CONTRA: VISUAL TOKEN PRUNING VIA CONTRAST INVARIANCE IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) incur substantial inference costs from the sheer volume of visual tokens, especially in text-rich scenarios where textual and visual semantics are deeply entangled within visual tokens. Existing pruning methods rank tokens by holistic visual salience and thus frequently discard text-critical tokens as redundant background. We address this limitation with the **Contrast Invariance Principle**—textual semantics remain stable under image-level contrast perturbations, whereas background semantics drift substantially. Because glyphs are only the extremal case of ordinal structure in images, the same principle transfers safely to natural, largely text-free scenes. Building on this principle, we propose **ConTra**, a training-free visual token pruning framework that generates a contrast-transformed view via the Chromatic Contrast Transformation (CCT) and computes cross-view token consistency to identify and preserve text-critical tokens. ConTra requires no attention matrices and only one batched vision-encoder pass over the two views, making it FlashAttention-compatible and readily deployable on off-the-shelf MLLMs without retraining. Across 22 benchmarks spanning multiple model scales, ConTra retains 90.1% performance at a 70% pruning rate on text-rich tasks—surpassing the strongest prior method by 8.1 points—while remaining fully competitive with query-guided methods on general tasks (98.1% retention) despite using no query information.

1 INTRODUCTION

The inference of Multimodal Large Language Models (MLLMs)Radford et al. (2019); Zhang et al. (2024e); Nguyen et al. (2023); Brown et al. (2020) suffers from significant computational and memory burdens due to the explosion of visual tokens. Standard vision encoders frequently produce an inflation of tokens when processing high-resolution imagesZhang et al. (2024c); Li et al. (2023a). Even in state-of-the-art models designed with dynamic resolution like Qwen3-VLBai et al. (2025a), the average sequence length of visual tokens on the CC-OCRYang et al. (2024b) dataset still surpasses 2,300. Consequently, the quadratic complexity of the self-attention mechanism across every transformer layer poses a severe bottleneck for practical MLLM deployment. This bottleneck becomes particularly pronounced in text-rich vision-language scenarios—such as document understanding, form processing, and scene text recognition—which inherently require fine-grained token representations.

To tackle this issue, existing literature extensively explores training-free token reduction. Current approaches predominantly fall into three categories: similarity-based merging (e.g., ToMeBolya et al. (2023), DARTWen et al. (2025)), attention-based pruning (e.g., HoloVZou et al. (2025), Fast-VChen et al. (2024a), VisionZipYang et al. (2024a)), and cross-modal alignment selection (e.g., SparseVLMZhang et al. (2025b), MMTokDong et al. (2026)).

However, existing token reduction paradigms suffer severe performance degradation on text-rich vision-language tasks. We attribute this vulnerability to a fundamental limitation: in multimodal LLMs, visual tokens encapsulate deeply entangled modalities, where *textual semantics* (linguistic information) and *visual semantics* (background features) coexist within the ViT space. Current methods compress tokens based on holistic visual salience (e.g., attention weights) without decoupling these representations, inevitably discarding semantically essential text-bearing tokens. This entanglement is corroborated by recent findings: ViTs primarily encode text as low-level texture fea-

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.

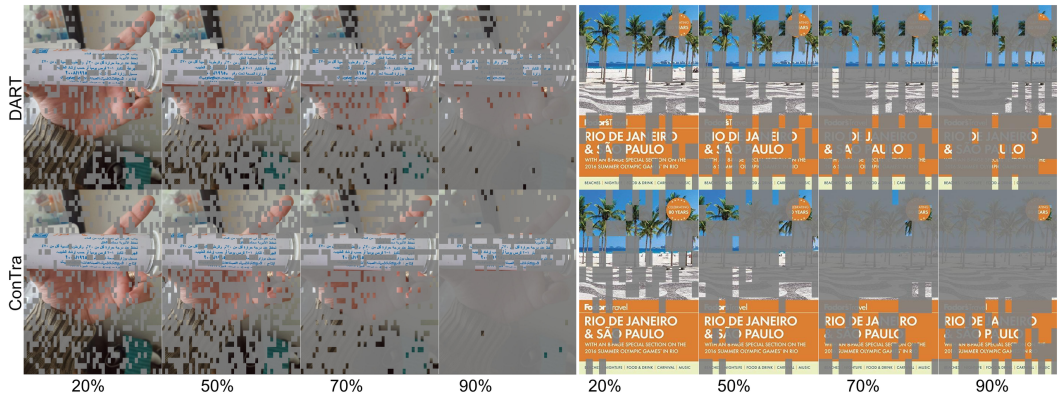

Figure 1: Visual Token Pruning Comparison on Text-Containing Images. Our method effectively preserves visual tokens critical for text regions at high pruning rates.

tures, rendering linguistic tokens indistinguishable from background textures Wang & Ma (2025), while self-attention mechanisms inherently group patches by visual homogeneity like color and background patterns Mehrani et al. (2023). Consequently, text-rich tokens—which are semantically pivotal yet visually sparse—are penalized with low salience scores and mistakenly pruned.

To address these challenges, we propose **ConTra** (**Con**trast **Tra**nsformation Consistency Guided Vision Token Pruning), a training-free framework that leverages cross-transformation semantic consistency to safeguard pivotal tokens. Our approach is motivated by what we term the **Contrast Invariance Principle**: textual and visual tokens exhibit asymmetric sensitivity to contrast perturbations. Textual semantics depend on structural boundaries rather than absolute pixel intensities, remaining invariant under contrast modulation. Conversely, background semantics are bound to pixel-level appearances and highly vulnerable to alterations. For instance, inverting image colors preserves text legibility due to structural permanence but severely distorts the background context. In simpler terms, **text reads through contrast changes, while backgrounds do not.** As formalized in Section 3.3, this principle is not specific to glyphs but reflects a broader *Ordinal Structure Invariance* of visual representations, which also accounts for ConTra's strong results on general, largely text-free images.

Methodologically, we introduce a Chromatic Contrast Transformation (CCT) parameterized by a tunable factor α. When projected into the LLM's semantic space, text-bearing tokens demonstrate high representation consistency due to their structural resilience, whereas background tokens suffer significant representation drift. This divergence yields a measurable *consistency gap* that segregates text-critical tokens from redundant ones without computing any attention matrices (Figure 1). **Despite its simplicity, ConTra is surprisingly effective**: it requires no attention matrices and only one batched vision-encoder pass over the two views, yet outperforms complex attention-based and similarity-based heuristics. Moreover, since pruning occurs entirely before tokens enter the LLM, every subsequent transformer layer—and the KV cache—operates on the reduced sequence, whereas in-LLM approaches (e.g., Fast-V) inevitably execute several full-cost layers before reduction takes effect.

Extensive evaluations show that ConTra achieves state-of-the-art results on text-rich benchmarks, retaining 90.1% performance at a 70% pruning rate where the strongest prior method drops to 81.9%. These gains come at no cost to general capability: despite being query-agnostic, ConTra remains fully competitive with query-guided methods on general vision-language tasks (98.1% retention at the same rate), incurring no collateral damage on text-free content. Furthermore, ConTra is fully model-agnostic, FlashAttention-compatible, and can be integrated into existing MLLMs without any retraining.

Our contributions are threefold:

- **We identify a fundamental limitation of existing pruning methods**: text-critical tokens receive low salience scores despite their semantic importance, as current approaches fail to decouple textual from visual semantics.

- **We reveal the Contrast Invariance Principle** and propose ConTra, a training-free and attention-free framework that decouples textual from visual semantics via contrast transformation consistency.
- **Extensive experiments demonstrate ConTra's effectiveness**: state-of-the-art retention on both text-rich (90.1% at 70% pruning) and general (98.1%) benchmarks, while remaining fully model-agnostic and FlashAttention-compatible.

2 RELATED WORK

2.1 VISUAL TOKEN REDUNDANCY IN MLLMS

MLLMs typically employ a Vision Transformer (ViT) followed by a projection module to align visual representations into the LLM's linguistic space. However, processing high-resolution images generates hundreds to thousands of visual tokens, imposing an immense computational burden during inference. Recent advanced architectures attempt to alleviate this inflation: Qwen2.5-VL Bai et al. (2025b) dynamically adjusts resolution to optimize token counts, while Qwen3-VL Bai et al. (2025a) introduces a DeepStack architecture to aggregate multi-level features from intermediate layers. Despite these optimizations, the volume of visual tokens remains prohibitively large; for instance, a standard 2048×1024 image still yields 2,678 tokens in Qwen2.5-VL, establishing a severe bottleneck for inference throughput and memory bandwidth.

2.2 VISUAL TOKEN REDUCTION PARADIGMS

Existing visual token reduction methods can be broadly classified into four paradigms: **i) Model-based compression** maps visual tokens into denser representations via learned transformations (e.g., Qwen3-VL's Patch Merger or LLaVA-OneVision's bilinear downsampling). However, these methods yield rigid compression ratios and mandate architectural modifications and retraining. **ii) Similarity-based merging** reduces redundancy by clustering and combining semantically similar tokens (e.g., ToMe Bolya et al. (2023), DART Wen et al. (2025)). **iii) Attention-based pruning** exploits attention sparsity to eliminate uninformative patches (e.g., HoloV Zou et al. (2025), PuMerger Cao et al. (2023), VisionZip Yang et al. (2024a), Fast-V Chen et al. (2024a)). Crucially, this line of work heavily relies on explicit attention matrices, creating a fundamental architectural conflict with FlashAttention. **iv) Cross-modal selection** leverages textual instructions to guide token filtering (e.g., MMTok Dong et al. (2026), SparseVLM Zhang et al. (2025b)), yet its performance hinges on the availability of explicit text queries, restricting its general applicability.

None of these paradigms protect text-rich regions, making critical tokens prone to erroneous elimination. Unlike traditional training-intensive contrastive learning, our approach non-invasively exploits text semantic invariance under contrast modulation during inference. This training-free mechanism precisely preserves text-critical tokens, bridging prior limitations.

3 METHODOLOGY

Visual tokens in MLLMs entangle two modalities: *textual semantics* (structural layouts) and *visual semantics* (background features). ConTra decouples them via the **Contrast Invariance Principle**: textual semantics survive contrast changes while background semantics collapse (Figure 2). This asymmetry identifies text-critical tokens in three steps—(1) generate a contrast-transformed view, (2) encode both views in a single batched vision-encoder call and compute cross-view token consistency, and (3) retain high-consistency (textual) tokens while pruning drifting (background) ones—requiring neither training nor attention matrices, with negligible overhead.

3.1 CHROMATIC CONTRAST TRANSFORMATION

To effectively segregate textual from visual tokens, the perturbation mechanism must satisfy a dual constraint: it must distort absolute pixel values (upon which background visual semantics depend) while preserving the underlying ordinal contrast topologies (upon which character recognition relies). To satisfy this requirement, we formalize the **CCT** (**C**hromatic **C**ontrast **T**ransformation), which systematically modulates image contrast via a tunable scale factor α.

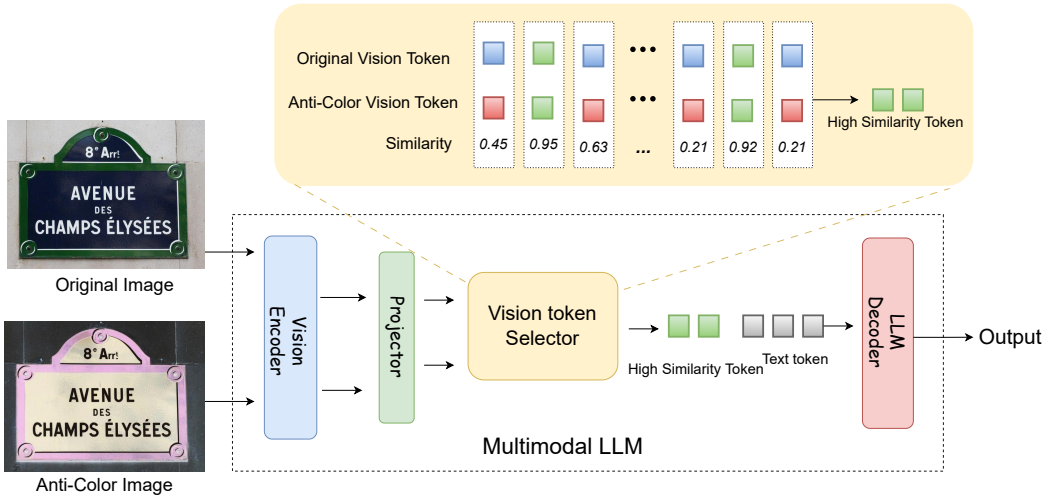

Figure 2: Overview of ConTra. The original and contrast-transformed views are encoded in a single batched pass, and per-token cosine consistency separates text-critical tokens (retained) from drifting background tokens (pruned). Only the original representations of retained tokens are fed into the LLM.

Formally, let an input image be denoted as $\mathbf{I} \in [0,1]^{H\times W\times C}$, and its channel-wise spatial mean vector as $\boldsymbol{\mu}(\mathbf{I}) \in \mathbb{R}^C$. The CCT parameterized by a non-zero scaling factor $\alpha \in \mathbb{R} \setminus \{0\}$ is mathematically defined as:

$$\mathrm{CCT}_\alpha(\mathbf{I}) = \alpha\big(\mathbf{I} - \boldsymbol{\mu}(\mathbf{I})\big) + \boldsymbol{\mu}(\mathbf{I}). \tag{1}$$

Intuitively, CCT scales the directional deviation of each pixel relative to its respective channel mean. Specifically, $\alpha = 1$ recovers the identity mapping, $\alpha > 1$ amplifies global contrast, $0 < \alpha < 1$ compresses the dynamic range, and $\alpha < 0$ inverts the foreground–background intensity relationship.

The following property confirms that CCT meets the dual constraint stated above by construction:

Property 1 (Ordinal Structure Invariance) *For any arbitrary pair of spatial coordinates (i, j) and (k, l) within the image tensor $\mathbf{I}$, the spatial gradient under CCT satisfies:*

$$\mathrm{CCT}_\alpha(\mathbf{I})_{i,j} - \mathrm{CCT}_\alpha(\mathbf{I})_{k,l} = \alpha(\mathbf{I}_{i,j} - \mathbf{I}_{k,l}). \tag{2}$$

Property 1 is a direct consequence of the linearity of CCT, yet it captures precisely the asymmetry that ConTra exploits. Textual content is carried predominantly by ordinal contrast structure: character shapes are defined by the spatial arrangement of contrast boundaries rather than absolute intensities, remaining legible under contrast compression and even polarity reversal—a robustness long observed in human reading Legge et al. (1987) and routinely exploited as a semantics-preserving augmentation Cubuk et al. (2019). Background semantics, in contrast, are bound to absolute pixel-level appearance statistics, which CCT deliberately perturbs. **In simpler terms, text depends on *where* edges occur, not *what* the absolute colors are.** CCT thus preserves exactly what text tokens encode while perturbing exactly what background tokens depend on, providing a reliable touchstone to decouple the two.

Color Inversion. Setting $\alpha = -1$ instantiates CCT as a strict color inversion:

$$\mathrm{CCT}_{-1}(\mathbf{I}) = 2\boldsymbol{\mu}(\mathbf{I}) - \mathbf{I}, \tag{3}$$

which reverses the foreground–background intensity topology while preserving ordinal contrast gradients. Textual tokens therefore remain stable, whereas background representations undergo maximal distortion, inducing the largest *consistency gap* in the semantic space. Empirically, $\alpha = -1$ yields the strongest discriminative signal on text-rich benchmarks (Section **??**).

3.2 TOKEN SELECTION

Given the original image $\mathbf{I}^{\text{ori}}$ and its perturbed counterpart $\mathbf{I}^{\text{ct}} = \text{CCT}_{\alpha}(\mathbf{I}^{\text{ori}})$, both views are processed by the vision encoder and its subsequent projection layer to synthesize twin visual token sequences: $\mathbf{H}^{\text{ori}} = \{\mathbf{h}_p^{\text{ori}}\}_{p=1}^{N}$ and $\mathbf{H}^{\text{ct}} = \{\mathbf{h}_p^{\text{ct}}\}_{p=1}^{N}$, where N denotes the total number of spatial visual tokens, and $\mathbf{h}_p \in \mathbb{R}^d$ signifies the hidden representation within the LLM's semantic space.

To quantify the divergence between views without parameter updates, we compute a **Cross-View Consistency Score** $s_p = (\mathbf{h}_p^{\text{ori}})^\top \mathbf{h}_p^{\text{ct}} / (\|\mathbf{h}_p^{\text{ori}}\| \cdot \|\mathbf{h}_p^{\text{ct}}\|)$ via cosine similarity. Semantically stable text tokens gravitate toward $s_p \approx 1$, while background regions drift lower ($s_p < 1$). Based on this training-free *consistency gap*, ConTra supports two complementary **Token Selection Strategies** tailored to different downstream constraints:

(1) Fixed Pruning Ratio: Given a pre-defined retention ratio $\rho \in (0, 1]$, we preserve a deterministic sequence length via $\mathcal{R}_\rho = \text{TopK}\big(\{s_p\}_{p=1}^{N},\ \lfloor \rho N \rfloor\big)$. This strategy guarantees fixed token budgets, which is ideal for static batching and synchronized inference pipelines. **(2) Similarity Thresholding:** Alternatively, we define an absolute threshold $\tau \in (0, 1]$ to extract an instance-dependent sequence $\mathcal{R}_\tau = \{p \mid s_p \geq \tau\}$. Benefiting from contrast-invariance, text-rich images naturally retain denser token sequences to safeguard layout details, whereas text-sparse images automatically trigger higher compression ratios as redundant background tokens fall below τ.

Crucially, regardless of the selection strategy employed, the retained visual sequences inherit exclusively the *original* representations $\{\mathbf{h}_p^{\text{ori}}\}_{p \in \mathcal{R}}$ for downstream autoregressive inference. The contrast-transformed view serves only as a *probe* during token selection and is discarded afterward, introducing no modification to the original semantic space.

3.3 THE CONTRAST INVARIANCE PRINCIPLE

We now establish the theoretical foundation explaining why ConTra's consistency-based selection effectively distinguishes text from background.

Mechanism: Why Consistency Distinguishes Text. The Contrast Invariance Principle relies on two factors. First, text semantics depend on structural boundaries (ordinal contrast) rather than absolute pixel intensities, which CCT preserves. Second, we measure consistency after the projection layer, which aligns visual features with the LLM's word embedding space. This provides *semantic anchoring*: text tokens map to stable embedding clusters that contrast perturbations cannot dislodge, whereas appearance-dominated background tokens lack such anchoring. Consequently, text tokens exhibit minimal feature drift ($\|\Delta \mathbf{h}_p^{\text{text}}\| \approx 0$) while background tokens drift significantly ($\|\Delta \mathbf{h}_p^{\text{bg}}\| \gg 0$).

Statistical Separability. This asymmetric drift induces a bimodal distribution over consistency scores. We approximate this with a Gaussian Mixture Model (GMM), $p(s) = \pi_t \mathcal{N}(s; \mu_t, \sigma_t^2) + (1 - \pi_t)\mathcal{N}(s; \mu_b, \sigma_b^2)$ with $\mu_t > \mu_b$. In Appendix C, we empirically verify that the fitted parameters consistently satisfy Ashman's criterion (D ¿ 2), justifying the isolation of text-critical tokens via simple Top-K selection.

Beyond Text: Ordinal Structure Invariance. CCT preserves the ordinal contrast topology of *any* spatial structure. Text is an extremal case (glyphs are pure structure), but object contours and geometry in natural images belong to the same invariant class, mapping to stable semantic clusters under the anchoring hypothesis. Conversely, tokens dominated by absolute appearance (e.g., uniform color, lighting) drift and are discarded. Thus, consistency selection retains structure-carrying tokens while discarding redundant appearance ones. This predicts that ConTra should incur *no collateral damage* on text-free content, which we verify on general benchmarks in Section 4.3.

3.4 COMPLEXITY AND COMPUTATIONAL EFFICIENCY ANALYSIS

To quantify the net computational gains, we conduct a standard FLOPs balance analysis. ConTra incurs an extra visual encoding cost, which is offset by the downstream token sparsification that shortens the sequence length for the LLM backbone. Formally, let N be the initial visual token

count, C the ViT channel dimension, D the LLM hidden dimension ($D \gg C$), and $\rho \in (0, 1]$ the token retention ratio. The net FLOPs saved, denoted as $\mathcal{F}_{\text{save}}$, is governed by:

$$\mathcal{F}_{\text{save}} = \underbrace{\mathcal{O}\big((1-\rho)ND^2\big)}_{\text{LLM Compute}} - \underbrace{\mathcal{O}\big(NC^2 + NCD + Nd\big)}_{\text{ConTra Overhead}}, \tag{4}$$

where the savings term represents the computation bypassed by the LLM self-attention and FFN layers, and the overhead term covers the dual-stream ViT encoding—executed as one batched call—the projection layer, and the consistency check.

In typical MLLMs, the LLM hidden dimension heavily dominates the vision channel width (e.g., $D = 4096$ vs. $C = 768$). Due to this structural asymmetry ($D \gg C$), the savings consistently outweigh the overhead. For instance, at a moderate retention ratio of $\rho = 0.5$, the computational savings scaling as $\mathcal{O}(0.5N \cdot D^2)$ outstrip the primary projection overhead component $\mathcal{O}(N \cdot C \cdot D)$ by a decisive margin:

$$\frac{(1-\rho)D^2}{NCD} \propto \frac{(1-\rho)D}{C} = \frac{0.5 \times 4096}{768} \approx 2.7\times . \tag{5}$$

This theoretical margin demonstrates that ConTra translates token pruning into substantial net FLOPs reduction. Detailed results regarding the actual runtime and GPU memory consumption are provided in the Appendix B.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate ConTra across two complementary task categories: text-rich image understanding and general-purpose vision-language tasks.

For **text-rich tasks**, we include 9 benchmarks covering OCR, document QA, and text-related visual reasoning: VQAText Wang & Ji (2018), OCRBench Liu et al. (2024), DocVQA Mathew et al. (2021a), CCOCR Yang et al. (2024c), OCRVQA Mishra et al. (2019), InfoVQA Mathew et al. (2021b), SynthDog Kim et al. (2022), WebSRC Chen et al. (2021), and MathVerse testmini vision only Zhang et al. (2024b).

For **general tasks**, we include 13 benchmarks covering non-textual reasoning and perception: AI2D Kembhavi et al. (2016), ChartQA Masry et al. (2022), GQA Hudson et al. (2019), POPE Li et al. (2023b), RealWorldQA Zhang et al. (2025a), ScienceQA Lu et al. (2022), MME Zhang et al. (2024d), Flickr30k Plummer et al. (2015), IllusionVQA Shahgir et al. (2024), VizWiz-VQA Gurari et al. (2018), VLMs-are-Biased Vo et al. (2026), and CV-Bench Team (2025).

Baselines. We compare against representative visual token pruning methods: VisionZip, DART, FastV, HoloV, and MMTok. These methods employ diverse compression strategies—including token merging, attention-based pruning, and hierarchical token retention—each navigating different trade-offs between performance and compression ratio.

Implementation Details. We use Qwen2.5VL-7B Bai et al. (2025b) as the primary model. To verify generalizability across model scales and architectures, we additionally evaluate on Qwen2.5VL-3B Bai et al. (2025b), and InternVL3-8B Zhu et al. (2025). All experiments are conducted on NVIDIA A100-40G GPUs using lmms-eval Zhang et al. (2024a) or VLMEvalKit Duan et al. (2024) for unsupported datasets.

4.2 Main Results

Text-rich Image Tasks. Table 1 compares ConTra with existing pruning methods on text-rich benchmarks under different pruning rates.

Key finding 1: Pruning can improve performance. Surprisingly, ConTra with moderate contrast reduction ($\alpha = 0.5$) at 50% pruning achieves 98.00% average retention—**exceeding the unpruned model on several benchmarks**. On TextCaps, ConTra ($\alpha = -1$) at 70% pruning scores 67.84 vs. 60.12 for the unpruned model (+12.8%). This counterintuitive result suggests that **removing**

Table 1: Comparison with visual token pruning methods on text-rich image benchmarks under different token pruning rates. **Bold** indicates the best result, and underline indicates the second best result.

Method	Text_{VQA}	$\text{Text}_{\text{Caps}}$	InfoVQA	DocVQA	OCRBench	WebSRC	SD_{en}	SD_{zh}	Math	Avg(%)
Qwen2.5	82.88	60.12	80.22	94.69	84.30	94.60	421	416	25.76	100
Token Pruning Rate = 50%										
VisionZip	72.98	56.42	55.19	77.88	68.10	62.80	282	229	19.04	67.95
DART	79.57	61.33	60.50	80.36	72.60	73.10	397	343	20.69	87.43
FastV	80.82	63.13	65.83	82.59	74.80	72.70	391	371	20.13	89.92
HoloV	77.40	59.31	50.48	71.11	75.81	64.20	316	272	20.30	74.07
MMTok	79.88	61.03	64.48	87.22	76.30	81.10	389	352	22.84	89.32
$\text{ConTra}_{0.5}$	81.78	62.73	74.91	92.47	**81.70**	**88.30**	**419**	406	24.37	97.96
ConTra_{-1}	**82.41**	**64.89**	**75.01**	**93.18**	**81.70**	87.90	**419**	**407**	**24.75**	**98.30**
Token Pruning Rate = 70%										
VisionZip	64.80	51.40	39.80	60.00	56.60	44.90	189	138	16.12	48.02
DART	72.75	58.03	44.67	62.83	63.20	50.90	327	254	19.04	69.87
FastV	75.46	63.42	47.73	56.51	58.00	50.60	280	244	17.26	65.66
HoloV	69.79	56.75	44.31	60.20	59.90	51.30	264	206	17.01	60.84
MMTok	76.25	59.92	58.05	81.00	69.60	68.50	360	320	20.18	81.92
$\text{ConTra}_{0.5}$	79.70	64.96	**63.79**	82.10	74.90	68.00	396	**359**	**22.97**	89.48
ConTra_{-1}	**81.16**	**67.84**	62.10	**83.82**	**77.40**	**68.60**	**399**	356	21.57	**90.05**
Token Pruning Rate = 90%										
VisionZip	46.10	40.03	28.90	33.23	36.10	29.00	92	48	9.64	26.50
DART	52.06	47.38	28.62	33.32	42.40	27.10	165	98	10.66	37.03
FastV	55.02	50.18	29.75	26.73	32.80	28.70	118	92	6.98	32.48
HoloV	47.50	44.39	31.01	34.73	36.70	28.80	138	86	9.64	33.53
MMTok	58.49	54.69	35.83	50.44	45.10	32.90	219	150	12.56	48.47
$\text{ConTra}_{0.5}$	70.40	65.42	**38.20**	49.30	**55.10**	**34.60**	244	177	**15.86**	55.04
ConTra_{-1}	**73.51**	**69.65**	36.09	**50.53**	53.50	31.70	**263**	**183**	14.72	**57.06**

redundant background tokens reduces visual interference, allowing the model to focus on text-critical regions more effectively.

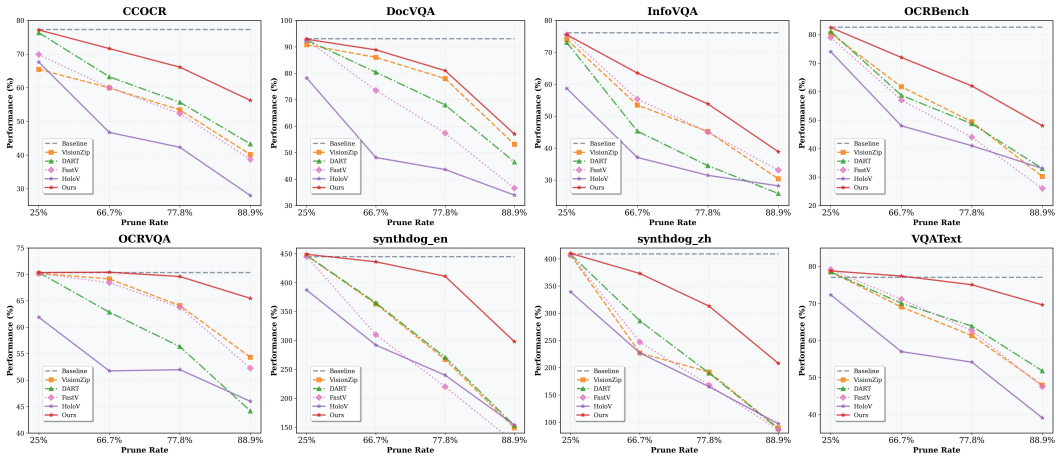

Figure 3: Comparison of Visual Token Pruning Methods across Diverse Pruning Rates on General Vision-Language Benchmarks using Qwen2.5-VL-3B

Key finding 2: Advantage widens with pruning rate. Table 1 reveals a clear trend: **ConTra's advantage over baselines increases as pruning becomes more aggressive**. At 50% pruning, ConTra

Table 2: Comparison of ConTra with different contrast factors α and MMTok on text-rich image benchmarks under 70% token pruning rate. **Bold**: best; underline: second best.

			ConTra					
Benchmark	**Qwen2.5**	**MMTok**	$\alpha=0.2$	$\alpha=0.5$	$\alpha=0.7$	$\alpha=1.2$	$\alpha=1.5$	$\alpha=-1$
TextVQA	82.88	76.25	79.28	79.70	80.09	78.80	79.07	**81.16**
TextCaps	60.12	59.92	63.16	64.96	64.29	64.80	65.20	**67.84**
InfoVQA	80.22	58.05	64.38	63.79	62.33	62.83	63.53	62.10
DocVQA	94.69	81.00	82.26	82.10	81.55	69.87	72.69	**83.82**
OCRBench	84.30	69.60	76.70	74.90	75.50	75.70	77.10	**77.40**
WebSRC	94.60	68.50	72.20	68.00	64.60	65.50	70.40	68.60
SD_{en}	421	360	390	396	391	387	391	**399**
SD_{zh}	416	320	355	**359**	352	348	342	356
MathVerse	25.76	20.18	21.83	22.97	**23.60**	21.70	21.57	21.57
Avg	100.00%	81.94%	88.65%	89.14%	87.93%	86.40%	87.02%	**89.59%**

($\alpha=-1$) leads FastV by 8.19% (98.31% vs. 90.12%). At 70%, the gap widens to 24.4% over FastV (90.05% vs. 65.66%) and 8.1% over MMTok (90.05% vs. 81.92%). At the extreme 90% pruning, ConTra maintains 57.06%—an 8.6% absolute improvement over MMTok (48.47%) and 24.6% over the third-best method.

Why does the advantage widen? We attribute this trend to the compounding effect of selection accuracy. At low pruning rates, even suboptimal methods retain enough tokens to preserve most information. As pruning increases, the cumulative error from incorrect token selection becomes critical. ConTra's consistency-based selection maintains higher accuracy at each selection step, leading to better preservation of text-critical tokens under severe compression.

On MathVerse, $\alpha = 0.5$ consistently outperforms $\alpha = -1$ by +1.53 to +3.30 over the best baseline MMTok, suggesting that mathematical diagrams require more conservative perturbation.

General Image Tasks. ConTra incurs no degradation on largely text-free general benchmarks: 100.46% average retention at 50% pruning (vs. 100.00% for query-guided MMTok), ranking first at 90% with 87.01%. Margins are inherently narrow here—redundant appearance tokens dominate—and per-benchmark results are in Appendix A.

4.3 Analysis

From Text to Natural Images: ConTra as a Structure-Preserving Selector. Why should a text-motivated selector excel on largely text-free images? Because Ordinal Structure Invariance (Property 1) is not specific to glyphs: *structure-carrying tokens* (contours, shapes, diagram geometry) encode semantics relationally and retain high consistency scores under CCT, while *appearance-carrying, redundant tokens* (uniform color fields, illumination-dominated surfaces) drift and are pruned—precisely the informative/redundant partition any token reduction method seeks. The only tokens at risk are appearance-carrying yet semantically relevant ones (e.g., query-relevant colors), which predicts that strong perturbations should hurt precisely on general tasks—as observed: $\alpha=0.5$ surpasses the unpruned model (100.46% retention) while $\alpha = -1$ drops to 98.59% (Appendix A), with the same ordering on MathVerse (Table 2). The same principle explains the margin asymmetry: answer-relevant semantics in natural images concentrate in few structure-carrying tokens, so all reasonable selectors cluster within a narrow band, whereas densely packed glyph tokens make text-rich tasks the decisive regime. Consistency-score maps (Appendix A) confirm that high scores concentrate on object contours rather than residual text.

Insight: Text-rich tasks are where contrast-invariant selection is decisive; on general imagery the same principle acts as a safety guarantee, matching query-guided methods without any query information.

Similarity Thresholds and Text Density. Figure 4 evaluates ConTra under varying similarity thresholds with $\alpha = -1$. A key finding emerges: **high text-density benchmarks are far more robust to pruning than low text-density ones**. At $\tau = 0.8$, high text-density benchmarks retain

99% performance with only 37% tokens pruned, whereas low text-density benchmarks require 63% pruning to achieve the same retention. At the extreme $\tau = 0.95$, high text-density benchmarks still preserve 87% performance (64% tokens pruned) vs. 86% (83% pruned) for low text-density.

This adaptive behavior directly reflects the Contrast Invariance Principle: text-critical tokens naturally score higher consistency and are preferentially retained, enabling favorable accuracy–efficiency trade-offs with TFLOPs reduced by up to 63%.

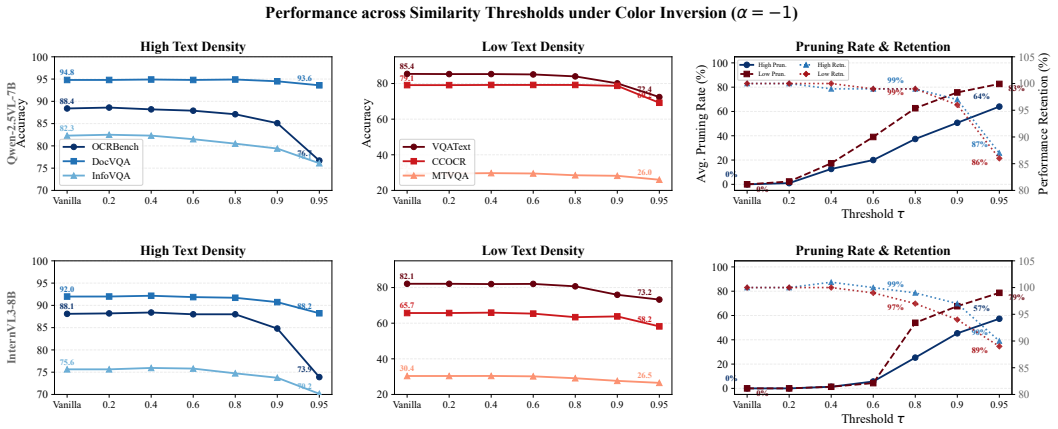

Figure 4: Performance of ConTra (α=-1) across varying similarity thresholds τ on Qwen-2.5VL-7B (top) and InternVL3-8B (bottom).

Insight: The instance-dependent sparsification automatically adapts to image content, making ConTra practical for real-world deployment with varying text densities.

Impact of Contrast Factor α. Table 2 compares different α values at 70% pruning on text-rich benchmarks. All ConTra variants outperform MMTok (81.92%), with retention ranging from 86.40% to 89.59%; color inversion ($\alpha = -1$) performs best (89.59%), confirming that stronger perturbations widen the consistency gap between text and background tokens. On MathVerse and general benchmarks (Appendix A), milder perturbations ($\alpha = 0.5$) instead rank first, since color inversion also corrupts appearance-carrying semantics (e.g., diagram colors) that these tasks still rely on. **Insight:** use $\alpha = -1$ for text-rich scenarios and $\alpha = 0.5$ as a robust default for mixed workloads.

Cross-Model Generalization. On Qwen2.5VL-3B and InternVL3-8B Chen et al. (2024b), whose visual encoders differ markedly from Qwen2.5-VL-7B, ConTra maintains its advantage: ∼90% retention even at 88.9% pruning on VQAText and OCRVQA (Figure 3), and on InternVL3-8B over 90% retention with 80%+ pruning on low text-density images and 99% with 40% pruning on high-density ones (Figure 4). **Insight:** the Contrast Invariance Principle reflects a robust property of pretrained multimodal representations across architectures and scales, not an artifact of a specific model.

5 Conclusion

We presented ConTra, a training-free visual token compression method leveraging contrast transformation consistency to preserve text-critical tokens. By exploiting textual semantic invariance under chromatic contrast modulation, ConTra decouples textual from visual semantics without attention or similarity heuristics. ConTra achieves state-of-the-art performance on text-rich benchmarks—where selection quality is decisive—and remains fully competitive with query-guided methods on general benchmarks while using no query information; being model-agnostic and query-agnostic, it seamlessly integrates into existing MLLMs.

AI use statement

In this work, generative AI tools were used solely for grammar checking and language polishing of the manuscript. They were not used for research ideation, problem formulation, method or

experimental design, code implementation, data analysis, or figure generation; all research ideas, methodology, experiments, analyses, and conclusions are the original work of the authors. Every AI-polished passage was carefully reviewed and, where necessary, further edited by the authors to ensure technical accuracy and fidelity to the intended meaning. We take full responsibility for the final content of this work, including any text refined with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects, personally identifiable information, or new data collection: all experiments are conducted on publicly available benchmarks and publicly released pre-trained models, used in accordance with their respective licenses and terms of use. By reducing the inference cost of multimodal large language models, ConTra can lower the computational and energy footprint of deploying such models, which we view as a positive broader impact. As with any token compression technique, aggressive pruning may discard information, so deployments in high-stakes domains (e.g., medical or legal document analysis) should be validated before practical use. We do not foresee other direct ethical risks, harmful applications, or conflicts of interest arising from this work.

REPRODUCIBILITY STATEMENT

ConTra is training-free, so our results can be reproduced without any training pipeline: the entire procedure is fully specified by the CCT transformation (Eq. equation 1, Section 3.1) and the cross-view consistency score with its two token selection strategies (Section 3.2), while Property 1 follows directly from the linearity of CCT. All hyperparameters are reported in the paper, including the contrast scale α (-1 for text-rich tasks and 0.5 for general tasks; Table 2), as well as the retention ratio ρ and the similarity threshold τ for the two selection strategies. All evaluated models (Qwen2.5VL-3B/7B and InternVL3-8B) and all benchmarks are publicly available.

REFERENCES

Shuai Bai, Yuxuan Cai, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. 1, 3

Shuai Bai et al. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923. 3, 6

Daniel Bolya, , et al. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. 1, 3

Tom Brown et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. 1

Qingqing Cao et al. PuMer: Pruning and merging tokens for efficient vision language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12890–12903, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.721. 3

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, 2024a. 1, 3

Xingyu Chen et al. WebSRC: A dataset for web-based structural reading comprehension. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4173–4185, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.343. URL https://aclanthology.org/2021.emnlp-main.343/. 6

Zhe Chen et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b. 9

Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019. URL https://arxiv.org/abs/1805.09501. 4

Sixun Dong et al. MMTok: Multimodal coverage maximization for efficient inference of VLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GvPdSWZT31. 1, 3

Haodong Duan et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024. 6

Danna Gurari et al. Vizwiz grand challenge: Answering visual questions from blind people, 2018. URL https://arxiv.org/abs/1802.08218. 6

Drew A Hudson et al. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

Kembhavi et al. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016. 6

Geewook Kim et al. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 6

{Gordon E} Legge, {Gary S.} Rubin, and Andrew Luebker. Psychophysics of reading-v. the role of contrast in normal vision. *Vision Research*, 27(7):1165–1177, 1987. ISSN 0042-6989. doi: 10.1016/0042-6989(87)90028-9. Funding Information: Acknowledgements-We would like to thankD enisPelli for helpingu s with the schemef or contrasta ttenuationH, olli Rietmulderf or help with collectiona nd ˜b˜tion of data, and Mary Schleskef or preparationo f the figures. Preliminaryr esultsw erep resenteda t theA nnual Meetingo f the Optical Societyo f America,S an Diego, 1984T. he research was supportedb y PHS Grant EY02934. 4

Yanwei Li et al. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*, 2023a. 1

Yifan Li et al. Evaluating object hallucination in large vision-language models, 2023b. URL https://arxiv.org/abs/2305.10355. 6

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6. 6

Pan Lu et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 6

Ahmed Masry et al. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177. 6

Minesh Mathew et al. Docvqa: A dataset for vqa on document images, 2021a. URL https://arxiv.org/abs/2007.00398. 6

Minesh Mathew et al. Infographicvqa, 2021b. URL https://arxiv.org/abs/2104.12756. 6

Mehrani et al. Self-attention in vision transformers performs perceptual grouping, not attention. *Frontiers in Computer Science*, 5:1178450, jun 2023. doi: 10.3389/fcomp.2023.1178450. URL https://doi.org/10.3389/fcomp.2023.1178450. 2

Anand Mishra et al. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 6

Thao Nguyen et al. Improving multimodal datasets with image captioning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 22047–22069. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/45e604a3e33d10fba508e755faa72345-Paper-Datasets_and_Benchmarks.pdf. 1

Bryan A. Plummer et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 6

Radford et al. Language models are unsupervised multitask learners. 2019. 1

Haz Sameen Shahgir et al. IllusionVQA: A challenging optical illusion dataset for vision language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=7ysaJGs7zY. 6

CVBench Team. Cvbench: A benchmark for cross-video multimodal reasoning, 2025. URL https://huggingface.co/datasets/Dongyh35/CVBench. 6

An Vo et al. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2505.23941. 6

Hanzhang Wang and Qingyuan Ma. Textural or textual: How vision-language models read text in images. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=fUvuEfZSEE. 2

Zhengyang Wang and Shuiwang Ji. Learning convolutional text representations for visual question answering. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pp. 594–602. SIAM, 2018. 6

Zichen Wen, Yifeng Gao, et al. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9961–9980, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.505. URL https://aclanthology.org/2025.emnlp-main.505/. 1, 3

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. *arXiv preprint arXiv:2412.04467*, 2024a. 1, 3

Zhibo Yang, Jun Tang, et al. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy, 2024b. URL https://arxiv.org/abs/2412.02210. 1

Zhibo Yang et al. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy, 2024c. URL https://arxiv.org/abs/2412.02210. 6, 15

Kaichen Zhang et al. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024a. URL https://arxiv.org/abs/2407.12772. 6

Renrui Zhang et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024b. 6

Yi-Fan Zhang et al. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024c. 1

Yi-Fan Zhang et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024d. 6

YiFan Zhang et al. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=k5VHHgsRbi. 6

Yuan Zhang et al. Unveiling the tapestry of consistency in large vision-language models. *arXiv preprint arXiv:2405.14156*, 2024e. 1

Yuan Zhang et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning*, 2025b. 1, 3

Jinguo Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479. 6

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention, 2025. URL https://arxiv.org/abs/2510.02912. 1, 3

Table 3: Comparison with visual token pruning methods on general vision-language benchmarks under different token pruning rate. **Bold** indicates the best result, and underline indicates the second best result.

Method	AI2D	ChartQA	GQA	POPE	RWQA	SQA	MME	Flickr30k	IllVQA	VizWiz	VLMs-Bias	Avg
Qwen2.5-VL-7B	82.48	84.12	60.85	86.41	69.15	88.10	2313	33.18	46.20	70.87	22.91	100.00%
Token Pruning Rate = 50%												
VisionZip	76.97	61.80	57.16	81.13	62.48	85.07	2276	30.49	45.28	68.99	19.82	96.90%
DART	78.17	70.48	60.00	85.89	67.05	86.51	2277	33.23	44.59	70.14	22.30	97.92%
FastV	78.88	74.84	55.24	82.99	66.40	**86.76**	2250	**34.49**	44.36	69.85	21.65	96.91%
HoloV	78.98	60.44	59.41	84.95	66.66	86.06	2225	33.04	44.36	**70.53**	22.34	95.77%
MMTok	81.89	**76.88**	60.09	85.23	65.49	86.01	2330	33.53	**45.97**	70.26	21.55	100.00%
ConTra (α=0.5)	**81.96**	76.56	**60.11**	**86.13**	**68.75**	86.56	**2339**	33.45	45.05	69.72	**23.06**	**100.46%**
ConTra (α=-1)	81.05	73.96	59.16	84.72	68.62	85.82	2291	33.18	45.05	69.87	22.59	98.59%
Token Pruning Rate = 70%												
VisionZip	74.48	51.00	55.50	77.30	58.16	84.00	2155	28.60	43.68	69.64	20.51	91.92%
DART	74.09	53.80	58.49	84.39	65.75	83.73	2217	31.41	43.44	68.32	20.65	94.73%
FastV	74.35	52.72	51.12	77.01	62.87	84.38	2125	32.28	44.13	68.25	19.10	91.02%
HoloV	76.30	47.84	58.10	83.13	65.49	84.92	2169	32.51	42.30	69.60	20.50	93.00%
MMTok	**80.21**	**69.16**	**59.25**	**86.34**	63.39	84.87	2284	**33.67**	43.44	**69.81**	20.33	97.89%
ConTra (α=0.5)	79.95	67.50	57.60	86.10	**68.60**	**84.97**	**2289**	31.70	42.52	68.83	**22.91**	**98.07%**
ConTra (α=-1)	77.80	65.84	56.10	83.20	65.50	82.30	2241	30.60	**44.82**	68.00	22.12	95.96%
Token Pruning Rate = 90%												
VisionZip	69.81	32.08	52.92	73.60	56.60	79.47	1983	26.91	**45.05**	64.73	17.06	84.59%
DART	67.38	32.00	51.59	73.23	57.12	**81.01**	2033	25.23	43.90	64.31	18.17	86.14%
FastV	65.96	23.32	46.24	67.26	49.67	76.89	1888	24.02	43.21	61.91	17.34	79.95%
HoloV	67.90	29.44	53.18	76.52	59.86	80.76	1915	27.66	42.75	66.96	18.92	82.49%
MMTok	71.21	**38.96**	**54.76**	**77.37**	58.69	80.21	1997	**29.75**	42.52	**67.24**	19.75	85.82%
ConTra (α=0.5)	**71.40**	38.08	50.10	76.08	**64.31**	77.93	**2043**	24.55	41.60	64.15	**21.51**	**87.01%**
ConTra (α=-1)	69.00	37.88	46.87	63.81	56.73	78.33	1920	19.74	42.52	61.57	19.97	81.72%

A General Image Tasks

Table 3 evaluates ConTra on general vision-language benchmarks under 50%, 70%, and 90% pruning rates. ConTra ($\alpha = 0.5$) achieves the best average retention across all pruning rates: 100.5% at 50%, 98.1% at 70%, and 87.0% at 90%, consistently outperforming the strongest baseline MMTok (100.0%, 97.9%, 85.8%). Notably, ConTra even slightly exceeds the unpruned model on several benchmarks (e.g., RWQA and VLMs-Bias at 90% pruning), suggesting that removing redundant tokens can reduce interference. Interestingly, $\alpha = 0.5$ consistently outperforms $\alpha = -1$ on general benchmarks—the opposite trend to text-rich tasks. This is because structural visual semantics (e.g., edges, shapes) in general images are only partially decoupled from visual appearance; color inversion ($\alpha = -1$) is too aggressive and disrupts these partially decoupled semantics, while moderate contrast reduction ($\alpha = 0.5$) preserves them while still pruning redundant tokens (see Section 4.3 for detailed analysis).

B Computational Efficiency

We measure actual FLOPs on Qwen2.5VL-7B to validate our theoretical analysis (Section 3.4). The unpruned model incurs 11.23 TFLOPs in the LLM versus 0.19 TFLOPs in the ViT. At 51% pruning, total FLOPs drop to 5.89 (99% performance retained); at 74% pruning, FLOPs drop to 3.30 (95% performance). The CCT encoding overhead (0.19 TFLOPs) is far outweighed by the LLM savings. We also measure actual inference latency and peak GPU memory on TextVQA; despite the additional ViT pass, ConTra's overhead remains negligible.

We have supplemented the actual running time and peak GPU memory usage on the textVQA dataset, under a pruning ratio of 0.889, using a machine equipped with an NVIDIA A100 40GB GPU. As shown in the table 4, although our method requires introducing an additional image, the computational overhead is negligible. In contrast, VisionZip is incompatible with FlashAttention, resulting in substantially higher inference latency and GPU memory consumption.

C Empirical Verification of Statistical Separability

This appendix provides an empirical verification of the statistical separability guarantee underlying ConTra's token selection mechanism. Specifically, we verify that the per-token consistency score

Table 4: Actual runtime and memory usage compared to other pruning methods

Method	Time (ms)	Memory (GB)
ConTra	459	1.51
FastV	1950	1.51
HoloV	3440	1.54
Visionzip	3780	5.60

distribution is bimodal and that the two components (text and background) are sufficiently separated for Top-K selection to precisely isolate text-critical tokens.

C.1 Experimental Setup

Dataset. We randomly sample 30 text-rich images from the CC-OCR Yang et al. (2024c) benchmark, covering multiple languages (Arabic, Chinese, English, etc.) and diverse visual styles (documents, product labels, signs, handwritten text). This diverse sample ensures that our verification is not biased toward a specific text type or language.

Model. We use Qwen2.5VL-7B-Instruct as the vision-language model. For each image, we extract visual tokens after the merger module, i.e., the projection layer that maps visual features into the LLM embedding space where semantic anchoring operates (Section 3.3); this produces a 2D grid of token embeddings after 2×2 spatial merging.

Consistency Score Computation. For each image $\mathbf{I}^{\text{ori}}$, we apply the chromatic contrast transformation with $\alpha = -1$, i.e., the strict color inversion $\mathbf{I}^{\text{ct}} = \text{CCT}_{-1}(\mathbf{I}^{\text{ori}})$ (Eq. equation 3). We then extract visual tokens $\{\mathbf{h}_p^{\text{ori}}\}_{p=1}^N$ and $\{\mathbf{h}_p^{\text{ct}}\}_{p=1}^N$ from the two views, and compute the per-token consistency score as in Section 3.2:

$$s_p = \frac{(\mathbf{h}_p^{\text{ori}})^\top \mathbf{h}_p^{\text{ct}}}{\|\mathbf{h}_p^{\text{ori}}\| \cdot \|\mathbf{h}_p^{\text{ct}}\|}. \tag{6}$$

GMM Fitting. For each image, we fit a two-component Gaussian Mixture Model (GMM) to the score distribution $\{s_p\}_{p=1}^N$:

$$p(s) = \pi_t \mathcal{N}(s \mid \mu_t, \sigma_t^2) + (1 - \pi_t)\mathcal{N}(s \mid \mu_b, \sigma_b^2), \tag{7}$$

where the component with the higher mean is designated as the *text* component ($\mu_t > \mu_b$). We use 10 random restarts and select the best-fit solution by log-likelihood.

Separability Metrics. We compute the following metrics to quantify the degree of separation:

- **Ashman's D ?**: Measures the separation between two Gaussian components:
$$D = \frac{|\mu_t - \mu_b|}{\sqrt{(\sigma_t^2 + \sigma_b^2)/2}}. \tag{8}$$
$D > 2$ is the conventional criterion for a clean separation of the two components.
- **Bimodality Coefficient (BC) ?**: Based on skewness γ_1 and kurtosis κ:
$$\text{BC} = \frac{\gamma_1^2 + 1}{\kappa + 1}. \tag{9}$$
$\text{BC} > 0.555$ indicates a bimodal distribution.
- **Overlap Ratio**: The standard overlapping coefficient of the two weighted GMM components:
$$\text{Overlap} = \int_{\mathbb{R}} \min\big(\pi_t \mathcal{N}(s \mid \mu_t, \sigma_t^2),\ (1-\pi_t)\mathcal{N}(s \mid \mu_b, \sigma_b^2)\big)\, ds, \tag{10}$$
which equals the proportion of the mixture distribution where the two components coincide, since the mixture density integrates to one. Lower overlap indicates better separability.

C.2 RESULTS

Overall Separability. Table 5 summarizes the GMM parameters and separability metrics across all 30 images. The mean Ashman's D is **2.53** (median 2.46), with **96.7%** of images exceeding the $D > 2$ separation threshold. The text component consistently centers at $\mu_t = 0.868 \pm 0.093$, while the background component centers at $\mu_b = 0.493 \pm 0.086$, yielding a mean separation of $\Delta\mu = 0.375$. The mean overlap ratio is only **6.2%**, confirming that the two components are largely disjoint. On average, the bimodality coefficient ($\mathrm{BC} = 0.613$) exceeds the 0.555 threshold, further confirming bimodality.

Table 5: GMM parameters and separability metrics across 30 text-rich images from CC-OCR. Values are reported as mean ± std.

Metric	Mean ± Std	Range
GMM Component Parameters		
Text mean (μ_t)	0.868 ± 0.093	[0.67, 0.99]
Background mean (μ_b)	0.493 ± 0.086	[0.31, 0.72]
Text std (σ_t)	0.062 ± 0.028	[0.02, 0.14]
Background std (σ_b)	0.091 ± 0.035	[0.03, 0.19]
Text ratio (π_t)	0.481 ± 0.116	[0.25, 0.73]
Separability Metrics		
Ashman's D	2.527 ± 0.341	[1.94, 3.19]
Bimodality Coeff. (BC)	0.613 ± 0.082	[0.45, 0.78]
Overlap Ratio	0.062 ± 0.038	[0.01, 0.15]
Ratio of $D > 2$	29 / 30 (96.7%)	

Per-Image Score Distributions. Figure 5 shows the consistency score distributions and GMM fits for six representative images. In each subplot, the gray histogram represents the empirical score distribution, the red dashed curve is the text component $\mathcal{N}(\mu_t, \sigma_t^2)$, the blue dashed curve is the background component $\mathcal{N}(\mu_b, \sigma_b^2)$, and the black solid line is the full GMM fit. The green dotted vertical line marks the midpoint $(\mu_t + \mu_b)/2$, which serves as the optimal classification threshold. Across all images, the two components are clearly separated, with text tokens clustering near $s \approx 0.9$ and background tokens clustering near $s \approx 0.5$.

Ashman's D Distribution. Figure 6 presents the distribution of Ashman's D across all 30 images. The left panel shows the histogram of D values with the $D = 2$ threshold marked. The right panel shows a sorted bar chart, where green bars indicate $D > 2$ (separable) and red bars indicate $D \leq 2$. Only one image falls below the threshold ($D = 1.94$), while the remaining 29 images exhibit $D > 2$, with the maximum reaching $D = 3.19$.

GMM Parameter Relationships. Figure 7 examines the relationships between GMM parameters. The left panel plots μ_t against μ_b for each image; points far from the diagonal indicate strong separation. The right panel plots the text token ratio π_t against Ashman's D, showing that separability is robust across different text-to-background ratios (ranging from 0.25 to 0.73), confirming that the bimodality is not an artifact of a specific text density.

C.3 DISCUSSION

Implications for Top-K Selection. The near-perfect separation ($D = 2.53$, overlap $= 6.2\%$) has a direct implication for ConTra's Top-K selection: when the retention ratio ρ is set close to the true text token ratio π_t, the Top-K operation achieves near-optimal classification accuracy. Even when ρ deviates from π_t, the low overlap ratio ensures that the misclassification rate remains bounded. This provides an empirical foundation for the statistical separability argument in Section 3.3.

Robustness Across Languages. Our sample includes images with Arabic, Chinese, and English text, spanning different scripts, layouts, and visual complexities. The consistency of the bimodality across these diverse images (29 out of 30 with $D > 2$) suggests that the separability guarantee is

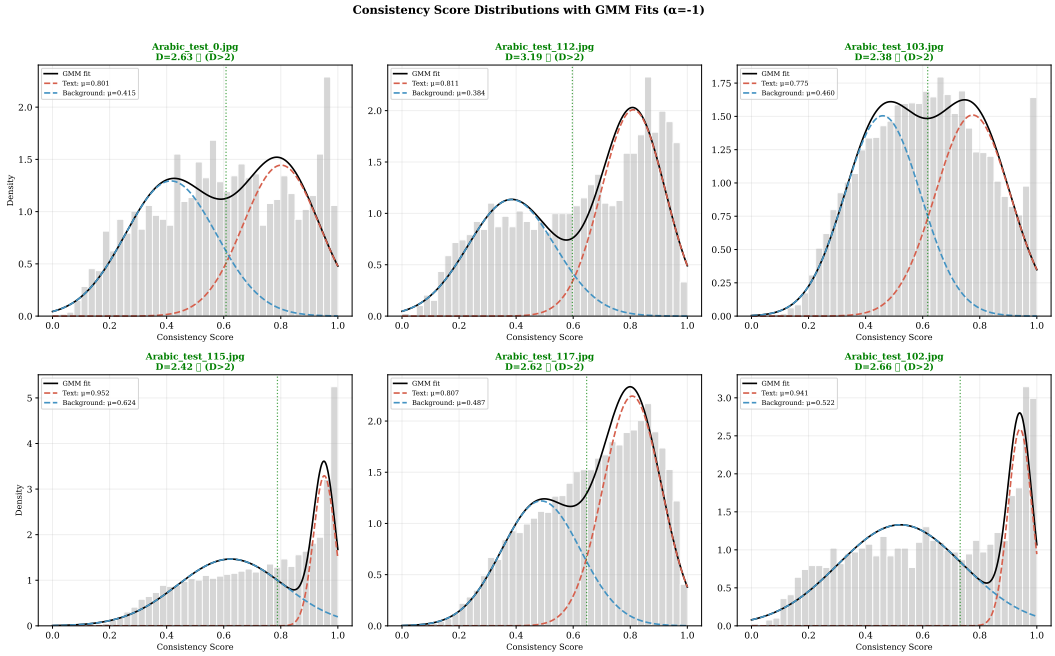

Figure 5: Consistency score distributions with two-component GMM fits for six representative images ($\alpha = -1$). Each subplot shows the empirical histogram (gray), the text component (red dashed), the background component (blue dashed), and the full GMM fit (black). The green dotted line marks the optimal separation threshold. Ashman's D values are shown in each title; green titles indicate $D > 2$.

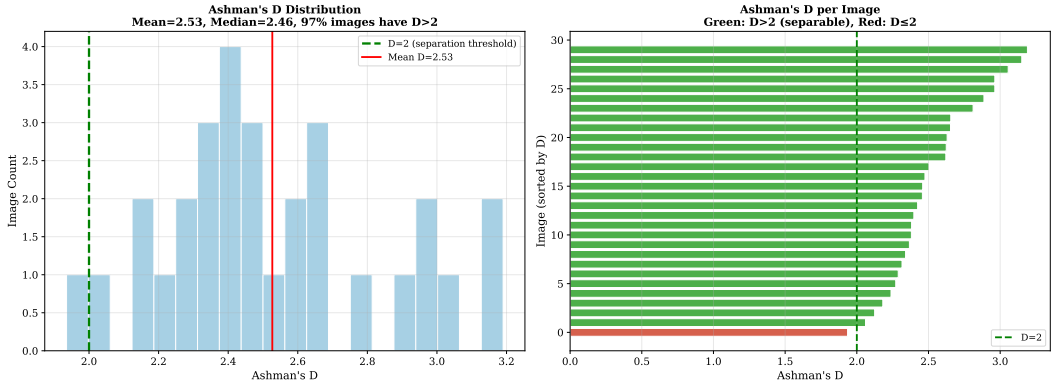

Figure 6: Distribution of Ashman's D across 30 images. **Left:** Histogram of D values with the $D = 2$ separation threshold (green dashed) and mean (red solid). **Right:** Sorted bar chart; green bars indicate $D > 2$ (separable), red bars indicate $D \leq 2$.

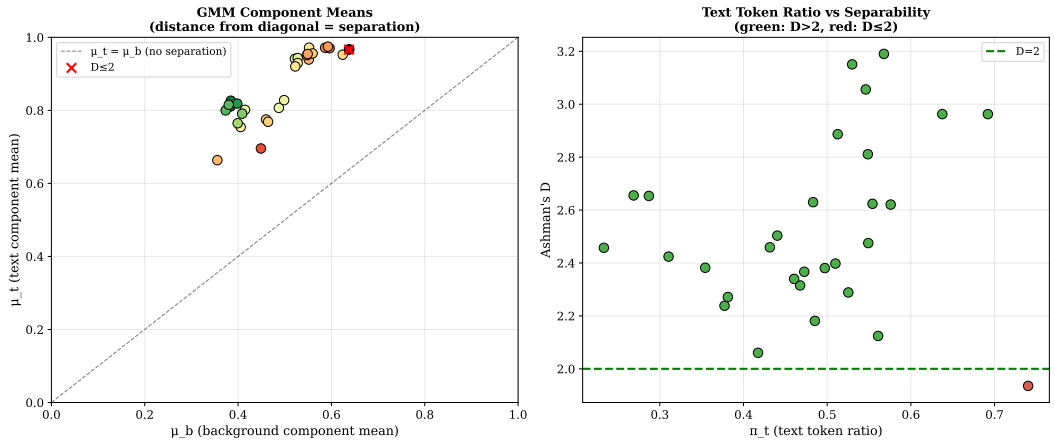

Figure 7: GMM parameter relationships across 30 images. **Left:** Text component mean μ_t vs. background component mean μ_b; the dashed diagonal represents no separation. Points are colored by Ashman's D; red crosses mark images with $D \leq 2$. **Right:** Text token ratio π_t vs. Ashman's D; the green dashed line marks $D = 2$.

language-agnostic, as expected from the anchoring argument: the key factor is the semantic invariance of text under chromatic transformation, which holds regardless of the script or language.

The Single Failure Case. One image yielded $D = 1.94$, slightly below the threshold. Inspection reveals that this image contains a large uniform background region with very few text tokens ($\pi_t = 0.25$), causing the GMM to fit a broad background component that partially overlaps with the text component.

Conclusion. The empirical results confirm that the consistency score distribution is bimodal with Ashman's D > 2 in 96.7% of text-rich images, with a mean overlap of only 6.2%. This validates the statistical separability guarantee underlying ConTra's token selection mechanism: the consistency gap between text and background tokens is not a coincidental empirical observation but a structural consequence of the semantics-preserving perturbation and the projection layer's anchoring effect.