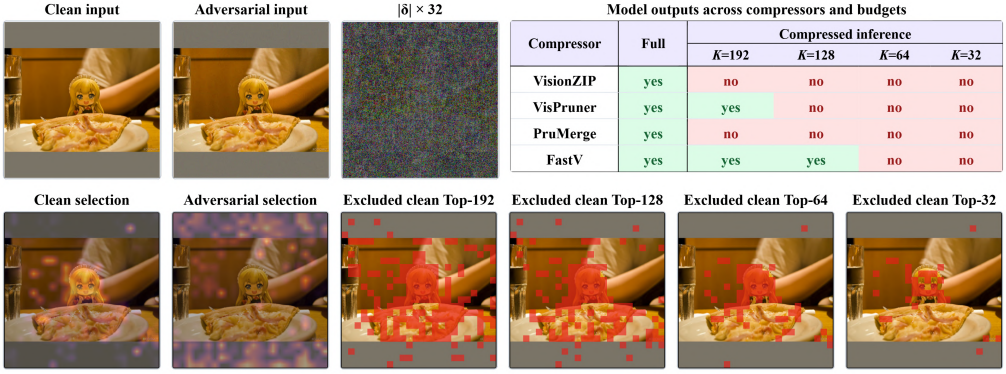
Question: Is there a pizza in the image?

Ground truth: yes

Clean input | Adversarial input | $|\delta| \times 32$

Model outputs across compressors and budgets

Compressor	Full	Compressed inference			
		K=192	K=128	K=64	K=32
VisionZIP	yes	no	no	no	no
VisPruner	yes	yes	no	no	no
PruMerge	yes	no	no	no	no
FastV	yes	yes	yes	no	no



TextVQA — **LLaVA-v1.5-7B**

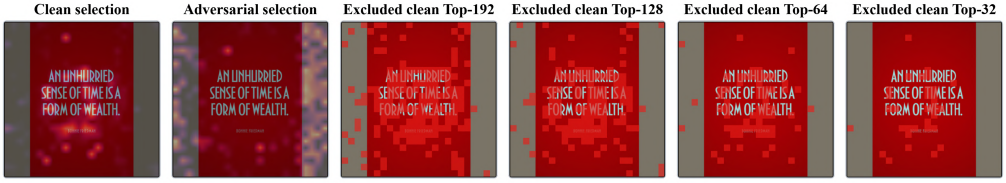
Question: what is the book about?

Ground truth: wealth / time

Clean input | Adversarial input | $|\delta| \times 32$

Model outputs across compressors and budgets

Compressor	Full	Compressed inference			
		K=192	K=128	K=64	K=32
VisionZIP	wealth	film	productivity	life	States
VisPruner	wealth	lives	Gogh	States	States
PruMerge	wealth	Olien	States	States	States
FastV	Time	Wealth	Money	Cooking	Religion



MME — **LLaVA-v1.5-7B**

Question: Is this artwork created by palma vecchio?

Ground truth: yes

Clean input | Adversarial input | $|\delta| \times 32$

Model outputs across compressors and budgets

Compressor	Full	Compressed inference			
		K=192	K=128	K=64	K=32
VisionZIP	yes	no	no	no	no
VisPruner	yes	no	no	no	no
PruMerge	yes	no	no	no	no
FastV	yes	no	no	no	no

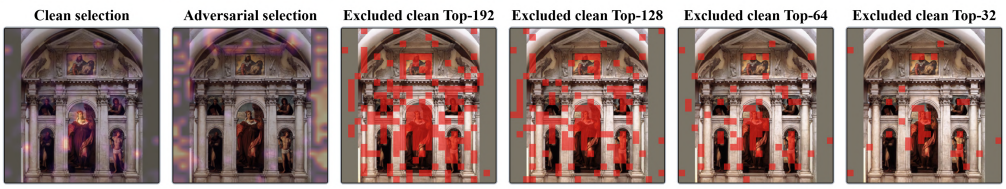


Figure 9: Qualitative examples of compression-specific failures on LLaVA-v1.5-7B across visual-token compressors and retention budgets.

POPE — **Qwen3-VL-8B-Instruct**

Question: Is there a cell phone in the image?

Ground truth: yes

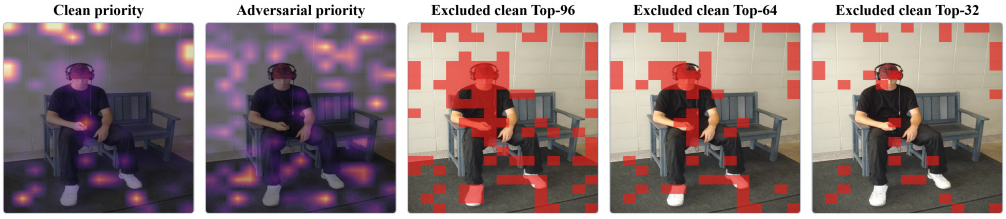
Clean input | Adversarial input | $|\delta| \times 32$



Model outputs across compressors and budgets

Compressor	Full	Compressed inference		
		K=96	K=64	K=32
VisionZIP	yes	no	no	no
VisPruner	yes	no	no	no
PruMerge	yes	no	no	no
FastV	yes	no	no	no

Clean priority | Adversarial priority | Excluded clean Top-96 | Excluded clean Top-64 | Excluded clean Top-32



TextVQA — **Qwen3-VL-8B-Instruct**

Question: what number can be seen?

Ground truth: 5

Clean input | Adversarial input | $|\delta| \times 32$



Model outputs across compressors and budgets

Compressor	Full	Compressed inference		
		K=96	K=64	K=32
VisionZIP	5	1	2	2
VisPruner	5	5	2	2
PruMerge	5	8	6	666
FastV	5	5	5	3

Clean priority | Adversarial priority | Excluded clean Top-96 | Excluded clean Top-64 | Excluded clean Top-32



MME — **Qwen3-VL-8B-Instruct**

Question: Is this a picture of Church of Saint Giles in Prague?

Ground truth: yes

Clean input | Adversarial input | $|\delta| \times 32$



Model outputs across compressors and budgets

Compressor	Full	Compressed inference		
		K=96	K=64	K=32
VisionZIP	yes	no	no	no
VisPruner	yes	no	no	no
PruMerge	yes	no	no	no
FastV	yes	no	no	no

Clean priority | Adversarial priority | Excluded clean Top-96 | Excluded clean Top-64 | Excluded clean Top-32

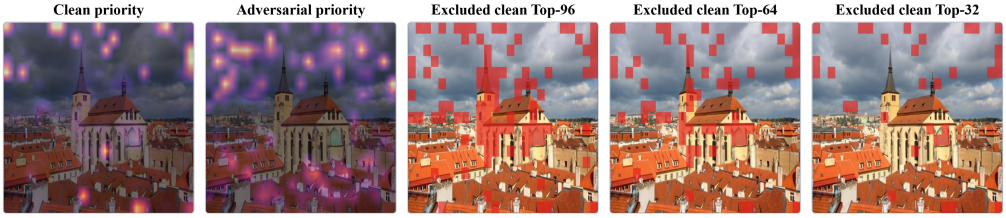


Figure 10: Qualitative examples of compression-specific failures on Qwen3-VL-8B-Instruct across visual-token compressors and retention budgets.

POPE — **InternVL3.5-8B**

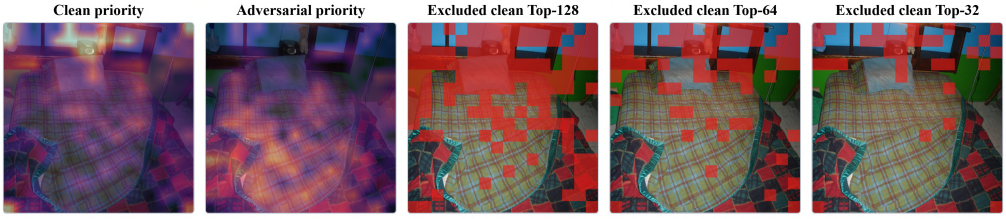
Question: Is there a dining table in the image?

Ground truth: no

Clean input | Adversarial input | $|\delta| \times 32$

Model outputs across compressors and budgets

Compressor	Full	Compressed inference		
		K=128	K=64	K=32
VisionZIP	no	no	yes	yes
VisPruner	no	yes	yes	yes
PruMerge	no	yes	yes	yes
FastV	no	yes	yes	yes



TextVQA — **InternVL3.5-8B**

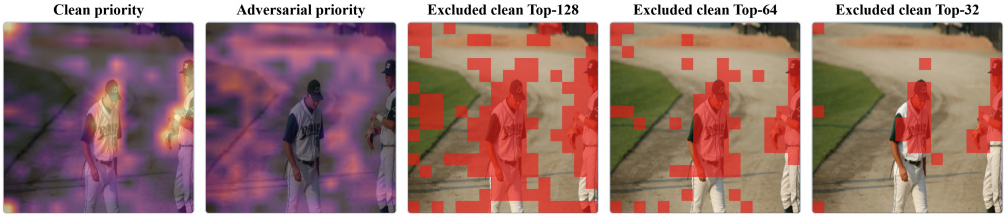
Question: what team name is on the player's jersey?

Ground truth: rays

Clean input | Adversarial input | $|\delta| \times 32$

Model outputs across compressors and budgets

Compressor	Full	Compressed inference		
		K=128	K=64	K=32
VisionZIP	Rays	braves	braves	GIANTS
VisPruner	Rays	rangers	red sox	whitecaps
PruMerge	Rays	braves	braves	braves
FastV	Rays	RAYS	braves	NEW YORK



MME — **InternVL3.5-8B**

Question: Is there a white bird in the image?

Ground truth: yes

Clean input | Adversarial input | $|\delta| \times 32$

Model outputs across compressors and budgets

Compressor	Full	Compressed inference		
		K=128	K=64	K=32
VisionZIP	yes	no	no	no
VisPruner	yes	no	no	no
PruMerge	yes	no	no	no
FastV	yes	yes	yes	no

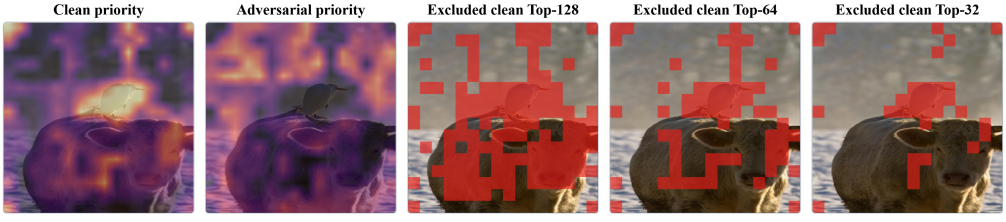


Figure 11: Qualitative examples of compression-specific failures on InternVL3.5-8B across visual-token compressors and retention budgets.