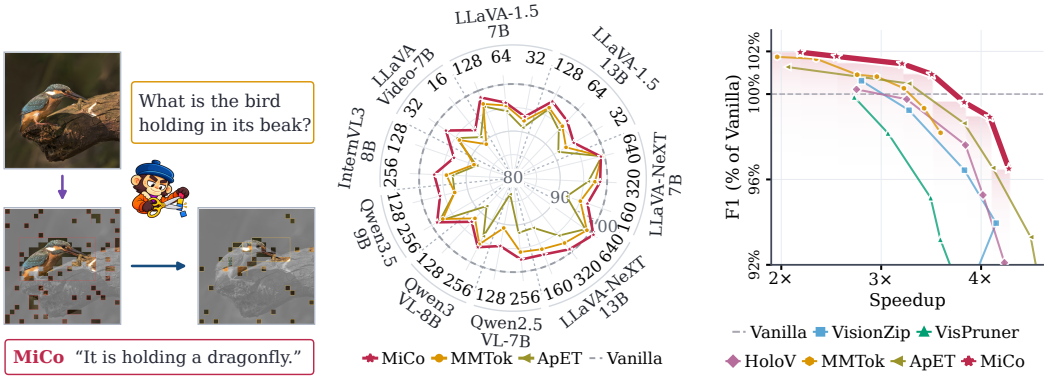
MiCo: Mutual Information Coverage Optimization through Semantic Erasure Modeling for Efficient MLLM Inference

Anonymous authors
Paper under double-blind review

Abstract

Multimodal large language models (MLLMs) have demonstrated impressive performance in multimodal understanding, but processing large numbers of visual tokens results in high computational costs. While many methods have been proposed to reduce the number of visual tokens, most of them rely on heuristics and are prone to discarding substantial visual information during pruning, leading to degradation in model performance. In this work, by using a semantic erasure model, we derive a general mutual information coverage objective from task log-loss and propose MiCo, a training-free two-stage pruning method. MiCo first uses visual signals to select a representative candidate pool before visual tokens enter the language model, then performs task-aware subset selection within it. At each stage, suitable observable proxies instantiate the derived objective as a monotone submodular coverage function, which MiCo greedily optimizes under the token budget. MiCo is evaluated on diverse MLLMs ranging from 7B to 13B parameters across a broad range of image and video benchmarks spanning general visual reasoning, fine-grained OCR and grounding, hallucination detection, and long-video understanding. MiCo consistently achieves the best performance across nearly all evaluated models under all pruning ratios. On LLaVA-NEXT-13B, MiCo uses only 5.6% visual tokens, retains 97.5% of baseline performance, and achieves a 3.8× inference speedup. Our experiments demonstrate the effectiveness of MiCo and our mutual information coverage objective for visual token pruning.



(a) MiCo's pruning **(b)** Cross-model performance **(c)** Quality–speed trade-off

Figure 1: MiCo: (a) two-stage pruning on LLaVA-1.5-7B at $T = 64$, retaining $576 \to 128 \to 46$ visual tokens for "What is the bird holding in its beak?"; (b) relative performance across eight image models and one video model in 22 model–budget settings, normalized to each model's unpruned reference (video budgets are per frame); (c) POPE F1 retention relative to Vanilla versus one-token inference speedup on LLaVA-NeXT-13B.

1 Introduction

Large language models (LLMs) have demonstrated strong capabilities in language understanding and reasoning (Achiam et al., 2023; Grattafiori et al., 2024; Team et al., 2023; Yang et al., 2025a).

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.

By integrating vision encoders with LLMs, multimodal large language models (MLLMs) leverage these capabilities to understand visual content (Liu et al., 2023a; Bai et al., 2023; Liu et al., 2024a;b; Bai et al., 2025a; Zhu et al., 2025). They convert visual inputs into token sequences that the LLM processes jointly with text. However, a single image can produce hundreds to thousands of visual tokens, and the number grows further with image resolution, the number of input images, and video frame count. Combined with the quadratic cost of self-attention, these long sequences dominate computation and memory, so reducing visual tokens while preserving model capabilities is a central challenge.

Many methods prune visual tokens to address this problem. Importance-ranking methods score tokens individually: FastV and SparseVLM use text-to-visual attention inside the language model, whereas PruMerge and VisionZip use CLS-token attention from the vision encoder (Chen et al., 2024a; Zhang et al., 2024b; Shang et al., 2025; Yang et al., 2025b). Subset-construction methods instead build a representative token subset: DivPrune and DART use diversity criteria to reduce redundancy, while MMTok and CoverPruner use coverage criteria to preserve multimodal and visual information (Alvar et al., 2025; Wen et al., 2025; Dong et al., 2026; Zhu et al., 2026). However, attention-based ranking often retains semantically redundant tokens, whereas the tokens kept by subset construction may be irrelevant to the current query. Even methods that combine the two paradigms (Zou et al., 2025; Deng et al., 2025) are motivated by heuristics rather than an analysis of the visual information the model needs, and degrade substantially at high pruning ratios.

To preserve the visual information that inference requires, we analyze how removing visual tokens changes the expected task log-loss and, through a semantic erasure model, derive a mutual information coverage objective. Its theoretical factors are not observable in a single forward pass, so we instantiate them with observable proxies and propose MiCo, a training-free two-stage method that greedily optimizes the resulting surrogate at each stage (Figure 1(a)); since the surrogate is monotone submodular, greedy selection attains at least a $(1 - 1/e)$ fraction of its optimum (Nemhauser et al., 1978). Across MLLMs of diverse architectures and a broad range of image and video benchmarks, MiCo achieves the strongest performance among compared methods on nearly all models at all pruning ratios (Figure 1(b)) while substantially accelerating inference (Figure 1(c)).

Our main contributions are as follows:

- We analyze the visual information required for inference from the expected task log-loss and, via a semantic erasure model, derive a general mutual information coverage objective for visual token pruning.
- We instantiate the objective with accessible proxies, obtaining a tractable submodular surrogate, and propose MiCo, a training-free method that optimizes it by greedy selection.
- Extensive experiments across MLLMs, benchmarks, and pruning ratios show that MiCo consistently preserves performance while substantially accelerating inference.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) (Liu et al., 2023a; Bai et al., 2023; Chen et al., 2024c) encode images and videos into visual tokens that the LLM processes jointly with text. Visual tokens usually far outnumber text tokens: LLaVA-1.5 (Liu et al., 2024a) uses 576 tokens for a 336×336 image, LLaVA-NeXT (Liu et al., 2024b) up to 2,880 at 672×672, dynamic-resolution models such as Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), and InternVL3 (Zhu et al., 2025) scale token counts with image size, and LLaVA-Video (Zhang et al., 2024c) exceeds 10K tokens for 64 frames. These long sequences dominate prefill computation and memory, making visual token reduction essential for efficient MLLM inference.

2.2 Visual Token Reduction for MLLMs

Training-free pruning methods fall into two categories. The first ranks individual tokens by attention. FastV (Chen et al., 2024a), PyramidDrop (Xing et al., 2024), and SparseVLM (Zhang et al., 2024b) use the attention that visual tokens receive from text inside the LLM, whereas PruMerge (Shang et al., 2025), VisionZip (Yang et al., 2025b), and VisPruner (Zhang et al., 2025a) use CLS-to-patch

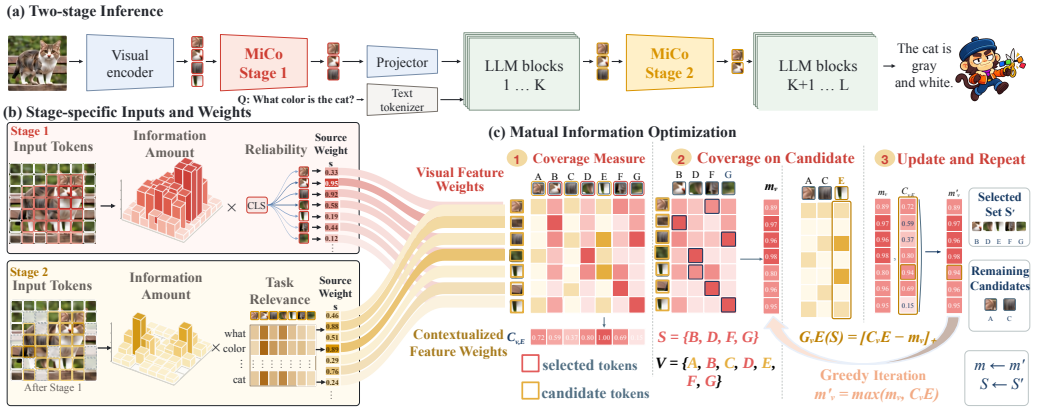


Figure 2: Overview of MiCo's two-stage pruning framework.

attention from the visual encoder; VScan (Zhang et al., 2026) combines both attention sources. The second constructs a token subset. Diversity-based methods, including DivPrune (Alvar et al., 2025), DART (Wen et al., 2025), and CDPruner (Zhang et al., 2025b), construct mutually distinct subsets to reduce redundancy, while coverage-based methods seek subsets that represent the original visual or multimodal token set. SCOPE, MMTok, CoverPruner, and ApET instantiate different forms of coverage or representative-set construction (Deng et al., 2025; Dong et al., 2026; Zhu et al., 2026; Ma et al., 2026). Attention-based selectors can retain redundant tokens from one semantic region, whereas set-level objectives without query conditioning may spend budget on task-irrelevant regions. Moving beyond such heuristic criteria, **MiCo** derives a weighted mutual information coverage objective and optimizes its observable proxy.

3 Method

3.1 From Task Loss to Recoverable Coverage

Task loss and recoverable information. Let π_{data} be the true joint distribution of visual information Z, prompt Q, shared geometric context G (the image layout and token positions, which pruning leaves unchanged), and discrete answer Y, and let π_{pred} be a conditional predictor. Fix $Q = q$ and $G = g$; all entropies and expectations are under π_{data} and are assumed finite. For expected log-loss $\mathcal{L}(\pi_{\mathrm{pred}}; Z, q, g) = \mathbb{E}[-\log \pi_{\mathrm{pred}}(Y \mid Z, q, g) \mid q, g]$, the Bayes-optimal risk is $R^*_Z(q, g) := \inf_{\pi_{\mathrm{pred}}} \mathcal{L}(\pi_{\mathrm{pred}}; Z, q, g) = H(Y \mid Z, q, g)$ (Gneiting & Raftery, 2007).

Let $\mathbf{U}_V = (U_v)_{v \in V}$ be the complete visual source with discrete semantic variables U_v, and $\widehat{\mathbf{U}}_V(S) = (\widehat{U}_v(S))_{v \in V}$ the information recoverable from retained tokens S. Their Bayes risks are R^*_{full} and R^*_S, respectively. We model task relevance by a semantic-query index J with $\Pr(J = v \mid q, g) = \Pr(J = v \mid Q = q)$, sampled independently of $(\mathbf{U}_V, \widehat{\mathbf{U}}_V(S))$ given q, g for each fixed S and recovery channel. Under $Y \perp \widehat{\mathbf{U}}_V(S) \mid \mathbf{U}_V, Q, G$, the log-loss/side-information identity (Jiao et al., 2015), entropy bound, and data-processing inequality (Polyanskiy & Wu, 2025, Theorems 3.4 and 3.7) yield

$$\begin{aligned} \Delta R^*(S; q, g) &:= R^*_S(q, g) - R^*_{\mathrm{full}}(q, g) \\ &\le H(\mathbf{U}_V \mid q, g) - I\Big(U_J; \widehat{U}_J(S) \mid J, q, g\Big) . \end{aligned} \tag{1}$$

Since $H(\mathbf{U}_V \mid q, g)$ is independent of S, maximizing the recovered information about the queried semantic unit minimizes this upper bound for a fixed query distribution.

Semantic erasure and source-weighted mutual-information coverage. To evaluate recovery at each source position, we introduce two binary gates. The coverage gate E_v indicates whether S supplies a usable representative, with probability $M_v(S; q, g) := \Pr(E_v = 1 \mid q, g)$. The reliability gate C_v indicates whether the source representation faithfully carries U_v; for a fixed recovery chan-

nel, we treat $\Pr(C_v = 1 \mid q, g)$ as invariant to S. Conditional on q, g, we assume $C_v \perp E_v$ and $(C_v, E_v) \perp U_v$. The semantic erasure model is

$$\widehat{U}_v(S) = \begin{cases} U_v, & C_v E_v = 1, \\ \perp, & C_v E_v = 0. \end{cases} \tag{2}$$

Here $\perp$ is an erasure symbol outside the alphabet of U_v. The independent gates succeed with probability $\Pr(C_v = 1 \mid q, g) M_v(S; q, g)$. The standard erasure-channel identity then gives $I(U_v; \widehat{U}_v(S) \mid q, g) = H(U_v \mid q, g) \Pr(C_v = 1 \mid q, g) M_v(S; q, g)$ (Polyanskiy & Wu, 2025). Expanding over J and substituting this identity, we define the task-relevant recovered information to maximize:

$$\begin{aligned} \mathcal{I}_{\text{task}}(S; q, g) &:= I\Big(U_J; \widehat{U}_J(S) \mid J, q, g\Big) \\ &= \sum_{v \in V} \Pr(J = v \mid Q = q) I\Big(U_v; \widehat{U}_v(S) \mid q, g\Big) \\ &= \sum_{v \in V} H(U_v \mid q, g) \Pr(C_v = 1 \mid q, g) \Pr(J = v \mid Q = q) M_v(S; q, g). \end{aligned} \tag{3}$$

In Eq. 3, the source contribution is weighted by $H(U_v \mid q, g)$, which measures the information uncertainty associated with position v, and by $\Pr(C_v = 1 \mid q, g)$ and $\Pr(J = v \mid Q = q)$, which measure, respectively, the reliability of its representation and the likelihood that the query requires it. The factor $M_v(S; q, g)$ measures how well the retained set covers that position. The result is a source-weighted coverage objective for visual-token pruning. The complete derivation is given in Appendix B.1–B.2.

3.2 MICO

The objective in Eq. 3 contains three theoretical factors that cannot be computed directly during a forward pass. We therefore approximate them with observable proxies, denoted by $\widehat{\mathcal{H}}_v$, $\widehat{\mathcal{R}}_v$, and $\widehat{\mathcal{P}}_v$.

We use the hidden-state norm $\widehat{\mathcal{H}}_v = \|x_v\|_2$ as a proxy for a token's information amount, since its magnitude reflects the strength of the encoded visual signal. We approximate representation reliability $\widehat{\mathcal{R}}_v$ with the vision encoder's CLS-token attention, since it measures how strongly the representation at v participates in global aggregation. At this point the visual tokens have not yet interacted with the text, so aggressive pruning could discard query-relevant tokens prematurely; Stage 1 therefore keeps a larger candidate pool, and a second selection inside the LLM uses text-to-visual attention as a query-dependent proxy for task relevance $\widehat{\mathcal{P}}_v$. The proxies are:

$$(\widehat{\mathcal{H}}_v, \widehat{\mathcal{R}}_v, \widehat{\mathcal{P}}_v) = \begin{cases} \big(\|x_v^{\text{vis}}\|_2, \overline{A}^{\text{vis}}_{\text{CLS}\to v}, 1\big), & \text{Stage 1}, \\ \big(\|x_v^{(\ell_\star)}\|_2, 1, \overline{A}^{(\ell_\star)}_{\text{text}\to v}\big), & \text{Stage 2}. \end{cases} \tag{4}$$

Here x_v^{vis} denotes the pre-projection visual feature, and $x_v^{(\ell_\star)}$ denotes the visual feature at the selected layer; $\overline{A}^{\text{vis}}_{\text{CLS}\to v}$ is the attention from the CLS token to token v in the vision encoder, averaged over heads, and $\overline{A}^{(\ell_\star)}_{\text{text}\to v}$ is the attention from the text tokens to token v at decoder layer $\ell_\star$, averaged over text queries and heads. In both stages, we estimate coverage using the best retained visual representative:

$$\widehat{M}_v(S) = \max_{s \in S} \left[\frac{x_v^\top x_s}{\|x_v\|_2 \|x_s\|_2} \right]_+, \tag{5}$$

where $[a]_+ = \max(a, 0)$ and $\widehat{M}_v(\varnothing) = 0$. The overall proxy objective sums the weighted coverage over all source positions:

$$\widehat{\mathcal{I}}_{\text{task}}(S) = \sum_{v \in V} \widehat{\mathcal{H}}_v \widehat{\mathcal{R}}_v \widehat{\mathcal{P}}_v \widehat{M}_v(S).$$

Starting from $S = \varnothing$, we use a greedy search to add the token $i^\star = \arg\max_{i \in V \setminus S} \Delta\widehat{\mathcal{I}}_{\text{task}}(i \mid S)$ until budget is reached, where $\Delta\widehat{\mathcal{I}}_{\text{task}}(i \mid S) = \widehat{\mathcal{I}}_{\text{task}}(S \cup \{i\}) - \widehat{\mathcal{I}}_{\text{task}}(S)$. The objective is normalized, monotone, and submodular, so greedy selection achieves a $(1 - 1/e)$ approximation

under the cardinality constraint $|S| \leq B$, where B is the stage budget (Krause & Golovin, 2014; Nemhauser et al., 1978). Appendices B.3 and C.1 give the procedure and proof; Appendix C.2 checks the proxies against their theoretical counterparts, and Table 7 evaluates substitutes.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate MiCo on MLLMs of different architectures, covering the LLaVA family (LLaVA-1.5-7B/13B (Liu et al., 2024a) and LLaVA-NeXT-7B/13B (Liu et al., 2024b)), the Qwen family (Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B-Instruct (Bai et al., 2025a), and Qwen3.5-9B (Qwen Team, 2026)), and InternVL3-8B (Zhu et al., 2025). For video understanding, we use LLaVA-Video-7B (Zhang et al., 2024c).

Benchmarks. We use ten image tracks for the LLaVA family and eight for Qwen-VL/InternVL (e.g., GQA, POPE, MME, MMBench, SEED, and MMMU), supplementary tracks for grounding, document, chart, and OCR understanding (e.g., RefCOCO, DocVQA, OCRBench, and VTC-Bench), and four video benchmarks (VideoMME, MVBench, LongVideoBench, and MLVU); see Appendix A.

Baselines. We compare MiCo with FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2024b), VisionZip (Yang et al., 2025b), DivPrune (Alvar et al., 2025), VisPruner (Zhang et al., 2025a), HoloV (Zou et al., 2025), VScan (Zhang et al., 2026), MMTok (Dong et al., 2026), and ApET (Ma et al., 2026).

Implementation details. We evaluate all methods with the public evaluation stack of each model family: the official LLaVA evaluation scripts for LLaVA-1.5 and LLaVA-NeXT, VLMEvalKit (Duan et al., 2024) for Qwen2.5-VL, Qwen3-VL, Qwen3.5, and InternVL3, and lmms-eval (Zhang et al., 2024a) for LLaVA-Video. Each model keeps its native processor, prompt format, and official scorer; we use batch size one with greedy decoding, and all pruning methods share the same token budgets and evaluation pipeline. For MiCo, Stage 1 keeps a candidate pool of $N_1 = 2T$ tokens on the visual-encoder features, and Stage 2 re-selects R of them after decoder layer K, with R chosen so that the average number of visual tokens over all decoder layers equals the budget T (Appendix C.3); the per-model K values are listed in Table 19. Details are provided in Appendix C.

4.2 LLAVA SERIES

We evaluate LLaVA-1.5 and LLaVA-NeXT at 7B and 13B scales with pruning ratios of 77.8%, 88.9%, and 94.4%. MiCo achieves the highest aggregate score among reported pruners in 11 of 12 model–budget settings (Tables 1 and 2); the exception is LLaVA-NeXT-7B at 77.8% pruning, where VScan is 0.3 points higher. At 94.4% pruning, it retains 93.3%/94.9% of baseline performance on LLaVA-1.5-7B/13B and 96.7%/97.5% on LLaVA-NeXT-7B/13B (99.3% on NeXT-13B at 88.9%). Complete results are in Appendix G.

4.3 ADVANCED ARCHITECTURES

Table 2 compares Qwen2.5-VL-7B, Qwen3-VL-8B, Qwen3.5-9B, and InternVL3-8B at $T = 256$ and $T = 128$. MiCo achieves the highest aggregate score in all eight model–budget settings. At $T = 128$, it retains 94.7%, 92.8%, 96.0%, and 93.3% of baseline performance, respectively. On Qwen2.5 and Qwen3, this exceeds MMTok by 3.9 and 1.0 percentage points; on InternVL3, the corresponding margin is 4.5 points. Qwen3.5 retains 98.7%/96.0% at $T = 256/128$ versus ApET's 98.1%/92.8%, a larger advantage at the tighter budget. These gains across model families support MiCo's applicability beyond LLaVA (full results in Appendix G).

4.4 LLAVA-VIDEO

Table 3 compares LLaVA-Video-7B at per-frame budgets $T = 32$ and $T = 16$. MiCo achieves the highest Acc at both budgets and leads seven of eight task–budget comparisons; at $T = 32$ it exceeds

Table 1: Performance on LLaVA-NeXT-13B across ten benchmarks.

Method	GQA	SQA-IMG	TextVQA	POPE	MME	MMB-EN	MMB-CN	SEED	AI2D	MMMU	Acc	Rel
Upper Bound — 2880 tokens (no pruning)												
Vanilla	64.4	73.1	63.2	85.3	1539.5	68.5	61.2	71.6	70.1	35.7	67.0	100.0%
Retain 640 tokens (77.8% pruned)												
FastV (ECCV'24)	60.9	71.7	60.7	80.2	1516.7	65.5	59.9	67.4	67.7	35.8	64.6	96.4%
DivPrune (CVPR'25)	63.5	72.2	59.2	86.5	1526.1	67.5	62.9	69.4	68.4	37.8	66.4	99.1%
VisPruner (ICCV'25)	62.6	71.4	**62.1**	85.2	1561.2	67.4	63.1	68.8	68.2	37.0	66.4	99.1%
VScan (TMLR'26)	62.8	72.2	61.8	85.2	1553.6	67.6	62.7	68.2	69.9	36.6	66.5	99.2%
MMTok (ICLR'26)	63.7	71.4	60.8	86.8	1540.9	66.6	62.5	69.7	68.1	**38.3**	66.5	99.2%
ApET (CVPR'26)	64.2	72.5	58.0	86.4	1475.4	67.7	62.5	69.5	**70.9**	36.7	66.2	98.8%
MiCo	**64.3**	**73.2**	61.9	**87.0**	**1573.9**	**69.0**	**63.4**	**71.4**	70.2	37.2	**67.6**	**100.9%**
Retain 320 tokens (88.9% pruned)												
FastV (ECCV'24)	54.6	70.5	55.4	63.6	1279.0	59.8	54.4	59.2	65.0	35.8	58.2	86.9%
DivPrune (CVPR'25)	61.8	72.3	57.6	85.2	1473.0	65.9	61.9	67.2	67.7	**37.2**	65.0	97.1%
VisPruner (ICCV'25)	60.8	70.1	60.3	81.1	1486.0	65.7	62.5	65.1	67.1	36.4	64.4	96.1%
VScan (TMLR'26)	61.0	**72.4**	59.3	82.0	1496.4	65.2	59.8	64.9	66.9	36.3	64.3	95.9%
MMTok (ICLR'26)	62.78	71.49	59.05	86.06	1501.1	65.0	61.7	67.7	67.9	**37.2**	65.4	97.6%
ApET (CVPR'26)	61.7	71.4	54.7	84.1	1455.0	65.6	60.1	65.5	**68.4**	36.4	64.1	95.6%
MiCo	**63.3**	71.3	**60.9**	**86.5**	**1543.3**	**67.9**	**63.2**	**69.5**	68.3	37.0	**66.5**	**99.3%**
Retain 160 tokens (94.4% pruned)												
DivPrune (CVPR'25)	60	71.4	56.3	81.9	1436.7	65.1	60.9	64.5	67.3	36.6	63.6	94.9%
VisPruner (ICCV'25)	58.4	71.2	**58.4**	76.2	1390.8	64.1	59.6	61.1	65.5	36.2	62.0	92.6%
VScan (TMLR'26)	58.2	70.3	55.0	75.5	1385.1	61.5	53.4	61.5	63.4	34.1	60.2	89.9%
MMTok (ICLR'26)	**62.02**	**72.19**	56.47	**85.52**	1465.7	65.3	60.3	65.5	67.3	**37.0**	64.5	96.3%
ApET (CVPR'26)	59.1	71.4	52.9	79.6	1342.7	62.1	55.8	62.3	65.6	36.8	61.3	91.4%
MiCo	62.0	71.8	**58.4**	85.0	**1488.4**	**66.2**	**63.1**	**67.6**	**68.0**	36.7	**65.3**	**97.5%**

Table 2: Relative performance (%) across LLaVA, Qwen-VL, and InternVL models; Vanilla is 100%.

Method	LLaVA									Qwen-VL						InternVL	
Model	1.5-7B			1.5-13B			NeXT-7B			2.5-VL-7B		3-VL-8B		3.5-9B		3-8B	
Budget T	128	64	32	128	64	32	640	320	160	256	128	256	128	256	128	256	128
FastV	92.6	75.5	–	94.1	85.0	–	95.4	80.4	–	93.0	77.9	84.3	63.7	84.0	62.8	95.5	81.5
DivPrune	96.8	94.4	91.3	96.7	93.2	91.2	99.0	97.2	94.2	94.5	90.3	93.6	88.6	95.9	91.8	93.3	87.7
VisPruner	96.6	93.3	87.8	96.6	93.6	88.5	98.7	95.6	91.0	91.8	88.0	94.2	86.8	97.0	90.1	89.4	81.1
VScan	97.8	96.0	91.8	97.3	96.0	88.7	**99.5**	96.2	90.3	94.8	90.6	95.4	90.1	–	–	92.9	87.3
MMTok	97.0	95.5	92.4	96.8	95.2	93.0	99.4	97.4	95.1	95.8	90.8	95.8	91.8	97.3	92.8	94.2	88.8
ApET	96.4	94.3	90.9	96.4	94.1	91.1	99.4	96.0	92.0	91.7	83.5	93.9	88.2	98.1	92.8	92.9	88.2
MiCo	**98.5**	**96.1**	**93.3**	**98.3**	**96.7**	**94.9**	99.2	**98.5**	**96.7**	**97.4**	**94.7**	**96.5**	**92.8**	**98.7**	**96.0**	**96.8**	**93.3**

ApET, the strongest baseline, by 3.1 points on MLVU and 1.7 Acc points overall, and at $T = 16$ it prunes 90.5% of visual tokens while retaining 93.0% of baseline performance with a 2.31× speedup.

4.5 Efficiency Study

At $T = 16$ on LLaVA-Video-7B (Table 3), MiCo retains 93.0% of baseline performance with a 2.31× complete-answer speedup, 75.8% fewer logical FLOPs, and the lowest peak memory of all compared methods (21.32 vs. 25.59 GiB for Vanilla; FastV, SparseVLM, and VisPruner exceed the unpruned model). Figure 1(c) shows 99.62% POPE F1 retention at a 3.83× one-token speedup on NeXT-13B; Figure 3 shows 99.61% at 2.93× on NeXT-7B and 98.25% VideoMME retention at 1.81× on Video, together with where MiCo's latency goes. Protocols and the full NeXT-7B/13B comparisons (Table 10) are in Appendix D.

4.6 Challenging Visual Understanding

Grounding. Grounding tests whether the retained tokens support spatial localization (parenthesized values are retention relative to Vanilla). On RefCOCO, RefCOCO+, and RefCOCOg, MiCo exceeds FastV, VisPruner, and ApET in all 12 combinations in Table 4; gRefCOCO results are in Appendix H.2.

Table 3: Performance and efficiency on LLaVA-Video-7B.

Method	Performance (%)↑						Efficiency↓		
	VideoMME	MVBench	LongVideo Bench	MLVU	Acc	Rel	Latency (ms)	Peak memory (GiB)	Logical FLOPs (T)
Vanilla	63.63	58.18	59.01	67.75	62.14	100.0%	1855.0	25.59	243.833
Budget $T = 32$ tokens per frame									
FastV	56.00	52.58	52.40	58.50	54.87	88.3%	975.1	41.30	75.951
SparseVLM	59.00	54.40	53.70	60.70	56.95	91.6%	1105.3	41.30	76.338
DivPrune	59.30	53.85	56.40	61.50	57.76	93.0%	**881.8**	22.61	74.738
VisPruner	60.15	55.20	56.69	62.46	58.63	94.3%	884.2	47.96	**74.732**
MMTok	59.67	54.90	57.07	61.96	58.40	94.0%	1098.9	22.61	74.734
ApET	59.63	55.33	57.44	62.89	58.82	94.7%	924.1	22.61	74.734
MiCo	**61.81**	**56.78**	**57.67**	**65.99**	**60.56**	**97.5%**	933.8	**21.32**	74.834
Budget $T = 16$ tokens per frame									
FastV	50.00	46.45	46.60	52.80	48.96	78.8%	885.0	41.30	60.937
SparseVLM	49.80	50.80	47.60	52.00	50.05	80.5%	1093.1	41.30	61.750
DivPrune	56.70	51.83	52.10	58.60	54.81	88.2%	**783.7**	22.61	59.002
VisPruner	56.41	52.15	52.28	57.14	54.50	87.7%	789.6	47.96	58.997
MMTok	55.96	52.18	54.38	59.27	55.45	89.2%	902.0	22.61	58.998
ApET	57.04	54.00	**56.32**	59.06	56.61	91.1%	818.2	22.61	58.994
MiCo	**59.78**	**54.80**	54.75	**61.93**	**57.82**	**93.0%**	804.7	**21.32**	**58.913**

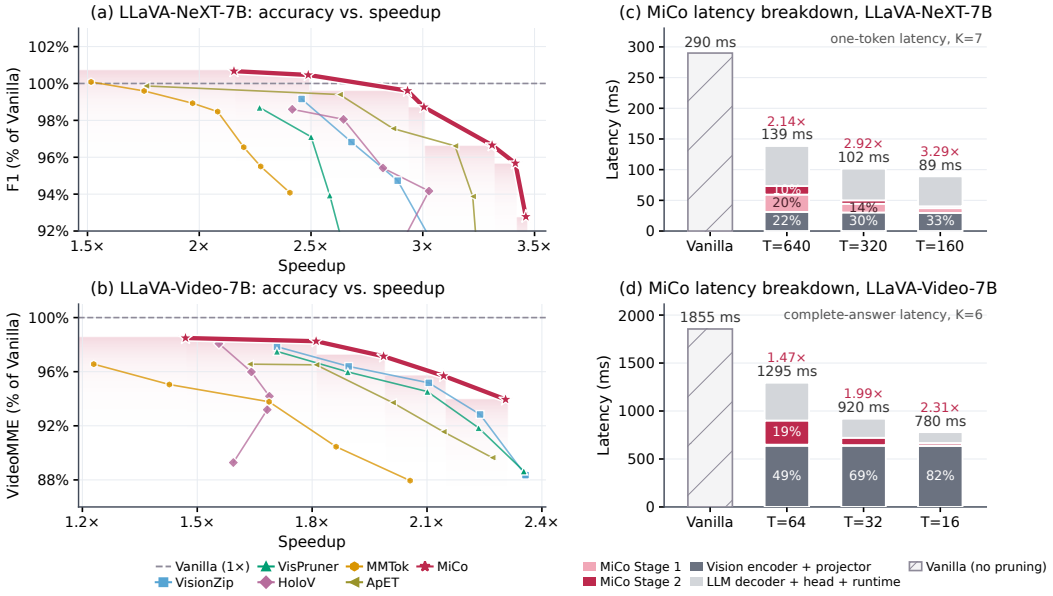


Figure 3: Performance–efficiency trade-offs (a, b) and MiCo's latency by component with speedups over Vanilla (c, d); NeXT one-token, Video complete-answer latency. Details in Appendix D.

Fine-grained understanding. Table 5 compares MiCo with FastV, VisPruner, and ApET on TextVQA, ChartQA, and DocVQA using Qwen2.5-VL-7B and Qwen3-VL-8B. MiCo leads all twelve displayed settings; at $T = 128$ on Qwen2.5 its ChartQA score is 70.68 versus 36.68 for VisPruner, the strongest displayed baseline. The full comparison with eight external methods, including settings led by HoloV or MMTok, is in Appendix H.3.

VTC-Bench. VTC-Bench (Liao et al., 2026) uses a fixed reference model to define downsampling-sensitive (Group A) and downsampling-robust (Group B) questions. We evaluate Qwen2.5-VL-7B on original images at $T = 256/128$ (Table 6; column percentages are the reference image-area reductions used for grouping, not token budgets). MiCo outperforms FastV, VisPruner, and ApET in all 20 settings, retaining 75.8–84.9% of baseline on Group A and 90.7–95.0% on Group B at $T = 128$. Details are in Appendix H.1.

Table 4: REC scores (%) on RefCOCO, RefCOCO+, and RefCOCOg.

Method	Qwen2.5-VL-7B			Qwen3-VL-8B		
	RefCOCO	RefCOCO+	RefCOCOg	RefCOCO	RefCOCO+	RefCOCOg
Vanilla	86.8 (100.0%)	79.4 (100.0%)	85.5 (100.0%)	92.4 (100.0%)	88.7 (100.0%)	90.4 (100.0%)
Retain 256 tokens (80.2% pruned)						
FastV	33.0 (38.0%)	28.3 (35.6%)	35.5 (41.5%)	67.2 (72.7%)	59.0 (66.5%)	68.1 (75.3%)
VisPruner	22.3 (25.7%)	17.8 (22.4%)	21.7 (25.4%)	77.8 (84.2%)	71.7 (80.8%)	77.0 (85.2%)
ApET	44.0 (50.7%)	37.1 (46.7%)	42.6 (49.8%)	51.4 (55.6%)	47.1 (53.1%)	49.7 (55.0%)
MiCo	**75.0** (86.4%)	**68.1** (85.8%)	**74.0** (86.5%)	**82.8** (89.6%)	**77.3** (87.1%)	**81.9** (90.6%)
Retain 128 tokens (90.1% pruned)						
FastV	7.1 (8.2%)	4.2 (5.3%)	4.9 (5.7%)	23.8 (25.8%)	19.6 (22.1%)	24.4 (27.0%)
VisPruner	11.3 (13.0%)	8.6 (10.8%)	10.6 (12.4%)	49.1 (53.1%)	43.0 (48.5%)	47.8 (52.9%)
ApET	17.9 (20.6%)	14.6 (18.4%)	15.4 (18.0%)	33.4 (36.1%)	29.5 (33.3%)	33.3 (36.8%)
MiCo	**50.0** (57.6%)	**44.1** (55.5%)	**47.7** (55.8%)	**64.5** (69.8%)	**57.8** (65.2%)	**65.6** (72.6%)

Table 5: TextVQA, ChartQA, and DocVQA scores (%).

Method	Qwen2.5-VL-7B						Qwen3-VL-8B					
	$T = 256$			$T = 128$			$T = 256$			$T = 128$		
	Text	Chart	Doc	Text	Chart	Doc	Text	Chart	Doc	Text	Chart	Doc
Vanilla	85.06	87.28	94.86	85.06	87.28	94.86	84.28	83.04	95.72	84.28	83.04	95.72
FastV	78.44	68.28	38.22	46.45	33.20	14.54	62.18	28.64	14.79	32.88	16.24	9.83
VisPruner	64.02	56.08	27.66	49.40	36.68	19.69	71.96	63.52	39.63	54.58	39.64	23.00
ApET	69.17	50.60	23.12	45.82	31.16	14.35	59.54	54.52	39.25	41.27	35.44	25.59
MiCo	**80.96**	**81.16**	**57.44**	**71.74**	**70.68**	**38.58**	**79.69**	**66.80**	**64.98**	**73.44**	**50.12**	**46.35**

Table 6: VTC-Bench Rel (%) on Qwen2.5-VL-7B. Entries show Group A (B): downsampling-sensitive (robust) questions under the reference model; $\Delta = |\mathrm{Rel}_B - \mathrm{Rel}_A|$ is the gap in percentage points.

Method	Reference image-area reduction				
	75.00%	88.89%	93.75%	96.00%	99.00%
Vanilla	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)
Token budget $T = 256$					
FastV	76.1 (90.8)	80.0 (91.7)	81.0 (92.4)	82.3 (92.5)	85.1 (93.9)
VisPruner	74.4 (89.7)	78.3 (91.8)	79.3 (93.2)	80.1 (93.6)	83.4 (94.8)
ApET	71.8 (87.6)	76.0 (90.0)	76.8 (91.8)	77.4 (92.5)	80.7 (93.4)
MiCo	**90.1** (**96.1**)	**91.6** (**96.7**)	**91.9** (**97.3**)	**92.4** (**97.3**)	**93.3** (**98.3**)
Token budget $T = 128$					
FastV	48.9 (66.8)	50.1 (70.1)	48.9 (73.6)	48.9 (75.0)	53.6 (76.3)
VisPruner	62.7 (82.0)	64.9 (85.9)	66.1 (88.7)	67.7 (89.6)	72.3 (90.9)
ApET	56.5 (77.8)	56.8 (82.1)	58.4 (85.6)	60.4 (87.0)	66.2 (88.2)
MiCo	**75.8** (**90.7**)	**79.1** (**92.6**)	**81.5** (**93.4**)	**82.3** (**93.7**)	**84.9** (**95.0**)

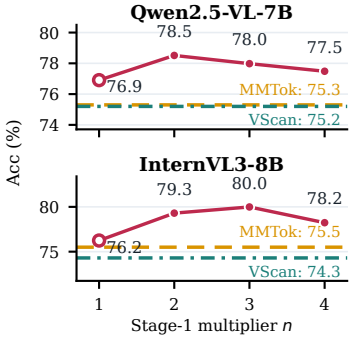


Figure 4: Stage-1 budget sensitivity at $T = 128$: eight-benchmark Acc for $n = 1$–4 (hollow: MiCo-S1 reference) vs. MMTok and VScan.

4.7 Ablation Study

We ablate the components of MiCo's two-stage design and the generality of the mutual information coverage objective.

Stage-1 budget and objective's effectiveness. We change the Stage-1 pool size nT at a fixed layer-average budget of $T = 128$ and report the eight-benchmark Acc on Qwen2.5-VL-7B and InternVL3-8B in Figure 4. MiCo exceeds MMTok and VScan at every n on both models, proving our gains do not depend on the pool size used in Stage 1. Specifically, at $n = 1$ the pool equals the budget and MiCo reduces to its single-stage implementation (MiCo-S1); every $n > 1$ improves on it, proving that the LLM-side refinement works, while $n = 1$ still beats MMTok and VScan, proving that our objective is promising even in a single stage. Appendix E.4 gives details, and Appendix G records MiCo-S1's results on more models.

Two-stage components. Figure 5 ablates the components of MiCo's two stages. The complete configuration achieves the highest seven-benchmark mean, $\mathrm{Acc}_7 = 78.4$, and Acc_7 generally decreases as more components are disabled, showing that all signals contribute.

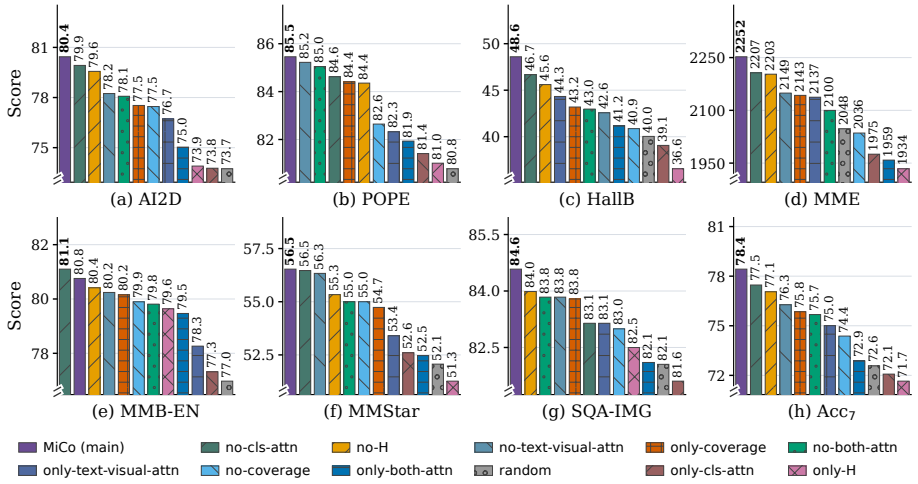


Figure 5: Two-stage component comparison on Qwen2.5-VL-7B at $T = 128$.

Proxy substitutions. We evaluate replacements for the information, coverage, reliability, and task-relevance proxies on Qwen2.5-VL-7B at $T = 128$ in Table 7. The information proxy is replaced with normalized hidden-state entropy, the coverage and reliability proxies with similarity measures over visual tokens, and the task-relevance proxy with similarity measures over text tokens. MiCo achieves the highest score, and all variants except the coverage replacement still outperform MMTok, supporting the applicability and robustness of our mutual information coverage objective beyond MiCo's proxy choices. Appendix E.2 details the replacements, and Appendix E.3 compares further task-relevance variants.

Table 7: Proxy substitutions on Qwen2.5-VL-7B at $T = 128$. The second line in each cell reports the score difference Δ from MMTok.

Replaced proxy	AI2D	POPE	HallB	MME	$\mathrm{MMB_{EN}}$	$\mathrm{MMB_{CN}}$	MMStar	SQA	Acc	Rel
MMTok (ICLR'26)	76.3	84.2	43.3	2120.7	79.8	77.5	53.7	81.4	75.3	90.8%
Replace Coverage	70.14 $\Delta -6.16$	78.10 $\Delta -6.10$	35.80 $\Delta -7.50$	1688.69 $\Delta -432.01$	75.69 $\Delta -4.11$	73.71 $\Delta -3.79$	47.20 $\Delta -6.50$	81.31 $\Delta -0.09$	68.30 $\Delta -7.00$	82.34% $\Delta -8.46$
Replace H	78.47 $\Delta +2.17$	84.49 $\Delta +0.29$	45.43 $\Delta +2.13$	2152.60 $\Delta +31.90$	<u>80.84</u> $\Delta +1.04$	78.26 $\Delta +0.76$	52.60 $\Delta -1.10$	83.04 $\Delta +1.64$	76.35 $\Delta +1.05$	92.04% $\Delta +1.24$
Replace Relevance	79.08 $\Delta +2.78$	**85.54** $\Delta +1.34$	45.88 $\Delta +2.58$	2142.01 $\Delta +21.31$	79.73 $\Delta -0.07$	<u>78.95</u> $\Delta +1.45$	56.47 $\Delta +2.77$	<u>84.28</u> $\Delta +2.88$	77.13 $\Delta +1.83$	92.98% $\Delta +2.18$
Replace Reliability	<u>79.83</u> $\Delta +3.53$	84.59 $\Delta +0.39$	<u>46.13</u> $\Delta +2.83$	<u>2207.75</u> $\Delta +87.05$	**81.10** $\Delta +1.30$	<u>78.95</u> $\Delta +1.45$	**56.73** $\Delta +3.03$	83.34 $\Delta +1.94$	<u>77.63</u> $\Delta +2.33$	<u>93.59%</u> $\Delta +2.79$
MiCo	**80.44** $\Delta +4.14$	<u>85.46</u> $\Delta +1.26$	**48.60** $\Delta +5.30$	**2252.37** $\Delta +131.67$	80.76 $\Delta +0.96$	**79.12** $\Delta +1.62$	<u>56.53</u> $\Delta +2.83$	**84.58** $\Delta +3.18$	**78.51** $\Delta +3.21$	**94.65%** $\Delta +3.85$

5 Conclusion

We derive a general visual token pruning objective from task log-loss through a semantic erasure model and instantiate its factors with observable proxies, yielding a monotone submodular coverage objective that MiCo optimizes greedily in two stages. Across diverse MLLMs and image/video benchmarks, MiCo retains strong performance while reducing inference cost. We hope the objective encourages better proxies and further improves the performance–speed trade-off.

AI use statement

We used large language models to aid and polish the writing, for retrieval and discovery of related work, and for research ideation and execution. The authors reviewed and verified all AI-assisted content and take full responsibility for the final content of this work.

Reproducibility statement

MiCo is training-free and all experiments use publicly available pretrained models and benchmarks. Implementation details, per-model configurations, evaluation protocols, and proofs are given in the appendix. We will open-source the code and configuration files so that all reported results can be reproduced.

Ethics statement

This work uses only publicly available pretrained models and standard public benchmarks, involves no human subjects or newly collected data, and introduces no capabilities beyond those of the underlying models. We do not foresee ethical concerns beyond those already associated with the base models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

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

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

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.

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

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. Scope: Saliency-coverage oriented token pruning for efficient multimodel llms. *arXiv preprint arXiv:2510.24214*, 2025.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. 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.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 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.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023. URL https://arxiv.org/abs/2308.16182.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Jiantao Jiao, Thomas A. Courtade, Kartik Venkat, and Tsachy Weissman. Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, 61(10):5357–5365, 2015. doi: 10.1109/TIT.2015.2462848. URL https://people.eecs.berkeley.edu/~courtade/pdfs/JiaoJustification_TIT2015.pdf.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2014. URL https://las.inf.ethz.ch/files/krause12survey.pdf.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

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, 2023.

Chenfei Liao, Wensong Wang, Zichen Wen, Xu Zheng, Yiyu Wang, Haocong He, Yuanhuiyi Lyu, Lutao Jiang, Xin Zou, Yuqian Fu, Bin Ren, Linfeng Zhang, and Xuming Hu. Are we using the right benchmark: An evaluation framework for visual token compression methods. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4236–4253, San Diego, California, United States, July

2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.195. URL https://aclanthology.org/2026.acl-long.195/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

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

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models, 2023b.

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

Qiankun Ma, Ziyao Zhang, Haofei Wang, Jie Chen, Zhen Song, and Hairong Zheng. ApET: Approximation-error guided token compression for efficient VLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ma_ApET_Approximation-Error_Guided_Token_Compression_for_Efficient_VLMs_CVPR_2026_paper.html.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971. URL https://link.springer.com/article/10.1007/BF01588971.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025. doi: 10.1017/9781108966351. URL https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf.

Qwen Team. Qwen3.5-9B model card. https://huggingface.co/Qwen/Qwen3.5-9B, 2026. Accessed September 23, 2026.

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

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. VScan: Rethinking visual token reduction for efficient large vision-language models. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=KZYhyilFnt.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024a.

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

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=BLLixcuZgl.

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*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

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

Qingchan Zhu, Weihang You, Hanqi Jiang, Changdi Yang, Tianming Liu, and Geng Yuan. Who speaks for the pruned? visual token pruning as coverage optimization. *arXiv preprint arXiv:2609.03158*, 2026. URL https://arxiv.org/abs/2609.03158.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/38b445e98823a02479582302ee3aa8e4-Abstract-Conference.html.

A Benchmark Definitions and Evaluation Protocol

This section defines the benchmark inventory, task coverage, and scoring conventions used in the experiments. The two main image suites overlap: the LLaVA tables use ten benchmark tracks, while the Qwen-VL and InternVL tables use eight. Their union is twelve core image tracks when the English and Chinese MMBench evaluations are counted separately. We additionally report eight supplementary image tracks (RefCOCO, RefCOCO+, RefCOCOg, gRefCOCO, ChartQA, DocVQA, OCRBench, and VTC-Bench) and four video benchmarks (VideoMME, MVBench, LongVideoBench, and MLVU). In total, this appendix covers twenty image tracks (twelve core and eight supplementary, counting VTC-Bench as one track) and four video benchmarks; supplementary tracks never enter a main-table Acc.

A.1 Core Image Benchmarks

The LLaVA tables use GQA, SQA-IMG, TextVQA, POPE, MME, MMBench-EN, MMBench-CN, SEED, AI2D, and MMMU; the Qwen-VL and InternVL tables use AI2D, POPE, HallusionBench, MME, MMBench-EN, MMBench-CN, MMStar, and SQA-IMG. A benchmark enters a model's Acc only if it appears in that model's main table.

General visual understanding and reasoning. GQA (Hudson & Manning, 2019) evaluates compositional visual reasoning over objects, attributes, and relations grounded in scene graphs. We use the balanced evaluation split and report exact-match accuracy; the standard split contains 12,578 questions. ScienceQA-IMG (Lu et al., 2022) is the image-grounded subset of the multimodal science multiple-choice benchmark. It requires selecting an answer after combining the question, explanatory context, and a diagram or image; we report multiple-choice accuracy on the released image subset. AI2D (Kembhavi et al., 2016) evaluates question answering over scientific diagrams and their associated text. Its released collection contains more than 5,000 diagrams and 15,000 questions; we report answer accuracy on the test split (3,088 questions). MMMU (Yue et al., 2024) contains 11.5K college-level multimodal questions spanning six broad disciplines and 30 subjects; we use the recorded single-image evaluation protocol on the validation split and report multiple-choice accuracy. MMStar (Chen et al., 2024b) is a 1,500-question diagnostic suite covering coarse and fine-grained perception, reasoning, and knowledge-intensive visual tasks; we report its official accuracy.

Text and broad multimodal understanding. TextVQA (Singh et al., 2019) tests reading and reasoning over scene text; the validation split contains 5,000 questions and is scored with the VQA consensus metric. MMBench (Liu et al., 2024c) evaluates 20 ability dimensions, including recognition, localization, OCR, and reasoning. MMBench-EN and MMBench-CN are treated as two separate tracks, evaluated with the released language-specific scorers and Circular accuracy. SEED-Bench (Li et al., 2024a) provides objective multiple-choice evaluation across multimodal capability dimensions; we use its image-only evaluation subset and report the official accuracy.

Hallucination diagnostics. POPE (Li et al., 2023) probes object hallucination with binary object-presence questions over its random, popular, and adversarial categories (3,000 released questions each). The reported metric follows each family's official evaluator: the LLaVA tables report the average F1 score over the three categories, as computed by the LLaVA-1.5 evaluation scripts, whereas the Qwen-VL and InternVL tables report the overall accuracy returned by VLMEvalKit, whose runner expands the category outputs to 9,000 scored entries. HallusionBench (Guan et al., 2024) is a diagnostic suite for language hallucination and visual illusion; we report VLMEvalKit's overall average, i.e., the mean of its all-question accuracy (aAcc), figure-level accuracy (fAcc), and question-pair accuracy (qAcc).

Perception and cognition. MME (Fu et al., 2023) covers perception and cognition in 14 subtasks with 2,374 questions over 1,187 images (two questions per image). The MME column keeps the benchmark's native summed score, whose scale differs by family: the LLaVA tables report the Perception score (maximum 2,000), following the LLaVA evaluation scripts, whereas the Qwen-VL and InternVL tables report the Perception+Cognition total (maximum 2,800) returned by VLMEvalKit.

In both cases the summed score is divided by 20 only when forming Acc, so MME enters the LLaVA aggregates on a 0–100 scale and the Qwen-VL/InternVL aggregates on a 0–140 scale.

A.2 Supplementary Image Benchmarks

Text, chart, and document understanding. ChartQA (Masry et al., 2022) requires visual and logical reasoning over charts. We use the test split and its relaxed accuracy metric (2,500 questions in the evaluation record). DocVQA (Mathew et al., 2021) evaluates question answering over document images; we use the validation split and report ANLS over 5,349 questions. OCRBench (Liu et al., 2023b) aggregates 29 OCR-oriented datasets covering text recognition, scene-text VQA, document VQA, key-information extraction, and handwritten mathematical expression recognition. We report its released aggregate score. The 100 OCRBench samples used to select the refinement layer (Appendix F.2) never enter a main-table Acc.

Referring expression comprehension. RefCOCO, RefCOCO+, and RefCOCOg evaluate localization of the region described by a natural-language expression. RefCOCO and RefCOCO+ distinguish person-centric testA and object-centric testB, while RefCOCOg uses longer, more descriptive expressions; we report the standard REC score on the released evaluation split for each dataset. The parenthesized values in Table 4 normalize each score by the matching Vanilla result. gRefCOCO extends this setting to expressions that refer to one object, multiple objects, or no target object (He et al., 2023). We use the released val, testA, and testB splits and report $\mathrm{Pr@}(F_1 = 1,\ \mathrm{GIoU} \geq 0.5)$, computed with the released GREC evaluation code; our protocol applies a generalized-IoU box-matching threshold in place of the $\mathrm{IoU} \geq 0.5$ threshold of the original metric definition.

Visual information preservation. VTC-Bench (Liao et al., 2026) measures whether a token-compressed model preserves visual information that is sensitive to image downsampling. Its fixed Qwen2-VL reference partitions each question into Group A (correct at full resolution but incorrect after downsampling) or Group B (correct in both conditions) for reference factors $d \in \{2, 3, 4, 5, 10\}$. We reuse these groups without modification, run Qwen2.5-VL-7B on the original images at $T = 256$ and $T = 128$, and report the official task scores, the group-wise mean, and the Group B–Group A gap. The reference image-area reductions (75.00%, 88.89%, 93.75%, 96.00%, and 99.00%) define the groups; they are not the pruning ratios of our target model.

A.3 Video Benchmarks

All video results use the same LLaVA-Video-7B input and decoding protocol described in Appendix C.3 (video input protocol). VideoMME (Fu et al., 2025) evaluates short, medium, and long videos across multiple visual domains; we use the no-subtitle setting and report the official aggregate percentage. MVBench (Li et al., 2024b) contains 20 multiple-choice tasks spanning spatial and temporal perception, action, object interaction, and episodic reasoning; we report the mean over its official task scores. LongVideoBench (Wu et al., 2024) tests long-context video-language understanding with interleaved video and text, including videos up to one hour; its official aggregate is returned as a fraction and is converted to a percentage for the tables. MLVU (Zhou et al., 2025) is a multi-task long-video suite covering diverse video genres and durations; we report its official macro-average over the seven task categories rather than pooling all questions. The four video scores are averaged with equal weight to form the video Acc.

A.4 Aggregation and Decoding Conventions

Unless a benchmark specifies a task-native metric above, scores are percentages and higher is better. For an image model, Acc is the unweighted mean of the benchmark tracks listed in its main table. MME contributes $\mathrm{MME}/20$ to this mean, while its native summed score remains visible in the MME column. Rel is the aggregate Acc divided by the corresponding Vanilla Acc and reported as a percentage. English and Chinese MMBench are separate terms in the mean; supplementary benchmarks and calibration examples never enter the main-table Acc unless explicitly stated in a table caption.

We use batch size one and deterministic greedy decoding without sampling. Each model family keeps its native processor, prompt format, image geometry, and official scorer. Empty answers and

generation-cap truncations remain in the relevant denominator. For video, we follow each benchmark's standard scoring protocol and then average the four resulting scores to obtain video Acc.

B Information-Theoretic Derivations

This appendix supplies the standard identities used in Section 3 and their application to our recovery model, followed by the optimization properties of MiCo's coverage proxy. The log-loss and erasure-channel identities are established results. MiCo applies them to the random semantic-query model defined in the main text.

Assumptions and scope. Throughout this appendix, Y and the semantic variables U_v are discrete, all relevant conditional entropies and losses are finite, and q, g are fixed. For each fixed retained set and recovery channel, the query index J is sampled independently of the source and recovered variables. The answer-side model assumes $Y \perp \widehat{\mathbf{U}}_V(S) \mid \mathbf{U}_V, q, g$. In the erasure model, the coverage and reliability gates are conditionally independent of each other and of U_v; reliability is held fixed with respect to S. The greedy result additionally fixes the proxy weights and similarities within one selector call.

B.1 Log-Loss, Bayes Risk, and Recoverable Information

Let π_{data} be the underlying data-generating joint distribution, including the answer Y, visual information Z, prompt Q, and geometric context G. Fix $Q = q$ and $G = g$, and use $\mid q, g$ as shorthand for conditioning on these values. A predictor supplies $\pi_{\mathrm{pred}}(\cdot \mid Z, q, g)$; the true conditional distribution $\pi_{\mathrm{data}}(\cdot \mid Z, q, g)$ is the analytical reference. All expectations and entropies below are under π_{data}, and the relevant losses and entropies are assumed finite.

For a discrete answer sequence $Y = (y_1, \ldots, y_M)$, including its end-of-sequence token, the negative log-likelihood is

$$\begin{aligned} \ell(\pi_{\mathrm{pred}}; Y, Z, q, g) &= -\log \pi_{\mathrm{pred}}(Y \mid Z, q, g) \\ &= -\sum_{t=1}^{M} \log \pi_{\mathrm{pred}}(y_t \mid y_{<t}, Z, q, g). \end{aligned} \tag{6}$$

Adding and subtracting the log of the true conditional distribution gives the standard entropy–divergence decomposition for logarithmic loss (Gneiting & Raftery, 2007, Example 3):

$$\begin{aligned} \mathcal{L}(\pi_{\mathrm{pred}}; Z, q, g) &= \mathbb{E}_{\pi_{\mathrm{data}}}[-\log \pi_{\mathrm{pred}}(Y \mid Z, q, g) \mid q, g] \\ &= \mathbb{E}_{\pi_{\mathrm{data}}}[-\log \pi_{\mathrm{data}}(Y \mid Z, q, g) \mid q, g] \\ &\quad + \mathbb{E}_{\pi_{\mathrm{data}}}\left[\log \frac{\pi_{\mathrm{data}}(Y \mid Z, q, g)}{\pi_{\mathrm{pred}}(Y \mid Z, q, g)} \,\middle|\, q, g\right] \\ &= H_{\pi_{\mathrm{data}}}(Y \mid Z, q, g) \\ &\quad + \mathbb{E}_{Z|q,g} D_{\mathrm{KL}}(\pi_{\mathrm{data}}(\cdot \mid Z, q, g) \,\|\, \pi_{\mathrm{pred}}(\cdot \mid Z, q, g)). \end{aligned} \tag{7}$$

The first term is the uncertainty remaining given the available information. The second is the predictor's discrepancy from the true conditional distribution. It is nonnegative and vanishes at $\pi_{\mathrm{pred}} = \pi_{\mathrm{data}}(\cdot \mid Z, q, g)$. Taking the infimum over all conditional predictive distributions yields the identity stated in Section 3:

$$R^*_Z(q, g) := \inf_{\pi_{\mathrm{pred}}} \mathcal{L}(\pi_{\mathrm{pred}}; Z, q, g) = H(Y \mid Z, q, g). \tag{8}$$

This infimum is an analytical reference, not an assumption that a fixed pretrained predictor attains it.

Substituting the complete visual source $\mathbf{U}_V$ and its reconstruction $\widehat{\mathbf{U}}_V(S)$ gives

$$\begin{aligned} R^*_{\mathrm{full}}(q, g) &= H(Y \mid \mathbf{U}_V, q, g), \\ R^*_S(q, g) &= H(Y \mid \widehat{\mathbf{U}}_V(S), q, g). \end{aligned} \tag{9}$$

Assume $Y \perp \widehat{\mathbf{U}}_V(S) \mid \mathbf{U}_V, q, g$. Applying the conditional-information identity for the benefit of side information under log-loss (Jiao et al., 2015, Corollary 1),

$$\begin{aligned}\Delta R^*(S; q, g) &= H(Y \mid \widehat{\mathbf{U}}_V(S), q, g) - H(Y \mid \mathbf{U}_V, q, g) \\ &= H(Y \mid \widehat{\mathbf{U}}_V(S), q, g) - H(Y \mid \mathbf{U}_V, \widehat{\mathbf{U}}_V(S), q, g) \\ &= I\Big(Y; \mathbf{U}_V \mid \widehat{\mathbf{U}}_V(S), q, g\Big) \\ &\le H\Big(\mathbf{U}_V \mid \widehat{\mathbf{U}}_V(S), q, g\Big) \\ &= H(\mathbf{U}_V \mid q, g) - I\Big(\mathbf{U}_V; \widehat{\mathbf{U}}_V(S) \mid q, g\Big).\end{aligned} \tag{10}$$

The second equality uses the conditional-independence assumption. The inequality follows because the semantic variables are discrete and conditional mutual information is at most the conditional entropy of either variable (Polyanskiy & Wu, 2025, Theorem 3.4). Since $H(\mathbf{U}_V \mid q, g)$ does not depend on S, maximizing the full-source recoverable mutual information minimizes this upper bound.

The same conditional mutual information is the expected KL divergence between the Bayes-optimal full and reconstructed-source answer distributions:

$$\begin{aligned}\Delta R^*(S; q, g) = \mathbb{E}_{\mathbf{U}_V, \widehat{\mathbf{U}}_V(S) \mid q, g} \\ D_{\mathrm{KL}}\Big(\pi_{\mathrm{data}}(\cdot \mid \mathbf{U}_V, q, g) \,\|\, \pi_{\mathrm{data}}(\cdot \mid \widehat{\mathbf{U}}_V(S), q, g)\Big).\end{aligned} \tag{11}$$

This identity concerns the true conditional distributions, not the outputs of an arbitrary fixed MLLM.

Task-weighted recovery and the answer-risk bound. The main text defines a semantic query by drawing J with $\Pr(J = v \mid q, g) = \Pr(J = v \mid Q = q)$, independently of $(\mathbf{U}_V, \widehat{\mathbf{U}}_V(S))$ given q, g. The retained set and recovery channel are fixed before J is drawn. Expanding the queried mutual information defined in equation 3 gives

$$\begin{aligned}\mathcal{I}_{\mathrm{task}}(S; q, g) &= I\Big(U_J; \widehat{U}_J(S) \mid J, q, g\Big) \\ &= \sum_{v \in V} \Pr(J = v \mid q, g) I\Big(U_v; \widehat{U}_v(S) \mid J = v, q, g\Big) \\ &= \sum_{v \in V} \Pr(J = v \mid Q = q) I\Big(U_v; \widehat{U}_v(S) \mid q, g\Big).\end{aligned} \tag{12}$$

The second equality expands the conditional mutual information over the values of J. The third equality uses $\Pr(J = v \mid q, g) = \Pr(J = v \mid Q = q)$ together with the conditional independence of J from $(\mathbf{U}_V, \widehat{\mathbf{U}}_V(S))$ given q, g, which lets the conditioning on $J = v$ be dropped. Thus the relevance weights arise from averaging over the queried source.

Both U_v and $\widehat{U}_v(S)$ are deterministic coordinate projections of their corresponding full vectors. Applying the data-processing inequality (Polyanskiy & Wu, 2025, Theorem 3.7) to each projection gives

$$\begin{aligned}I\Big(U_v; \widehat{U}_v(S) \mid q, g\Big) &\le I\Big(\mathbf{U}_V; \widehat{U}_v(S) \mid q, g\Big) \\ &\le I\Big(\mathbf{U}_V; \widehat{\mathbf{U}}_V(S) \mid q, g\Big).\end{aligned} \tag{13}$$

The query probabilities are nonnegative and sum to one, so

$$\begin{aligned}\mathcal{I}_{\mathrm{task}}(S; q, g) &\le \sum_{v \in V} \Pr(J = v \mid Q = q) I\Big(\mathbf{U}_V; \widehat{\mathbf{U}}_V(S) \mid q, g\Big) \\ &= I\Big(\mathbf{U}_V; \widehat{\mathbf{U}}_V(S) \mid q, g\Big).\end{aligned} \tag{14}$$

This bounds the queried mutual information by whole-source recovered information without requiring independence among the source variables U_v. Combining it with equation 10 yields

$$\begin{aligned}\Delta R^*(S; q, g) &\le H(\mathbf{U}_V \mid q, g) - I\Big(\mathbf{U}_V; \widehat{\mathbf{U}}_V(S) \mid q, g\Big) \\ &\le H(\mathbf{U}_V \mid q, g) - \mathcal{I}_{\mathrm{task}}(S; q, g),\end{aligned} \tag{15}$$

which establishes equation 1. With the query distribution fixed, the entropy term is independent of S, so maximizing $\mathcal{I}_{\text{task}}$ minimizes this looser upper bound. The query distribution is a modeling choice; a retained set that minimizes this bound need not minimize the answer risk itself.

Log-loss interpretation of the semantic query. A predictor given J, q, g has prior Bayes risk $R^*_{\text{query,prior}}(q, g) = H(U_J \mid J, q, g)$. Observing the recovered unit reduces its risk to

$$\begin{aligned} R^*_{\text{query,recovered}}(S; q, g) &= H(U_J \mid \widehat{U}_J(S), J, q, g) \\ &= \sum_{v \in V} \Pr(J = v \mid Q = q) H(U_v \mid q, g) - \mathcal{I}_{\text{task}}(S; q, g). \end{aligned} \tag{16}$$

The first term is the prior Bayes risk $R^*_{\text{query,prior}}(q, g)$ and is independent of S. Maximizing $\mathcal{I}_{\text{task}}$ therefore minimizes the Bayes log-loss of the modeled semantic query. Its prediction target is U_J, rather than the answer Y. For the answer task, equation 14 and equation 1 instead establish that this criterion minimizes a looser upper bound on excess Bayes risk, not the answer loss of a fixed pretrained MLLM.

B.2 Semantic Erasure with Coverage and Reliability Gates

Fix the prompt $Q = q$, geometric context $G = g$, and retained set S. The coverage probability is $M_v(S; q, g) = \Pr(E_v = 1 \mid q, g)$, and the representation reliability is $\Pr(C_v = 1 \mid q, g)$. For a fixed recovery channel, we treat this reliability probability as invariant to S. As stated in Section 3, C_v and E_v are conditionally independent, and the pair (C_v, E_v) is independent of U_v given q, g. These are assumptions of the semantic erasure model.

The event $C_v E_v = 1$ is exactly the event that both binary gates equal one. Conditional independence therefore gives

$$\begin{aligned} \Pr(C_v E_v = 1 \mid q, g) &= \Pr(C_v = 1, E_v = 1 \mid q, g) \\ &= \Pr(C_v = 1 \mid q, g) \Pr(E_v = 1 \mid q, g) \\ &= \Pr(C_v = 1 \mid q, g) M_v(S; q, g), \\ \Pr(C_v E_v = 0 \mid q, g) &= 1 - \Pr(C_v = 1 \mid q, g) M_v(S; q, g). \end{aligned} \tag{17}$$

Let $R_v = C_v E_v$. The erasure symbol $\perp$ lies outside the alphabet of U_v, so observing $\widehat{U}_v(S)$ also identifies R_v. If $R_v = 1$, the output reveals U_v exactly. If $R_v = 0$, the output is $\perp$ and the independence assumption leaves uncertainty $H(U_v \mid q, g)$. Thus

$$\begin{aligned} H\Big(U_v \mid \widehat{U}_v(S), q, g\Big) &= \Pr(R_v = 1 \mid q, g) \cdot 0 \\ &\quad + \Pr(R_v = 0 \mid q, g) H(U_v \mid q, g) \\ &= \left[1 - \Pr(C_v = 1 \mid q, g) M_v(S; q, g)\right] H(U_v \mid q, g). \end{aligned} \tag{18}$$

Subtracting this residual entropy yields

$$\begin{aligned} I\Big(U_v; \widehat{U}_v(S) \mid q, g\Big) &= H(U_v \mid q, g) - H\Big(U_v \mid \widehat{U}_v(S), q, g\Big) \\ &= H(U_v \mid q, g) \Pr(C_v = 1 \mid q, g) M_v(S; q, g), \end{aligned} \tag{19}$$

which is the standard erasure-channel mutual-information identity (Polyanskiy & Wu, 2025, Example 33.6) with success probability $\Pr(C_v = 1 \mid q, g) M_v(S; q, g)$. Substituting it into the random-query expansion in equation 12 gives equation 3. The argument requires no independence among the source variables U_v. The coverage probability remains abstract in this derivation; Section 3.2 specifies MiCo's coverage model and observable proxies.

B.3 Coverage Properties and Greedy Approximation

The following result applies the standard weighted facility-location and greedy-maximization arguments (Krause & Golovin, 2014; Nemhauser et al., 1978) to MiCo's observable objective. Fix one selector call on a finite token set V with features $\{x_v\}_{v \in V}$. Its source weights $\widehat{W}_v := \widehat{\mathcal{H}}_v \widehat{\mathcal{R}}_v \widehat{\mathcal{P}}_v \geq 0$

(Eq. 4) and similarities $\kappa_X(v,i) = [\cos(x_v, x_i)]_+ \in [0,1]$ remain fixed as the retained set grows. The objective is

$$\widehat{\mathcal{I}}_{\text{task}}(S) = \sum_{v \in V} \widehat{W}_v \widehat{M}_v(S), \qquad \widehat{M}_v(S) = \max_{s \in S} \kappa_X(v,s), \tag{20}$$

with $\widehat{M}_v(\varnothing) = 0$.

Proposition 1 (Coverage properties and greedy guarantee)**.** *The objective in equation 20 is normalized, nonnegative, monotone, and submodular. Let $b = \min(B, |V|)$. If $b = 0$, the empty set is optimal. For $b \geq 1$, let $S^\star$ maximize this objective over $S \subseteq V$ with $|S| \leq b$, and let S_b be the set returned after b exact greedy additions starting from the empty set. Then*

$$\begin{aligned} \widehat{\mathcal{I}}_{\text{task}}(S_b) &\geq \left[1 - \left(1 - \frac{1}{b}\right)^b\right] \widehat{\mathcal{I}}_{\text{task}}(S^\star) \\ &\geq (1 - 1/e)\widehat{\mathcal{I}}_{\text{task}}(S^\star). \end{aligned} \tag{21}$$

Proof. Normalization follows from $\widehat{M}_v(\varnothing) = 0$. Nonnegative weights and $0 \leq \widehat{M}_v(S) \leq 1$ give

$$0 \leq \widehat{\mathcal{I}}_{\text{task}}(S) \leq \sum_{v \in V} \widehat{W}_v.$$

For $A \subseteq D \subseteq V$, taking a maximum over the larger set gives $\widehat{M}_v(A) \leq \widehat{M}_v(D)$, so the objective is monotone. To establish submodularity, consider an unselected token $i \in V \setminus D$. Its marginal gain is

$$\begin{aligned} \Delta\widehat{\mathcal{I}}_{\text{task}}(i \mid A) &= \sum_{v \in V} \widehat{W}_v [\kappa_X(v,i) - \widehat{M}_v(A)]_+ \\ &\geq \sum_{v \in V} \widehat{W}_v [\kappa_X(v,i) - \widehat{M}_v(D)]_+ \\ &= \Delta\widehat{\mathcal{I}}_{\text{task}}(i \mid D). \end{aligned} \tag{22}$$

This is the diminishing-returns definition of submodularity. Multiplication by fixed nonnegative source weights preserves the inequality.

For the approximation bound, let S_t denote the first t greedy selections, with $S_0 = \varnothing$ and $0 \leq t < b$. Monotonicity, diminishing returns, and the greedy choice imply

$$\begin{aligned} \widehat{\mathcal{I}}_{\text{task}}(S^\star) - \widehat{\mathcal{I}}_{\text{task}}(S_t) &\leq \widehat{\mathcal{I}}_{\text{task}}(S_t \cup S^\star) - \widehat{\mathcal{I}}_{\text{task}}(S_t) \\ &\leq \sum_{i \in S^\star \setminus S_t} \Delta\widehat{\mathcal{I}}_{\text{task}}(i \mid S_t) \\ &\leq b\big[\widehat{\mathcal{I}}_{\text{task}}(S_{t+1}) - \widehat{\mathcal{I}}_{\text{task}}(S_t)\big]. \end{aligned} \tag{23}$$

Writing the remaining objective gap as $D_t = \widehat{\mathcal{I}}_{\text{task}}(S^\star) - \widehat{\mathcal{I}}_{\text{task}}(S_t)$ gives $D_{t+1} \leq (1 - 1/b)D_t$. Since $D_0 = \widehat{\mathcal{I}}_{\text{task}}(S^\star)$, iterating for b steps proves the first inequality in equation 21. The second follows from $(1 - 1/b)^b \leq e^{-1}$. □

Greedy selection attains at least a $(1 - 1/e)$ fraction of the optimal coverage value. Equivalently, the objective gap is at most $\widehat{\mathcal{I}}_{\text{task}}(S^\star)/e$. It is not a bound on answer loss or benchmark error. The proof requires nonnegative similarities, not a positive-semidefinite similarity matrix.

The guarantee is local to each fixed selector call. In Stage 2, the comparison optimum uses the candidate pool and contextualized features produced after Stage 1. When crops or tiles receive separate budgets, it applies within each group with its assigned budget. It does not establish the same approximation ratio for joint optimization of both stages or for unrestricted token allocation across groups. If $B \geq |V|$, all available tokens are retained and the solution is exact.

Algorithm 1 MiCo

Require: Visual features $X = \{x_v\}_{v \in V}$; nonnegative proxies $\widehat{\mathcal{H}}, \widehat{\mathcal{R}}, \widehat{\mathcal{P}}$; retained budget B
Ensure: Retained visual-token indices S in their original order

$\widehat{W}_v \leftarrow \widehat{\mathcal{H}}_v \widehat{\mathcal{R}}_v \widehat{\mathcal{P}}_v$ for every $v \in V$
$\kappa(v, i) \leftarrow [\cos(x_v, x_i)]_+$ for all $v, i \in V$
$S \leftarrow \varnothing$; $\widehat{M}_v \leftarrow 0$ for every $v \in V$
for $t = 1$ to $\min(B, |V|)$ **do**
  $i^\star \leftarrow \arg\max_{i \in V \setminus S} \sum_{v \in V} \widehat{W}_v [\kappa(v, i) - \widehat{M}_v]_+$
  $S \leftarrow S \cup \{i^\star\}$
  $\widehat{M}_v \leftarrow \max\{\widehat{M}_v, \kappa(v, i^\star)\}$ for every $v \in V$
end for
return S sorted by the original token order

C Implementation and Evaluation Details

C.1 Shared MiCo Selector

Both pruning stages use Algorithm 1. Its inputs are the visual features at the current stage, three nonnegative proxy vectors, and a retained-token budget. In both stages, coverage is the positive cosine similarity between visual tokens in that stage's feature space. Text-to-visual attention affects the Stage-2 weights, not the coverage kernel. Feature magnitudes are measured before the normalization used to compute cosine similarity.

The proxy-based version of equation 3 is $\widehat{\mathcal{I}}_{\text{task}}(S; X) = \sum_{v \in V} \widehat{W}_v \widehat{M}_v(S)$, where $\widehat{W}_v = \widehat{\mathcal{H}}_v \widehat{\mathcal{R}}_v \widehat{\mathcal{P}}_v$, $\widehat{M}_v(S) = \max_{s \in S} \kappa_X(v, s)$, and $\widehat{M}_v(\varnothing) = 0$. The marginal gain used to select the next token is

$$\Delta\widehat{\mathcal{I}}_{\text{task}}(i \mid S; X) = \sum_{v \in V} \widehat{W}_v [\kappa_X(v, i) - \widehat{M}_v(S)]_+. \tag{24}$$

The fixed-call assumptions, submodularity proof, and approximation guarantee are given in Appendix B.3; this section focuses on the executable selector.

Table 8 summarizes the stage-specific proxy assignments. The features x_v^{vis} and $x_v^{(\ell_\star)}$ denote visual-encoder outputs and contextualized visual states, respectively. Here $\overline{A}^{\text{vis}}_{\text{CLS} \to v}$ averages visual CLS attention over heads, and $\overline{A}^{(\ell_\star)}_{\text{text} \to v}$ averages text-to-visual attention over available text queries and heads. For vision encoders without a CLS token, the architecture-native pooled visual-attention analogue replaces CLS attention. Setting a proxy to one gives that factor uniform weight; feature norms and attention are proxies rather than direct measurements of semantic entropy or reliability.

Table 8: Stage-specific inputs to MiCo's shared weighted-coverage selector.

Stage	Energy $\widehat{\mathcal{H}}_v$	Reliability $\widehat{\mathcal{R}}_v$	Task relevance $\widehat{\mathcal{P}}_v$
Stage 1	$\|x_v^{\text{vis}}\|_2$	$\overline{A}^{\text{vis}}_{\text{CLS} \to v}$	1
Stage 2	$\|x_v^{(\ell_\star)}\|_2$	1	$\overline{A}^{(\ell_\star)}_{\text{text} \to v}$

C.2 Empirical Checks for the Proxy Assignments

The theoretical factors are latent, so these experiments validate the observable proxies against operational references rather than claiming to measure the factors exactly. The checks below are diagnostic evidence for the intended role of each proxy; they do not change the selector or its weights.

Stage-1 information amount. We use dense COCO-Stuff labels to construct an empirical entropy reference. The image pool is split into disjoint A/B halves of 2,500 COCO val2017 images. For each aspect bin a and native visual position v, A supplies category counts $n_{a,v,c}$ over the 182 stored

COCO-Stuff classes. With the fixed Jeffreys smoothing used in the analysis,

$$\widehat{p}_{a,v,c} = \frac{n_{a,v,c} + 0.5}{N_{a,v} + 0.5 \times 182}, \qquad \widehat{H}_{a,v} = -\sum_{c=1}^{182} \widehat{p}_{a,v,c} \log \widehat{p}_{a,v,c},$$

where $N_{a,v} = \sum_c n_{a,v,c}$ excludes the dataset's void label. On the independent B pool, we extract the raw pre-projection Stage-1 feature $x_{i,a,v}$ and compare $\widehat{H}_{a,v}$ with its position-wise mean norm $\overline{n}_{a,v} = |B_{a,v}|^{-1} \sum_{i \in B_{a,v}} \|x_{i,a,v}\|_2$. Figure 6 shows the two-sided 1% trimmed fit, which gives Spearman $\rho = 0.36832$; the pre-specified primary estimate is the untrimmed $\rho = 0.35943$ (Tukey $1.5 \times$ IQR trimming: 0.31021). Thus the evidence is a moderate position-level association for the Stage-1 norm, not a token-level identity with $H(U_v \mid q, g)$: the corresponding per-token association is $\rho = 0.00699$, and this diagnostic contains no prompt variable. It therefore estimates an $H(U_v \mid G)$-type quantity and is used only to support the task-agnostic Stage-1 proxy.

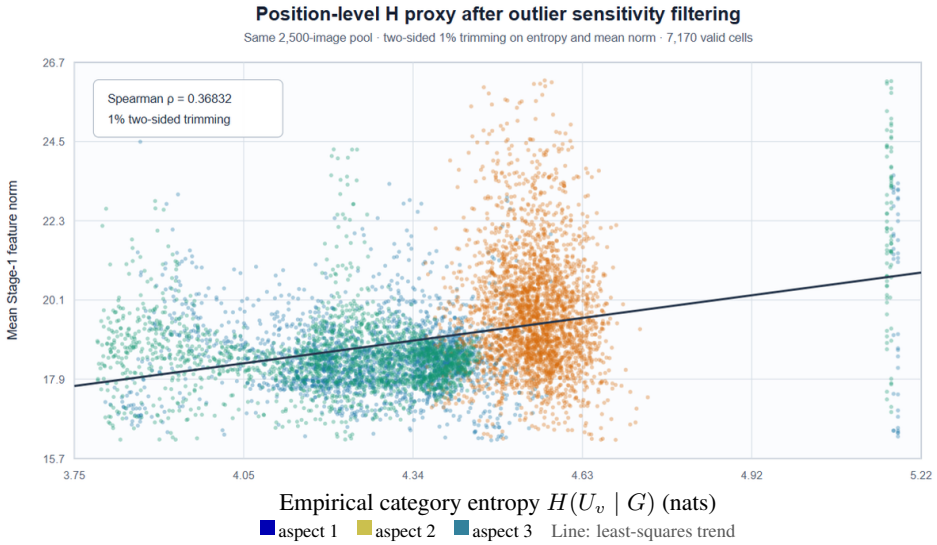


Figure 6: Position-level validation of the Stage-1 information-amount proxy. Dense-label entropy from COCO-Stuff is compared with the mean raw Stage-1 feature norm at matched aspect bin and native position. The displayed fit uses a two-sided 1% trim ($\rho = 0.36832$); the pre-specified primary estimate is untrimmed ($\rho = 0.35943$). This is a geometry-conditioned $H(U_v \mid G)$ diagnostic rather than a direct measurement of prompt-conditioned entropy.

Stage-1 representation reliability. We use annotated-category decoding as an operational reference for whether a visual representation preserves semantic content. On LLaVA-NeXT-7B, we train linear and MLP probes on 1,688 clean COCO objects, select checkpoints on 382 validation objects, and test on 200 objects from new, disjoint images. Each probe receives the unit-normalized mean visual feature within the target region; neither CLS attention nor feature norm is supplied as an extra input. We use three fixed head seeds and compare target-region masking with equal-area context masking. Decoding correctness is an observable diagnostic, not a direct observation of the latent reliability gate C_v.

Figure 7 provides three complementary checks. Panel (a) shows that target masking reduces category accuracy from 98.0% to 81.4% at 75% replacement, whereas equal-area context masking leaves it at 97.9%; full target masking reduces accuracy to 64.3%. This establishes that the decoding reference is sensitive to target content. Panels (b) and (c) then test whether the reliability proxy orders successful and unsuccessful decoding trials. Within each fixed object, masking arm, strength, and head seed, concordance is the fraction of correct–incorrect pairs whose CLS-attention score is higher on the correct trial, with half credit for ties. Scores average CLS-to-patch attention over heads and the fixed target region; feature norm is a comparison score averaged over the same region. Concordance is computed from six perturbation repeats, then averaged across defined seeds, strengths, and images, with no-association reference 0.5. All-correct and all-incorrect strata have undefined concordance and are not assigned a chance score.

For target masking, CLS-attention concordance is 0.557 (95% CI [0.479, 0.629]) with the linear probe and 0.564 ([0.484, 0.636]) with the MLP, compared with feature-norm estimates of 0.459 and 0.499. Thus the positive attention trend persists under both probe capacities, providing suggestive support for its reliability role. Both intervals include 0.5, so this is not a statistically established above-chance association. Target concordance uses 57 and 59 informative images, respectively; context concordance uses only 4 and 3 and is correspondingly uncertain. Confidence intervals use 1,000 image-cluster bootstrap resamples; all 200 images contribute to decoding accuracy.

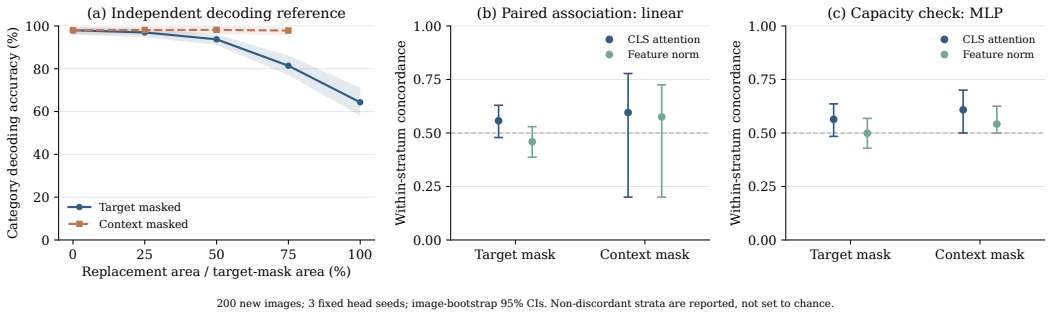


Figure 7: Three complementary diagnostics for Stage-1 reliability on LLaVA-NeXT-7B. (a) Category decoding is sensitive to target masking, but largely stable under context masking. (b) CLS attention has a positive concordance point estimate with decoding success under a linear probe. (c) The same direction persists with an MLP probe. Shading and error bars show image-bootstrap 95% confidence intervals over 200 test images and three fixed head seeds; concordance excludes strata without both successes and failures. The attention trends are suggestive, not conclusive, because the target-mask intervals include the 0.5 reference.

Stage-2 task relevance. For each held-out RefCOCO image–query pair, we convert the annotated box–cell overlap into a normalized target distribution over the Stage-1 survivor cells. This is an operational query-relevance target conditioned on the image and survivor pool, not a direct observation of J. We compare it with the text-to-visual attention averaged over text queries and heads, without post-hoc calibration. Figure 8 shows the local attention-mass interval $[0, 0.08]$, using the original rank-decile bins without rebinning. Across the complete held-out set of 480 queries from 240 images, the mean per-query Spearman correlation is $\rho = 0.0697$ (image-bootstrap 95% CI $[0.0433, 0.0976]$), indicating a weak positive spatial association. The higher-attention bins outside this view have decreasing target mass, so the local trend does not establish full-range probability calibration.

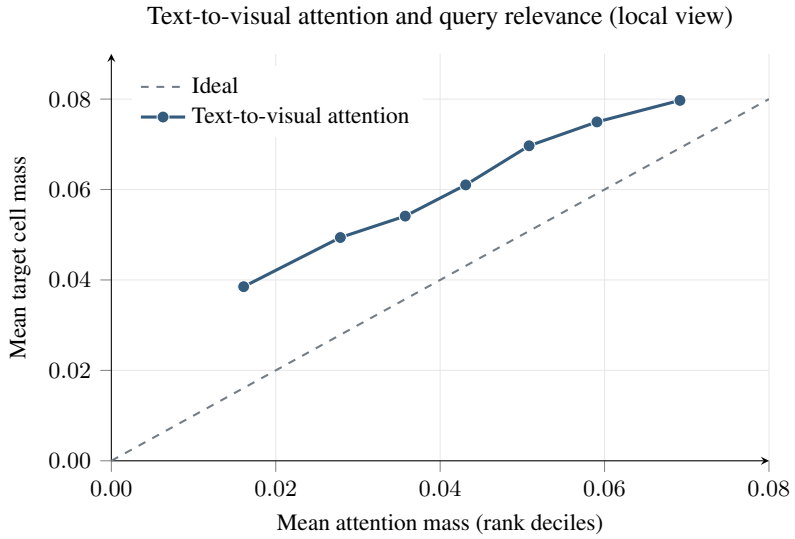


Figure 8: Local-range view of the Stage-2 task-relevance diagnostic on LLaVA-NeXT-7B, restricted to attention masses 0–0.08. Points retain the original rank-decile means; the dashed line is the ideal calibration reference. The target is normalized RefCOCO box–cell overlap conditioned on the Stage-1 survivor pool. Higher-attention bins are outside the displayed range; the correlation reported in the text uses the complete held-out data.

C.3 Architecture and Budget Mapping

Let N_0 be the original visual-token count and T the target layer-average count. Unless otherwise stated, Stage 1 retains a pool of $N_1 = \min(2T, N_0)$ tokens. With L decoder blocks, refinement uses the text-to-visual attention of the one-based block $K = \ell_\star$ and prunes immediately after it, where $1 \le \ell_\star < L$, so that the first $\ell_\star$ blocks process N_1 visual tokens and the remaining $L - \ell_\star$ blocks process R. Before integer rounding and clipping, matching the target average gives

$$T = \frac{\ell_\star N_1 + (L - \ell_\star)R}{L}. \tag{25}$$

Solving for the post-refinement count and enforcing integer and range constraints yields

$$R = \mathrm{clip}\left(\mathrm{round}\,\frac{TL - \ell_\star N_1}{L - \ell_\star}, 1, N_1\right). \tag{26}$$

Rounding and clipping can make the realized layer average differ from T. The selector can be applied directly to budget T when decoder-stage refinement is omitted. Table 9 gives the architecture-specific schedules. Its listed R assumes the pool $N_1 = 2T$ before integer rounding; video budgets are per frame.

Table 9: MiCo budgets and refinement layers by architecture; video budgets are per frame.

Backbone	Decoder blocks L	Refinement $\ell_\star$	Stage-1 keep	Stage-2 keep R
LLaVA-1.5/NeXT-7B	32	7	$2T$	$\mathrm{round}(18T/25)$
LLaVA-1.5-13B	40	8	$2T$	$\mathrm{round}(3T/4)$
LLaVA-NeXT-13B	40	14	$2T$	$\mathrm{round}(6T/13)$
Qwen2.5-VL-7B	28	2	$2T$	$\mathrm{round}(12T/13)$
InternVL3-8B	28	4	$2T$	$\mathrm{round}(5T/6)$
Qwen3-VL-8B	36	6	$2T$	$\mathrm{round}(4T/5)$
Qwen3.5-9B	32	12	$2T$	$\mathrm{round}_{\mathrm{even}}(2T/5)$
LLaVA-Video-7B	28	6	$2T$	$\mathrm{round}(8T/11)$

For multi-crop LLaVA-NeXT inputs and dynamically tiled InternVL inputs, the Stage 1 budget is distributed over the natural visual groups. Integer division and a minimum of one token per group can make the realized count differ slightly from T or $2T$. LLaVA, Qwen3-VL, and Qwen3.5

recompute R using equation 26 and the realized Stage 1 survivor count. For Qwen2.5-VL and InternVL, R is instead computed from the nominal $N_1 = 2T$ rather than the realized survivor count, so their realized layer average can differ slightly when clipping or group rounding prevents exactly $2T$ survivors.

Video input protocol. LLaVA-Video-7B is evaluated through lmms-eval with the video repository's native task prompts, generation settings, and official scorers. Each video is uniformly sampled to at most 64 frames (`force_sample=False`, so shorter videos contribute fewer frames), and every frame is encoded by the SigLIP vision tower and spatially pooled by the model's native pooling to 169 visual tokens, so an unpruned 64-frame input carries 64×169 visual tokens. Video budgets are nominal per-frame layer averages: $T = 64/32/16$ correspond to pruning 62.1%/81.1%/90.5% of the 169 tokens per frame. MiCo applies the Stage 1 pool ($2T$ per frame) and the refinement mapping of equation 26 to the concatenated frame tokens, and Table 12 reports the realized per-frame layer averages. Efficiency measurements fix 64 frames at 384×384 resolution with FP16 and SDPA on one A800 80GB GPU (Appendix D).

C.4 Computational Complexity

Following the standard accounting for token-pruning selectors (Zhang et al., 2024b), constructing a dense visual-token similarity matrix for N tokens of width d costs $O(N^2 d)$ time and $O(N^2)$ memory. Exact greedy coverage with a retained budget B adds $O(BN^2)$ arithmetic for the marginal updates. Thus, for the two-stage path with original token count N_0, Stage 1 pool N_1, and final budget R, the selector-specific work scales as $O(N_0^2 d_1 + N_1 N_0^2 + N_1^2 d_2 + R N_1^2)$, up to implementation constants, where d_1 and d_2 are the widths of the vision-encoder features and the decoder hidden states. For video inputs, Stage 1 is applied to each frame separately, so the first two terms use the per-frame token count (169 for LLaVA-Video-7B) and grow only linearly with the number of frames, whereas Stage 2 operates on the concatenated survivor pool $N_1 = F \cdot 2T$ over all F frames; the quadratic Stage 2 terms therefore dominate the selector cost for long videos (Appendix D).

We validate this scaling on the executed implementations by counting logical dense-equivalent FLOPs, including the vision encoder, projector, decoder, output head, and both selector stages. These counts are an operation-accounting check rather than hardware-instruction counts; the per-component breakdown and synchronized latency validation are given in Appendix D and Table 13.

C.5 Qwen3.5-9B: Hybrid-Decoder Adaptation

Qwen3.5-9B (Qwen Team, 2026) has 32 decoder blocks, repeating three Gated DeltaNet blocks followed by one full-attention block. We retain all decoder blocks and remove only visual tokens. Both evaluated budgets ($T = 256$ and $T = 128$) use the same configuration: a Stage 1 pool of $N_1 = \min(2T, N_0)$ tokens (multiplier $n = 2$, as for the other models), refinement after the 1-based block $K = 12$, which is a full-attention block, and full-attention QK relevance averaged over all non-visual prompt positions.

Stage 1 follows the CLS-free Qwen adaptation. We observe the final vision-block attention, pool its query and key axes over the native 2×2 merger groups, and average over query groups and heads. The coverage features are the means of the four pre-merger patch features in each group. Their norms multiply the pooled attention to form source weights for the shared greedy coverage selector. The selected group indices retain the corresponding native merged embeddings; the vision encoder and merger execute unchanged, and selection occurs before the first decoder block.

Stage 2 uses block 12's native Q/K projections, Q/K normalization, rotary positions, and grouped-query key expansion. For each non-visual prompt query and head, we normalize QK scores over the current visual keys, then average the probabilities over queries and heads. The block executes normally. The block's output visual hidden states serve as the Stage 2 coverage features, and their ℓ_2 norms multiplied by these relevance scores give the Stage 2 source weights; the weights are divided by their sum (uniform weights are used if the sum is zero), which does not change the greedy selection. Stage 1 attention weights are not multiplied in again. We keep only the selected visual states, keep all text states, retain each surviving token's original multimodal (mRoPE) position index, and run blocks 13–32 on the shortened sequence.

The second-stage budget is the instance of equation 26 with $L = 32$ and $K = 12$, $R = \mathrm{round}\left((32T - 12N_1)/20\right)$ with $N_1 = \min(2T, N_0)$, where round is round-half-to-even; instead of clipping, the Qwen3.5 path raises an error if R falls outside $[1, N_1]$. For $N_0 \geq 2T$, the schedules are $512 \to 102$ at $T = 256$ and $256 \to 51$ at $T = 128$, realizing layer averages of 255.750 and 127.875. During cached decoding each block keeps the prefill state it computed: blocks 1–12 keep KV caches (full-attention blocks) or recurrent states (Gated DeltaNet blocks) built from the full N_1-token sequence and are not truncated after pruning, whereas blocks 13–32 hold caches built from the pruned sequence. Generated tokens receive position indices that continue from the original, unpruned prompt length.

The evaluated scope is one unpadded image per request, batch size one, SDPA, greedy cached generation, thinking disabled, and a 2,048-token generation cap. Empty and capped answers remain in the scoring denominator. Each configuration has 23,715 predictions: AI2D 3,088; POPE 5,127 expanded to 9,000 category entries; HallusionBench 951; MME 2,374 questions (1,187 images, two questions each); MMBench EN/CN 4,329 entries and 1,164 Circular groups each; MMStar 1,500; and ScienceQA-IMG 2,017. MMStar's released questions are used unchanged, including items whose answer options are malformed in the original release.

As with Qwen2.5-VL and Qwen3-VL, we use one fixed configuration at both budgets (here $n = 2$, $K = 12$; per-model layers are listed in Table 19) and report the resulting eight-task evaluation. Acc follows the common Qwen convention (MME$/20$).

D Efficiency Measurement Details

D.1 Measurement Protocol

The main text reports one-token latency for the LLaVA-NeXT backbones and complete-answer latency for LLaVA-Video, together with peak allocated memory and logical FLOPs. Latency uses synchronized CUDA wall time after ten warm-up requests; FLOPs are counted in a separate pass and are not used as a latency proxy. Loading, preprocessing, and host-to-device transfer are outside the generation timer. Budgets are nominal layer-average targets; the token counts realized by the executed decoder schedules are reported for the MiCo breakdowns in Tables 11 and 12.

The NeXT-7B MiCo rows use the default MiCo implementation, which fuses the greedy selector and runs the language model with FlashAttention2; the backend changes kernel fusion and memory traffic, but not the selector objective, token budget, or selected tokens. We report logical dense-equivalent FLOPs rather than hardware instruction counts. The accounting includes the executed vision path, projector, decoder, output head, and pruning stages; fused attention is counted by its dense-equivalent arithmetic, while indexing, sorting, synchronization, and diagnostic-only work are excluded consistently.

D.2 Backbone-level efficiency results

LLaVA-NeXT backbones. Table 10 reports median latency, peak-memory, and logical-FLOP comparisons at $T = 640, 320, 160$ for NeXT-7B and NeXT-13B under one protocol (one A800 80GB GPU, batch size one, one generated token, ten warm-ups and 50 measured requests). The two backbones are placed side by side for direct comparison; the Pareto figures use arithmetic-mean timing, so their speedups differ slightly from median-based ratios. Under the layer-average accounting of Appendix C.3, FastV and SparseVLM have no feasible $T = 160$ schedule on NeXT-7B and only a degenerate one on NeXT-13B (Appendix G), so their $T = 160$ rows are left blank.

Table 10: LLaVA-NeXT-7B and LLaVA-NeXT-13B efficiency comparison (median latency; speedups and reductions are relative to each backbone's Vanilla row).

Method	LLaVA-NeXT-7B			LLaVA-NeXT-13B		
	Latency (ms) (Speedup)	Peak memory (GiB) (Reduction)	Logical FLOPs (T) (Reduction)	Latency (ms) (Speedup)	Peak memory (GiB) (Reduction)	Logical FLOPs (T) (Reduction)
Vanilla	290.0 (1.00×)	15.8 (0.0%)	45.6 (0.0%)	484.1 (1.00×)	28.4 (0.0%)	85.1 (0.0%)

Continued on next page

Table 10 (continued): LLaVA-NeXT-7B and LLaVA-NeXT-13B efficiency comparison

Method	LLaVA-NeXT-7B			LLaVA-NeXT-13B		
	Latency (ms) (Speedup)	Peak memory (GiB) (Reduction)	Logical FLOPs (T) (Reduction)	Latency (ms) (Speedup)	Peak memory (GiB) (Reduction)	Logical FLOPs (T) (Reduction)
Budget $T = 640$ tokens per image						
FastV	113.0 (2.57×)	16.0 (-1.7%)	11.8 (74.1%)	154.9 (3.13×)	28.1 (1.1%)	20.9 (75.4%)
SparseVLM	137.1 (2.11×)	16.0 (-1.7%)	11.8 (74.2%)	176.3 (2.75×)	28.1 (1.1%)	21.0 (75.3%)
DivPrune	108.7 (2.67×)	14.2 (9.7%)	11.7 (74.4%)	**154.2** (3.14×)	**25.9** (8.8%)	20.8 (75.6%)
VisionZip	118.8 (2.44×)	15.2 (3.9%)	11.7 (74.4%)	173.6 (2.79×)	27.0 (4.9%)	20.8 (75.6%)
VisPruner	128.0 (2.27×)	14.7 (7.1%)	11.7 (74.4%)	179.2 (2.70×)	26.5 (6.7%)	20.8 (75.6%)
HoloV	117.4 (2.47×)	15.2 (3.9%)	11.7 (74.3%)	176.1 (2.75×)	27.0 (4.9%)	20.8 (75.6%)
MMTok	208.3 (1.39×)	16.1 (-2.2%)	11.7 (74.4%)	259.2 (1.87×)	28.0 (1.4%)	20.8 (75.6%)
ApET	125.4 (2.31×)	23.3 (-48.0%)	10.5 (77.1%)	185.9 (2.60×)	37.8 (-33.1%)	**18.7** (78.0%)
MiCo	135.5 (2.14×)	13.8 (12.8%)	11.6 (74.5%)	222.4 (2.18×)	27.0 (4.9%)	20.8 (75.6%)
Budget $T = 320$ tokens per image						
FastV	107.0 (2.71×)	16.0 (-1.7%)	7.4 (83.8%)	128.1 (3.78×)	28.5 (-0.4%)	12.5 (85.3%)
SparseVLM	130.4 (2.22×)	19.0 (-20.1%)	7.4 (83.8%)	173.6 (2.79×)	32.0 (-12.7%)	12.3 (85.5%)
DivPrune	86.5 (3.35×)	14.0 (11.1%)	7.3 (84.0%)	**112.0** (4.32×)	**26.0** (8.5%)	12.3 (85.5%)
VisionZip	102.4 (2.83×)	15.2 (3.9%)	7.3 (84.1%)	126.5 (3.83×)	27.0 (4.9%)	12.2 (85.7%)
VisPruner	112.4 (2.58×)	14.7 (7.1%)	7.3 (84.0%)	139.5 (3.47×)	26.5 (6.7%)	12.3 (85.5%)
HoloV	102.6 (2.83×)	15.2 (3.9%)	7.3 (84.1%)	125.8 (3.85×)	27.0 (4.9%)	12.2 (85.7%)
MMTok	151.9 (1.91×)	16.1 (-2.2%)	7.3 (84.0%)	181.4 (2.67×)	28.0 (1.4%)	12.3 (85.5%)
ApET	103.4 (2.80×)	16.0 (-1.2%)	6.8 (85.0%)	124.7 (3.88×)	28.5 (-0.4%)	**11.5** (86.5%)
MiCo	99.4 (2.92×)	13.5 (14.1%)	7.2 (84.3%)	151.1 (3.20×)	27.0 (4.9%)	12.2 (85.7%)
Budget $T = 160$ tokens per image						
FastV	–	–	–	–	–	–
SparseVLM	–	–	–	–	–	–
DivPrune	81.9 (3.54×)	14.0 (11.2%)	5.1 (88.8%)	**92.8** (5.22×)	**25.9** (8.8%)	8.1 (90.5%)
VisionZip	97.1 (2.99×)	15.2 (3.9%)	5.1 (88.8%)	114.1 (4.24×)	27.0 (4.9%)	8.0 (90.6%)
VisPruner	113.1 (2.56×)	14.7 (7.1%)	5.1 (88.8%)	124.6 (3.89×)	26.5 (6.7%)	8.1 (90.5%)
HoloV	101.1 (2.87×)	15.2 (3.9%)	5.1 (88.8%)	114.1 (4.24×)	27.0 (4.9%)	8.0 (90.6%)
MMTok	131.9 (2.20×)	16.1 (-2.2%)	5.1 (88.8%)	144.3 (3.35×)	28.0 (1.4%)	8.1 (90.5%)
ApET	91.4 (3.17×)	14.0 (11.1%)	5.0 (89.0%)	108.4 (4.47×)	26.1 (8.1%)	**7.9** (90.7%)
MiCo	88.1 (3.29×)	13.4 (14.8%)	5.0 (89.0%)	123.6 (3.92×)	27.0 (4.9%)	8.0 (90.6%)

D.3 Component-level breakdown

Component timing. Tables 11 and 12 report the NeXT-7B and Video-7B timing breakdowns with the same row structure. The vision encoder, the projector, and the three pruning spans (Stage 1 selection, Stage 2 relevance, and Stage 2 selection) are CUDA event intervals recorded inside one generate call; the remaining span covers the decoder layers, the output head, and runtime/control work, so each column sums to the profiled generate span. The headline latency and peak-memory values are those reported in Tables 10 and 3; the component spans below are diagnostic. For NeXT-7B, the profile uses one A800 80GB GPU with $K = 7$, ten warm-ups, and 50 measured MME requests per budget. Within its remaining span, the decoder layers and final norm take 45.49/48.66/61.29 ms at $T = 160/320/640$, the output head 0.42/0.53/0.82 ms, and runtime/control 2.94/2.55/3.09 ms; the instrumented passes take 90.17/102.20/138.91 ms of wall time against un-instrumented reference passes of 87.95/98.75/134.92 ms before and 87.06/100.95/135.03 ms after profiling, so observer overhead is about 1–4%. For Video-7B, the spans are diagnostic means over two timed requests per budget (after warm-up) on the profiling cohort, a fixed set of videos sampled at 64 frames, and the decoder was not instrumented separately (†): its remaining span is the generate span minus the named spans. The two backbones differ in which stage dominates the pruning cost. On NeXT-7B, Stage 1 selects $2T$ tokens from the 2,880 multi-crop tokens while Stage 2 selects R from those survivors, so Stage 1 runs more greedy iterations and is the larger span (28.3 versus 13.9 ms at $T = 640$). On Video-7B, Stage 1 runs independently per frame over 169 tokens (batched across the 64 frames), whereas Stage 2 operates on the concatenated pool of $64 \times 2T$ survivors; its pairwise similarities and greedy loop over up to 8,192 tokens therefore dominate the pruning cost (247.6 ms at $T = 64$) and fall by roughly $3\times$ with each halving of T, while the per-frame Stage 1 stays at 5–11 ms.

Table 11: MiCo timing on LLaVA-NeXT-7B ($K = 7$). Component spans are means over 50 measured requests from paired instrumented profiles; complete-call latency is measured separately.

Component / measurement (ms unless noted)	$T = 160$	$T = 320$	$T = 640$
Vision encoder	28.87	29.59	30.48
Multimodal projector	0.62	0.62	0.62
Stage 1 selector	7.76	14.23	28.27
Stage 2 relevance (text attention)	0.39	0.09	0.10
Stage 2 selector	2.66	5.47	13.84
Pruning subtotal	*10.81*	*19.79*	*42.21*
Decoder, output head and runtime	48.84	51.73	65.21
Profiled generate span	89.14	101.73	138.52
Formal complete-call latency	88.1	99.4	135.5
Peak allocated memory (GiB)	13.4	13.5	13.8
Actual visual tokens per image (layer average)	159.84	319.69	640.16

Table 12: MiCo timing on LLaVA-Video-7B ($K = 6$). Component spans are diagnostic means over two timed requests after warm-up; complete-answer latency is measured separately.

Component / measurement (ms unless noted)	$T = 16$	$T = 32$	$T = 64$
Vision encoder	629.82	629.80	630.01
Multimodal projector	8.16	8.17	8.17
Stage 1 selector	4.62	6.65	10.94
Stage 2 relevance (text attention)	1.17	1.98	4.02
Stage 2 selector	22.94	74.28	247.64
Pruning subtotal	*28.73*	*82.91*	*262.60*
Decoder, output head and runtime[†]	113.27	199.04	394.25
Profiled generate span	779.98	919.92	1295.03
Formal complete-answer latency	804.7	933.8	1262.5
Peak allocated memory (GiB)	21.32	21.32	21.32
Actual visual tokens per frame	16.00	31.99	64.00

Component FLOPs. Tables 13 and 14 report the corresponding model and pruning arithmetic at the same budgets as the timing tables ($T = 160/320/640$ for NeXT-7B and per-frame $T = 16/32/64$ for Video-7B), separating model arithmetic from pruning arithmetic. Runtime/control work is excluded from FLOPs. The NeXT-7B counter further splits the pruning arithmetic into Stage 1 and Stage 2, matching the timing rows; the Video counter reports the Stage-1 and Stage-2 proxy computations as one row and the Triton-implemented greedy selection of both stages as a separate row. The Video breakdown is counted on the fixed 64-frame profiling cohort, so its totals differ by less than 1% from the native-input totals reported in Table 3.

Table 13: MiCo FLOPs breakdown on LLaVA-NeXT-7B ($K = 7$); percentages are shares of total FLOPs. Pruning arithmetic is attributed to Stage 1 and Stage 2 separately.

Component (GFLOPs)	$T = 160$	$T = 320$	$T = 640$
Vision encoder	1836.289 (36.63%)	1836.289 (25.57%)	1836.289 (15.78%)
Multimodal projector	120.879 (2.41%)	120.879 (1.68%)	120.879 (1.04%)
Decoder layers and final normalization	3002.246 (59.90%)	5136.922 (71.54%)	9520.805 (81.81%)
Vocabulary output head	48.350 (0.96%)	78.496 (1.09%)	139.052 (1.19%)
Stage 1: scores, similarity and selection	3.729 (0.07%)	4.047 (0.06%)	4.684 (0.04%)
Stage 2: extra attention / relevance	0.113 (0.002%)	0.206 (0.003%)	0.391 (0.003%)
Stage 2: scores, similarity and selection	0.881 (0.02%)	3.651 (0.05%)	15.713 (0.14%)
Stage 2 subtotal	0.994 (0.02%)	3.857 (0.05%)	16.104 (0.14%)
Model subtotal	5007.764 (99.91%)	7172.586 (99.89%)	11617.024 (99.82%)
Pruning subtotal	4.723 (0.09%)	7.904 (0.11%)	20.788 (0.18%)
Runtime/control (excluded from FLOPs)		–	
Total (GFLOPs)	**5012.486 (100.00%)**	**7180.490 (100.00%)**	**11637.813 (100.00%)**
Total (TFLOPs)	**5.012486**	**7.180490**	**11.637813**

Table 14: MiCo FLOPs breakdown on LLaVA-Video-7B ($K = 6$) on the fixed 64-frame profiling cohort; percentages are shares of total FLOPs. The two pruning paths are counted jointly; the Triton greedy term is listed separately.

Component (GFLOPs)	$T = 16$	$T = 32$	$T = 64$
Vision encoder	41228.062 (69.53%)	41228.062 (54.73%)	41228.062 (37.19%)
Multimodal projector	1585.032 (2.67%)	1585.032 (2.10%)	1585.032 (1.43%)
Visual pooling / plumbing	0.155 (0.000%)	0.155 (0.000%)	0.155 (0.000%)
Decoder layers and final normalization	15444.005 (26.05%)	30518.109 (40.51%)	63934.284 (57.68%)
Vocabulary output head	929.368 (1.57%)	1738.100 (2.31%)	3357.739 (3.03%)
Pruning operators (Stage 1 and Stage 2, jointly counted)	109.016 (0.18%)	261.046 (0.35%)	745.496 (0.67%)
Triton greedy selection	0.088 (0.000%)	0.386 (0.001%)	1.687 (0.002%)
Model subtotal	59186.622 (99.82%)	75069.459 (99.65%)	110105.272 (99.33%)
Pruning subtotal	109.104 (0.18%)	261.432 (0.35%)	747.183 (0.67%)
Runtime/control (excluded from FLOPs)		–	
Total (GFLOPs)	**59295.726 (100.00%)**	**75330.891 (100.00%)**	**110852.455 (100.00%)**
Total (TFLOPs)	**59.296**	**75.331**	**110.852**

E Additional Ablation Studies

E.1 Two-Stage Component Ablation

Table 15 reports AI2D, POPE, HallB, MME, MMB-EN, MMStar, and SQA-IMG for Figure 5, sorted by increasing Acc_7. Acc_7 averages exactly these seven benchmarks, using POPE accuracy

and MME divided by 20, and differs from the eight-benchmark Acc in the main performance tables. The main figure retains the English MMBench evaluation; Figure 9 shows MMB-CN, ChartQA, and OCRBench, none of which contributes to Acc_7. All means use unrounded evaluation records. Separate columns specify Stage-1 CLS-attn and Stage-2 text-to-visual attention. The run groups and their full-component references are distinguished in the table note. The random setting samples tokens uniformly without replacement; it differs from the legacy all-off setting, in which every token receives the same score and the selector degenerates to top-k in index order. All settings use $K = 2$ and a Stage-1 pool of $N_1 = 2T$. Y/N mark whether a component is enabled; Stage-1 CLS attention and Stage-2 text-to-visual attention have separate columns, whereas H and coverage are toggled jointly in both stages. The twelve plotted settings retain fixed colors and hatches, with bars sorted by score within each panel. Axis breaks mark truncated ranges; individual MME panels retain the original scale.

Within the paired-removal runs, disabling coverage lowers Acc_7 from 77.1 to 74.4 and disabling both attention signals lowers it to 75.7, the two largest drops among the tested switches; the feature-norm term H has the smallest effect in this collection. These paired-removal runs were collected with an earlier Stage-1 attention extraction, so their absolute scores sit slightly below the main-evaluation configuration ($\text{Acc}_7 = 78.4$); the single-component, random, and split-attention runs (the last disable only Stage-1 CLS attention or only Stage-2 text-to-visual attention) use the main-evaluation extraction. Comparisons are therefore most informative within a collection, and the table marks the collection of every row. Table 15 reports the individual switches, scores, and run distinctions. Stage-1 CLS-attn uses the native pooled analogue on Qwen2.5. MiCo does not lead every individual benchmark: no-cls-attn scores higher on MMB-EN.

Table 15: Two-stage component scores on Qwen2.5-VL-7B at $T = 128$, $K = 2$, and a Stage-1 pool of $N_1 = 2T$.

Setting	CLS-attn (S1)	Text-visual attn (S2)	H	Coverage	AI2D	POPE	HallB	MME	MMB EN	MMStar	SQA IMG	Acc_7
all-off	N	N	N	N	70.5	67.8	34.6	1662.24	72.4	46.7	77.6	64.7
only-H	N	N	Y	N	73.9	81.0	36.6	1933.90	79.6	51.3	82.5	71.7
only-cls-attn	Y	N	N	N	73.8	81.4	39.1	1975.37	77.3	52.6	81.6	72.1
random	N	N	N	N	73.7	80.8	40.0	2047.53	77.0	52.1	82.1	72.6
only-both-attn	Y	Y	N	N	75.0	81.9	41.2	1958.55	79.5	52.5	82.1	72.9
no-coverage	Y	Y	Y	N	77.5	82.6	40.9	2035.51	79.9	55.0	83.0	74.4
only-text-visual-attn	N	Y	N	N	76.7	82.3	44.3	2137.36	78.3	53.4	83.1	75.0
no-both-attn	N	N	Y	Y	78.1	85.0	43.0	2099.88	79.8	55.0	83.8	75.7
only-coverage	N	N	N	Y	77.5	84.4	43.2	2142.60	80.2	54.7	83.8	75.8
no-text-visual-attn	Y	N	Y	Y	78.2	85.2	42.6	2149.09	80.2	56.3	83.8	76.3
no-H	Y	Y	N	Y	79.6	84.4	45.6	2202.81	80.4	55.3	84.0	77.1
no-cls-attn	N	Y	Y	Y	79.9	84.6	46.7	2206.65	81.1	56.5	83.1	77.5
MiCo	Y	Y	Y	Y	80.4	85.5	48.6	2252.37	80.8	56.5	84.6	**78.4**

Note. CLS-attn denotes Qwen2.5's native pooled visual-attention analogue, not a literal CLS token. Text-visual-attn denotes text-to-visual attention in Stage 2; both-attn denotes both signals. MiCo uses the main-evaluation record. The no-H, no-both-attn, no-coverage, and all-off settings come from an earlier paired-removal run with a different Stage-1 attention extraction, whose full-component reference has $\text{Acc}_7 = 77.1$; the only-component, random, and split-attention runs use the main-evaluation extraction. Comparisons are most informative within one run. Scores and means use unrounded evaluation records; table values are rounded for display. In only-cls-attn and only-text-visual-attn, the stage without attention samples uniformly without replacement; random does so in both stages (one run, seed 0). All-off uses equal-score top-k selection and remains appendix-only.

E.2 Proxy Substitutions

Table 7 compares four proxy replacements on Qwen2.5-VL-7B at $T = 128$. The H replacement scores each token by its normalized hidden-state entropy instead of its feature norm, in both stages. The coverage replacement drops the pairwise coverage term and scores each token by its similarity to the mean visual feature (visual-mean similarity), in both stages. The reliability replacement replaces the Stage-1 CLS-attention proxy with each token's similarity to the global pooled visual token (global-token similarity), and the relevance replacement replaces the Stage-2 text-to-visual attention with each token's mean similarity to the text tokens (mean text–visual similarity). Acc averages AI2D, POPE accuracy, HallB, MME/20, MMBench-EN/CN, MMStar, and SQA-IMG; Rel normalizes Acc by the unpruned reference of 82.95. Rows are ordered by increasing Acc, with MiCo last. Each column highlights the best and second-best distinct displayed values; MiCo

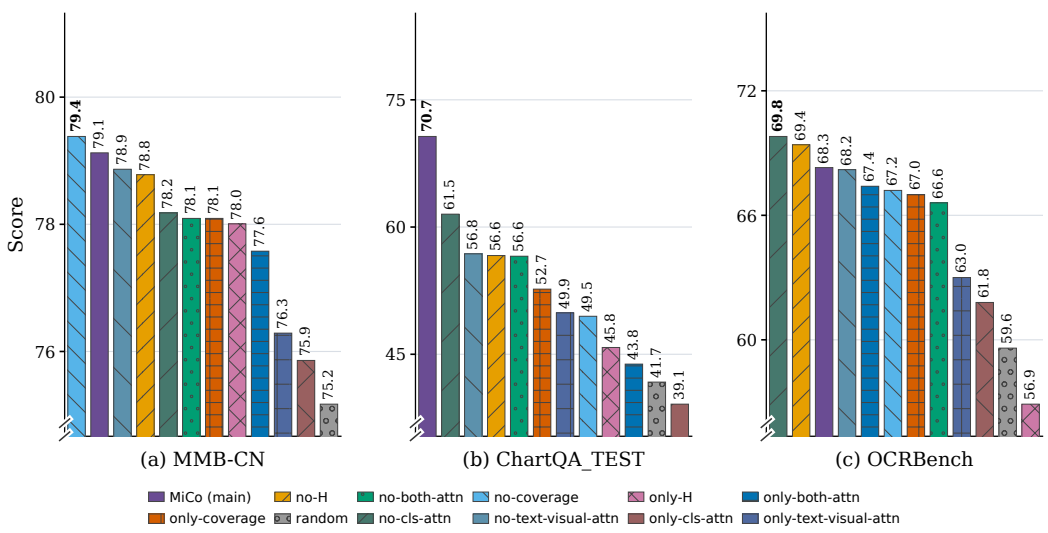


Figure 9: Additional component comparisons on Qwen2.5-VL-7B at $T = 128$; these tasks are excluded from Acc_7.

leads the aggregate, but not every individual benchmark. The complete task-relevance comparison below retains the additional uniform, inverse (one minus) mean text–visual similarity, and final-token alternatives.

E.3 Stage-2 Task-Relevance Ablation

We replace only the Stage-2 task-relevance weight on Qwen2.5 at $T = 128$ (Table 16). All variants use $K = 2$ and a Stage-1 pool of $2T$; uniform relevance sets the weight to one. Averaging text-to-visual attention over all text queries gives the highest Acc, 78.5, compared with 76.6 for uniform relevance, 77.1 for mean text–visual similarity, and 77.2 for its complement. Using only the final text token reaches 78.3. These results favor attention-based task relevance over raw cross-modal similarity, with a small further gain from aggregating text queries. All five variants exceed MMTok's Acc of 75.3, the strongest external baseline at this budget.

The aggregate uses the eight-benchmark Acc of the Qwen main tables, with MME divided by 20. POPE reports overall accuracy, MMBench uses Circular evaluation, and SQA uses image-question accuracy.

Table 16: Stage-2 task-relevance ablation on Qwen2.5-VL-7B at $T = 128$, $K = 2$, and a Stage-1 pool of $N_1 = 2T$.

Method / relevance	T	AI2D	POPE	HallB	MME	MMB EN	MMB CN	MMStar	SQA IMG	Acc	Rel
MMTok (ICLR'26)	128	76.3	84.2	43.3	2120.7	79.8	77.5	53.7	81.4	75.3	90.8%
uniform	128	78.2	85.2	42.6	2149.09	80.2	78.9	56.3	83.8	76.6	92.3%
text-vision-sim	128	79.1	85.5	45.9	2142.01	79.7	79	**56.5**	84.3	77.1	93.0%
1-(text-vision-sim)	128	79.4	**86.2**	45.9	2170.42	80.3	78.7	55.5	83.4	77.2	93.1%
last-token-attn	128	**80.5**	85.6	48.3	2235.97	80.6	79	56.1	84.3	78.3	94.4%
MiCo (all-text-attn)	128	80.4	85.5	**48.6**	**2252.37**	**80.8**	**79.1**	**56.5**	**84.6**	**78.5**	**94.7%**

E.4 Stage-1 Budget Ablations

To test the trade-off between a broad candidate pool and later-layer compression, we vary the Stage 1 pool $N_1 = nT$ for $n = 1, \ldots, 5$ on Qwen2.5-VL-7B while keeping the layer-average budget fixed (the figures show $n = 1$–4; the tables also include $n = 5$). Increasing the initial pool leaves fewer tokens for the remaining decoder layers under this constraint. Qwen2.5 uses 28 decoder blocks and

$K = 2$; for each n, the Stage-2 count R is recomputed from equation 26 with $N_1 = nT$, so the layer average stays at T (Appendix C.3).

Figure 10 shows four illustrative benchmarks and the full eight-benchmark mean at $T = 128$. All four pool sizes ($n = 1$–4) exceed the strongest external baseline on the displayed metrics. The eight-benchmark mean is higher at $n = 2$–4 than at $n = 1$ and peaks at $n = 2$, supporting a moderate initial candidate pool. This improvement is not uniform on each task: MMB-EN at $n = 3$ is 0.09 points below $n = 1$. The tables below report all benchmark results at both budgets, including $n = 5$, alongside the InternVL3 comparison.

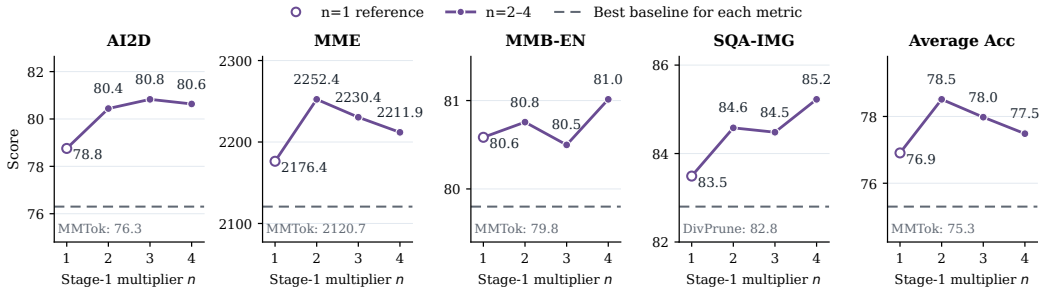


Figure 10: Stage-1 budget sensitivity on Qwen2.5-VL-7B at $T = 128$: four illustrative tasks and the eight-task mean.

Tables 17 and 18 report the full Stage-1 budget ablations for Qwen2.5-VL-7B and InternVL3-8B at $T = 256$ and $T = 128$, including $n = 5$. Figures 10 and 11 display $n = 1$–4 for Qwen2.5 and InternVL3, respectively. The $n = 1$ rows repeat the main-evaluation MiCo-S1 records of Tables 21 and 21, which both figures also use as their $n = 1$ reference. InternVL3 uses 28 decoder blocks and $K = 4$, and its displayed Average Acc peaks at $n = 3$. Its plotted $n = 2$ point is the main-evaluation MiCo record from Table 21 (Acc 79.3); Table 18 instead lists the separately collected budget-ablation run for that setting (Acc 79.2).

Each figure displays four illustrative benchmarks chosen after inspecting the results, with different selections for the two models. These panels do not summarize every task: the complete eight-benchmark mean is shown separately, and the tables retain all benchmark scores. On both models, all four pool sizes exceed the strongest external baseline in this mean, and $n = 2$–4 score higher than $n = 1$. Individual tasks do not always follow this pattern; for example, Qwen2.5 MMB-EN at $n = 3$ is 0.09 points below $n = 1$. Dashed lines show the strongest external pruner per metric; hollow $n = 1$ points identify the main-evaluation MiCo-S1 reference. Panel ranges differ, and MME retains its original scale outside the mean. Acc uses unrounded scores with MME/20. InternVL distributes the token budget evenly across its dynamic tiles, assigning any remainder to the first tiles.

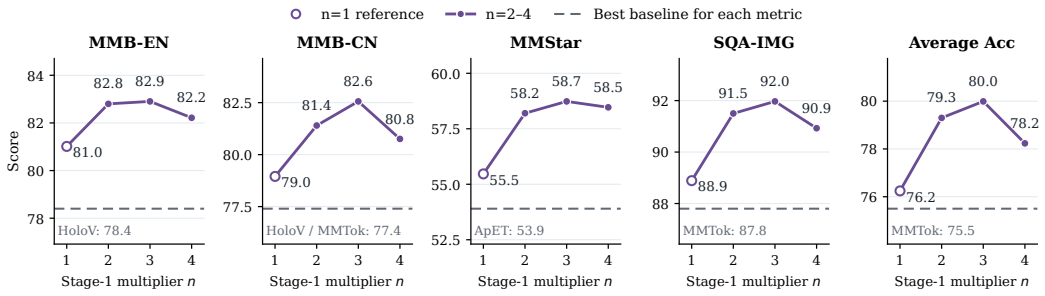


Figure 11: Stage-1 budget sensitivity on InternVL3-8B at $T = 128$: four illustrative tasks and the eight-task mean.

Table 17: Stage-1 budget ablation on Qwen2.5-VL-7B with refinement at $K = 2$.

S1 setting	T	S1 tokens	S2 R	AI2D	POPE	HallB	MME	MMB EN	MMB CN	MMStar	SQA IMG	Acc
n=1 (T)	256	256	256	81.7	86.1	48.6	2282.2	82.0	80.5	59.5	85.5	79.8
n=2 (2T)	256	512	236	83.8	86.3	51.2	2291.0	83.2	81.3	60.1	86.2	80.8
n=3 (3T)	256	768	217	82.4	86.4	50.6	2289.3	82.3	81.4	60.7	86.4	80.6
n=4 (4T)	256	1024	197	82.4	86.1	50.0	2266.5	82.4	81.3	59.6	86.3	80.2
n=5 (5T)	256	1280	177	82.0	85.6	49.9	2250.1	81.9	80.6	59.1	85.8	79.7
n=1 (T)	128	128	128	78.8	84.3	46.3	2176.4	80.6	78.4	54.6	83.5	76.9
n=2 (2T)	128	256	118	80.4	85.5	48.6	2252.4	80.8	79.1	56.5	84.6	78.5
n=3 (3T)	128	384	108	80.8	85.0	46.5	2230.4	80.5	78.7	56.3	84.5	78.0
n=4 (4T)	128	512	98	80.6	84.2	45.3	2211.9	81.0	78.4	54.5	85.2	77.5
n=5 (5T)	128	640	89	79.9	84.0	45.4	2203.2	80.2	77.8	54.9	85.1	77.2

Table 18: Stage-1 budget ablation on InternVL3-8B with refinement at $K = 4$.

S1 setting	T	S1 tokens	S2 R	AI2D	POPE	HallB	MME	MMB EN	MMB CN	MMStar	SQA IMG	Acc
n=1 (T)	256	256	256	80.1	90.8	44.1	2279.1	83.7	83.2	59.7	93.4	81.1
n=2 (2T)	256	512	213	83.0	90.2	45.1	2293.7	85.0	84.5	63.0	95.6	82.6
n=3 (3T)	256	768	171	82.9	90.0	45.3	2302.5	85.3	84.7	63.5	95.6	82.8
n=4 (4T)	256	1024	128	82.1	89.9	43.2	2281.9	84.8	83.7	61.7	94.6	81.8
n=5 (5T)	256	1280	85	79.7	89.0	40.7	2260.5	82.6	83.0	59.5	92.6	80.0
n=1 (T)	128	128	128	73.9	89.3	36.8	2112.5	81.0	79.0	55.5	88.9	76.2
n=2 (2T)	128	256	107	77.5	89.7	40.4	2252.3	82.6	81.4	57.8	91.7	79.2
n=3 (3T)	128	384	85	79.4	89.6	40.6	2284.1	82.9	82.6	58.7	92.0	80.0
n=4 (4T)	128	512	64	78.2	88.6	38.7	2159.8	82.2	80.8	58.5	90.9	78.2
n=5 (5T)	128	640	43	75.6	87.8	37.5	2149.2	80.0	78.4	54.5	88.3	76.2

F Refinement-Layer Selection and Sensitivity

F.1 Layer Configuration Summary

We use one fixed, architecture-specific refinement layer in each reported configuration. Table 19 summarizes the selected layers, and Appendix F.2 gives the selection procedure and representative examples.

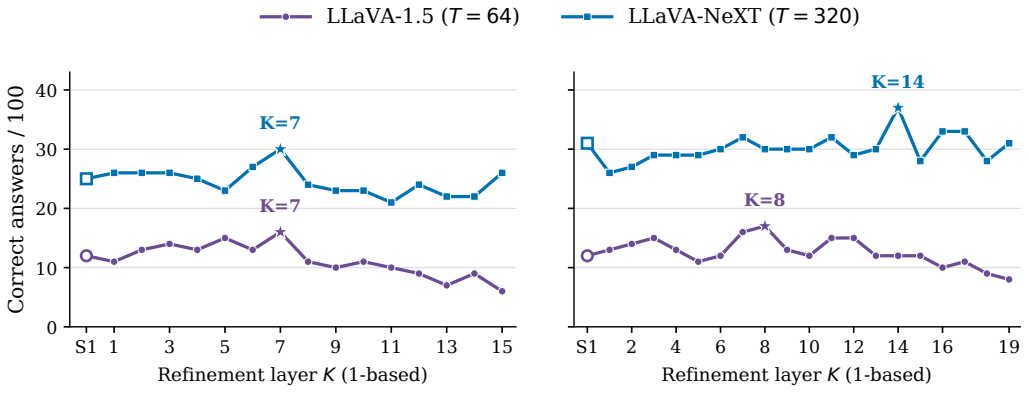
Table 19: Refinement-layer configurations used in the reported experiments.

Model	**Refinement-layer configuration**
LLaVA-1.5-7B	$K = 7$
LLaVA-1.5-13B	$K = 8$
LLaVA-NeXT-7B	$K = 7$
LLaVA-NeXT-13B	$K = 14$
Qwen3-VL-8B	$K = 6$
InternVL3-8B	$K = 4$
Qwen2.5-VL-7B	$K = 2$
Qwen3.5-9B	$K = 12$
LLaVA-Video-7B	$K = 6$

Detailed selection examples are given in Appendix F.2, while the complete LLaVA sweep tables are collected with the full image-model results.

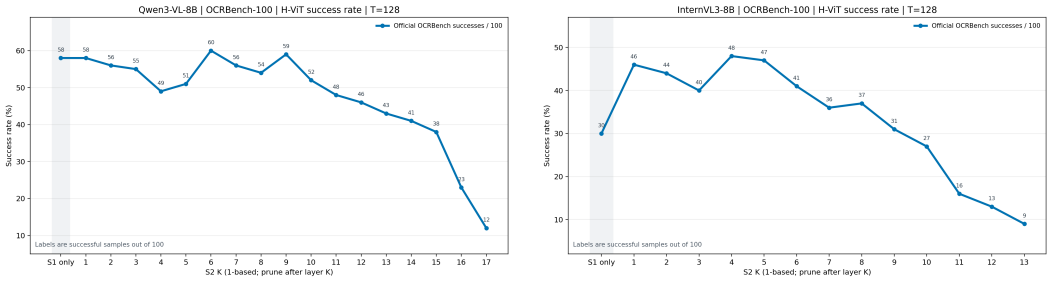
F.2 IMAGE-MODEL CALIBRATION

For the representative OCR-based checks on Qwen3-VL and InternVL3, we hold the same 100 OCRBench samples and target budget $T = 128$ fixed, sweep the one-based decoder-layer cutoff K over candidate blocks, run the same two-stage pruning pipeline for each candidate, and select the layer with the highest OCR accuracy. The LLaVA calibration plots use the corresponding fixed sample and budget settings stated in the captions. Answer-token KL divergence is reported as a supplementary diagnostic. Figure 12 shows the size-grouped LLaVA calibration, while Figures 13a and 13b give representative Qwen3-VL-8B and InternVL3-8B examples at $T = 128$. Accuracy curves count correct answers; stars mark the selected maxima, and hollow entries mark the MiCo-S1 (Stage-1-only) references. Layer indices are one-based, with pruning after K.



(a) 7B models (b) 13B models

Figure 12: Refinement-layer selection on 100 OCRBench samples: LLaVA-1.5 at $T = 64$ and LLaVA-NeXT at $T = 320$.



(a) Qwen3-VL-8B **(b)** InternVL3-8B

Figure 13: Representative refinement-layer selection for Qwen3-VL-8B and InternVL3-8B on 100 OCRBench samples at $T = 128$.

F.3 LLAVA REFINEMENT-LAYER SWEEPS

Figures 14, 15, and 16 report the refinement-layer sweeps at $T = 32/160$, $64/320$, and $128/640$, respectively, for LLaVA-1.5/LLaVA-NeXT. Each figure groups models by parameter scale, with one curve for LLaVA-1.5 and one for LLaVA-NeXT. The selected layers remain $K = 7/8$ for LLaVA-1.5-7B/13B and $K = 7/14$ for LLaVA-NeXT-7B/13B; they are not reselected from the downstream scores. The left/right panels show 7B/13B models and sweep $K = 1$–$15/1$–19, respectively. Diamonds mark calibration-selected layers; stars mark sweep maxima, including ties. Curves use the ten-benchmark mean Acc_{10} defined in Appendix G, and refinement occurs after the one-based layer K.

At the most aggressive budgets ($T = 32$ for LLaVA-1.5 and $T = 160$ for LLaVA-NeXT), the calibration-selected layers on LLaVA-NeXT-7B/13B are within 0.23/0.05 Acc points of the observed sweep maxima; the corresponding gaps on LLaVA-1.5-7B/13B are 1.19/0.45 points. The

calibration-selected layer therefore need not maximize the downstream ten-benchmark mean, although it is close to the observed peak on both NeXT models.

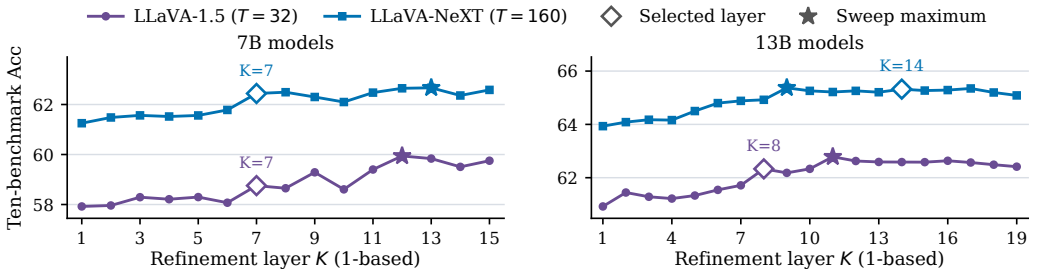


Figure 14: Refinement-layer sweeps at 94.4% pruning: LLaVA-1.5 ($T = 32$) and LLaVA-NeXT ($T = 160$).

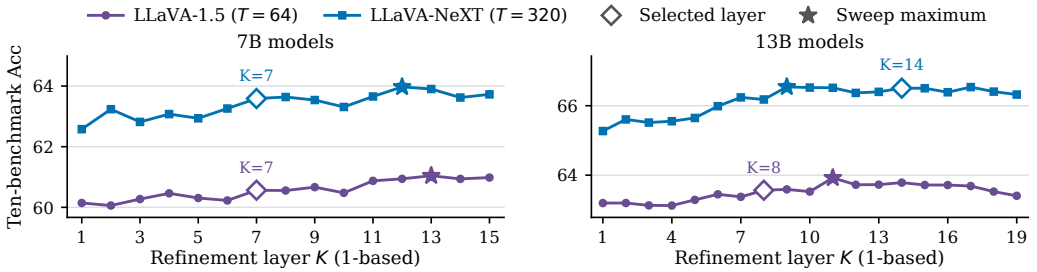


Figure 15: Refinement-layer sweeps at 88.9% pruning: LLaVA-1.5 ($T = 64$) and LLaVA-NeXT ($T = 320$).

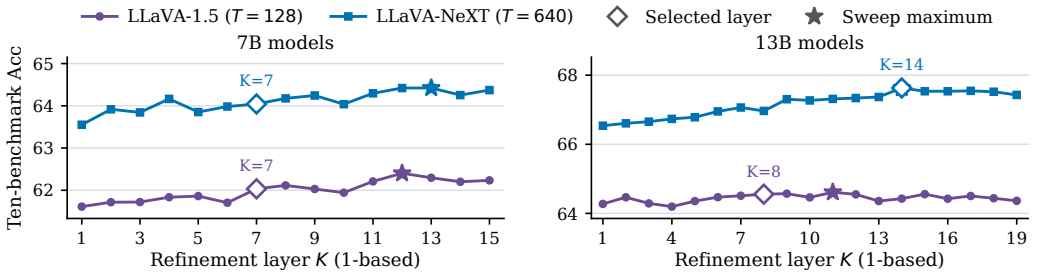


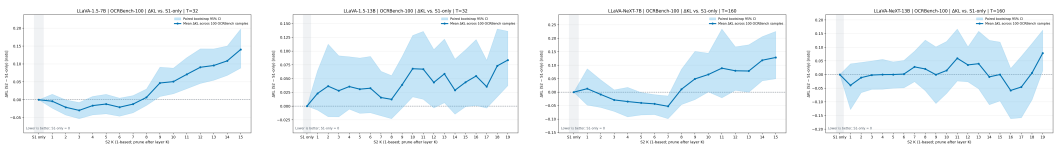
Figure 16: Refinement-layer sweeps at 77.8% pruning: LLaVA-1.5 ($T = 128$) and LLaVA-NeXT ($T = 640$).

F.4 Answer-Token KL Diagnostics

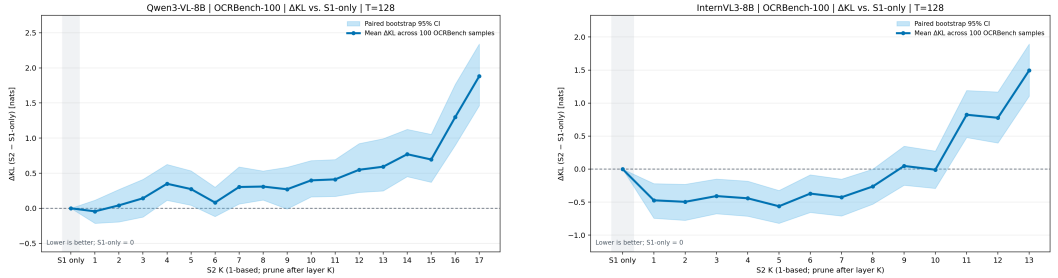
Figure 17 reports answer-token KL divergence as a complementary diagnostic, showing the change in KL from the unpruned baseline relative to the Stage-1-only reference. For LLaVA, the refinement layers were selected from the OCRBench accuracy curves at $T = 64$ (LLaVA-1.5) and $T = 320$ (LLaVA-NeXT) in Figure 12; the KL diagnostic was run at the most aggressive budgets, $T = 32$ and $T = 160$, to check the selected layers there. The representative Qwen3 and InternVL3 panels use $T = 128$ in both the accuracy and KL figures. Each ΔKL subtracts Stage-1-only KL from the refined model's answer-token KL to the unpruned baseline; lower is better. Shading denotes paired-bootstrap 95% confidence intervals.

F.5 Video Refinement Layer

For LLaVA-Video, we select the refinement layer once for the video setting and fix it at $K = 6$ for all reported video budgets; it is not re-tuned per budget or per benchmark. The video results therefore use the same fixed configuration throughout the reported efficiency and performance comparisons.



(a) LLaVA-1.5-7B **(b)** LLaVA-1.5-13B **(c)** LLaVA-NeXT-7B **(d)** LLaVA-NeXT-13B



(e) Qwen3-VL-8B **(f)** InternVL3-8B

Figure 17: Layer-wise ΔKL on 100 OCRBench samples (lower is better); shading shows paired-bootstrap 95% intervals.

G Complete Image-Model Results

The main text reports detailed results for LLaVA-NeXT-13B and summarizes the remaining image models by family. For lookup, this appendix keeps the complete benchmark scores in two cross-page tables: one for the LLaVA series and one for Qwen-VL and InternVL. Each model is a labeled block within its family table, with the original budgets, methods, MiCo-S1 rows, and highlighting preserved.

G.1 LLaVA Series

The first longtable consolidates LLaVA-1.5-7B/13B and LLaVA-NeXT-7B/13B. All four models share the same ten benchmark columns, followed by the aggregate Acc and relative retention (Rel). FastV and SparseVLM keep every visual token through their first two decoder layers, so at the 94.4% budget ($T = 32$ on LLaVA-1.5, $T = 160$ on LLaVA-NeXT) the layer-average accounting of Appendix C.3 leaves them no feasible schedule on the 32-layer 7B models and only a degenerate one on the 13B models; they are therefore not reported at that budget, and Tables 10 and 28 leave the same entries blank.

Table 20: Complete benchmark results across the LLaVA series, including MiCo-S1.

Method	GQA	SQA-IMG	TextVQA	POPE	MME	MMB-EN	MMB-CN	SEED	AI2D	MMMU	Acc	Rel
LLaVA-1.5-7B												
Upper Bound — 576 tokens (no pruning)												
Vanilla	61.9	69.5	58.2	85.9	1508.8	64.7	58.1	66.0	55.5	35.0	63.0	100.0%
Retain 128 tokens (↓77.8%)												
FastV (ECCV'24)	54.0	69.2	56.4	68.2	1376.5	**63.0**	55.9	59.7	53.6	34.3	58.3	92.6%
SparseVLM (ICML'25)	57.3	69.0	56.3	83.1	1401.3	62.6	56.9	61.7	54.7	35.9	60.8	96.4%
DivPrune (CVPR'25)	59.4	68.6	55.9	**87.0**	1411.2	61.5	54.8	62.4	54.2	35.2	61.0	96.8%
VisionZip (CVPR'25)	57.6	68.7	56.9	83.3	1437.3	62.1	57.0	61.6	54.5	35.9	60.9	96.7%
VisPruner (ICCV'25)	58.3	68.8	56.9	84.4	1425.5	61.6	55.9	61.8	53.7	35.8	60.8	96.6%
VScan (TMLR'26)	59.0	68.6	57.2	85.0	1425.6	61.7	**57.1**	**64.19**	**55.5**	**36.7**	61.63	97.8%
HoloV (NeurIPS'25)	57.4	68.1	55.7	82.3	1437.4	61.9	56.7	61.3	54.6	35.1	60.5	96.0%
MMTok (ICLR'26)	59.2	68.9	56.8	86.5	1419.9	60.9	55.5	63.3	54.1	35.0	61.1	97.0%
ApET (CVPR'26)	59.1	68.1	54.3	86.3	1430.6	61.4	54.9	63.0	54.1	34.6	60.7	96.4%
MiCo-S1	59.2	68.6	57.2	85.8	1433.3	61.6	56.8	62.7	54.3	35.2	61.3	97.3%
MiCo	**60.0**	**69.3**	**57.6**	85.9	**1474.8**	62.4	56.5	63.7	54.9	36.3	**62.0**	**98.5%**
Retain 64 tokens (↓88.9%)												

Continued on next page

Table 20 (continued): Complete benchmark results across the LLaVA series, including MiCo-S1

Method	GQA	SQA-IMG	TextVQA	POPE	MME	MMB-EN	MMB-CN	SEED	AI2D	MMMU	Acc	Rel
FastV (ECCV'24)	46.0	**70.1**	51.6	35.4	962.0	50.1	42.1	46.8	50.6	34.7	47.5	75.5%
SparseVLM (ICML'25)	52.0	69.2	52.1	69.7	1192.8	58.3	49.6	53.9	52.5	**36.3**	55.3	87.8%
DivPrune (CVPR'25)	57.7	68.0	54.5	85.4	1365.8	60.1	52.3	60.1	53.8	34.8	59.5	94.4%
VisionZip (CVPR'25)	55.1	69.0	55.5	77.0	1371.5	60.1	55.4	57.7	53.3	35.4	58.7	93.2%
VisPruner (ICCV'25)	56.0	68.5	55.8	80.6	1338.0	58.8	54.3	58.4	53.5	35.0	58.8	93.3%
VScan (TMLR'26)	57.9	69.1	56.1	85.0	1375.0	**61.3**	55.7	61.34	**54.4**	35.6	60.52	96.0%
HoloV (NeurIPS'25)	55.0	68.6	54.9	76.8	1362.5	59.2	55.6	58.2	51.9	34.8	58.3	92.6%
MMTok (ICLR'26)	**58.3**	68.7	55.9	**85.6**	1403.6	59.3	53.8	**61.5**	53.1	35.0	60.1	95.5%
ApET (CVPR'26)	57.4	68.6	53.2	84.6	1376.1	59.4	53.5	60.5	53.4	34.8	59.4	94.3%
MiCo-S1	57.7	68.5	**56.4**	83.3	1401.2	60.0	**56.0**	60.9	53.9	35.8	60.2	95.6%
MiCo	58.1	69.7	**56.4**	84.7	**1426.8**	60.4	55.7	61.0	54.1	34.2	**60.6**	**96.1%**
Retain 32 tokens (↓94.4%)												
DivPrune (CVPR'25)	54.9	68.6	52.9	81.5	1333.7	57.6	49.1	56.9	53.0	34.0	57.5	91.3%
VisionZip (CVPR'25)	51.8	69.1	53.1	69.4	1263.6	57.0	50.3	53.2	51.7	35.0	55.4	87.9%
VisPruner (ICCV'25)	52.2	67.9	53.2	73.4	1209.6	57.1	49.8	53.1	51.6	34.2	55.3	87.8%
VScan (TMLR'26)	54.9	**69.3**	53.8	79.8	1298.1	58.7	51.8	57.3	52.6	**35.4**	57.85	91.8%
HoloV (NeurIPS'25)	52.8	69.0	53.7	70.3	1261.7	58.2	51.5	54.8	53.1	33.1	56.0	88.8%
MMTok (ICLR'26)	**56.2**	69.0	53.5	**82.9**	1336.9	57.8	49.0	**59.5**	53.0	34.4	58.2	92.4%
ApET (CVPR'26)	54.7	68.5	50.9	80.6	1331.0	57.0	50.4	57.2	52.8	34.0	57.3	90.9%
MiCo-S1	55.6	69.0	54.6	80.5	1354.7	**59.6**	53.2	58.1	52.6	32.9	58.4	92.7%
MiCo	55.9	68.7	**55.2**	80.9	**1355.0**	59.2	**54.0**	58.2	**53.6**	34.1	**58.8**	**93.3%**
LLaVA-1.5-13B												
Upper Bound — 576 tokens (no pruning)												
Vanilla	63.3	72.8	61.2	86.0	1533.2	68.5	63.5	68.2	60.8	36.4	65.7	100.0%
Retain 128 tokens (↓77.8%)												
FastV (ECCV'24)	58.3	74.2	58.6	75.5	1460.6	66.1	62.3	55.5	58.1	36.9	61.9	94.1%
SparseVLM (ICML'25)	59.6	**74.3**	59.3	85.0	1487.9	**68.4**	62.6	65.3	58.1	37.4	64.4	98.1%
DivPrune (CVPR'25)	59.2	72.8	58.0	**86.8**	1457.7	66.3	60.7	64.1	57.7	36.8	63.5	96.7%
VisionZip (CVPR'25)	57.8	73.8	58.9	82.5	1448.2	66.8	62.3	63.7	57.0	**37.9**	63.3	96.4%
VisPruner (ICCV'25)	58.3	74.0	59.1	83.9	1427.0	67.1	62.4	63.9	58.1	36.7	63.5	96.6%
VScan (TMLR'26)	59.3	73.4	58.6	85.1	1469.2	65.9	62.4	65.8	**58.6**	37.1	63.97	97.3%
HoloV (NeurIPS'25)	58.1	73.4	57.9	82.0	1445.1	66.0	61.3	63.9	56.8	36.6	62.8	95.6%
MMTok (ICLR'26)	59.0	73.7	59.0	86.3	1454.8	66.1	61.9	64.5	57.6	35.3	63.6	96.8%
ApET (CVPR'26)	59.0	73.2	56.7	85.9	1467.9	65.2	60.7	65.3	57.4	36.4	63.3	96.4%
MiCo-S1	59.5	74.1	59.0	85.8	1458.1	66.7	**62.8**	64.7	57.0	36.4	63.9	97.2%
MiCo	**60.3**	73.7	**59.6**	86.6	**1509.3**	67.0	62.4	**66.0**	58.1	36.4	**64.6**	**98.3%**
Retain 64 tokens (↓88.9%)												
FastV (ECCV'24)	51.9	73.1	53.4	57.0	1233.6	59.2	55.1	54.6	55.5	36.8	55.8	85.0%
SparseVLM (ICML'25)	55.9	73.0	57.1	77.9	1374.3	65.2	60.3	60.8	56.7	36.4	61.2	93.2%
DivPrune (CVPR'25)	57.9	71.7	57.4	84.5	1430.7	59.2	55.1	62.2	57.0	35.7	61.2	93.2%
VisionZip (CVPR'25)	56.1	74.2	57.5	76.0	1397.4	64.3	61.1	60.2	56.8	35.9	61.2	93.1%
VisPruner (ICCV'25)	56.4	74.2	58.0	79.8	1396.1	63.5	59.6	60.7	57.1	36.2	61.5	93.6%
VScan (TMLR'26)	58.6	73.5	**58.7**	84.3	1446.8	64.7	61.5	63.5	57.2	**36.9**	63.12	96.0%
HoloV (NeurIPS'25)	56.1	**74.3**	57.3	75.7	1403.0	63.8	60.2	60.5	56.7	35.3	61.0	92.9%
MMTok (ICLR'26)	58.5	72.5	57.8	84.5	1449.7	64.1	59.4	63.5	56.9	35.7	62.5	95.2%
ApET (CVPR'26)	58.1	72.1	55.6	83.9	1425.9	62.9	59.2	62.7	56.2	36.1	61.8	94.1%
MiCo-S1	58.2	73.1	58.1	82.6	1437.3	65.6	**61.8**	63.3	**57.5**	35.6	62.8	95.5%
MiCo	**58.7**	73.2	58.4	**86.4**	**1490.8**	**65.9**	61.1	**63.9**	57.1	36.4	**63.6**	**96.7%**
Retain 32 tokens (↓94.4%)												
DivPrune (CVPR'25)	56.2	70.9	54.6	79.2	1356.9	61.7	57.2	59.5	57.2	35.0	59.9	91.2%
VisionZip (CVPR'25)	52.7	72.7	55.1	67.0	1254.7	60.5	55.7	55.8	56.4	35.6	57.4	87.4%
VisPruner (ICCV'25)	53.4	71.5	56.0	72.1	1261.6	61.6	56.4	56.0	56.2	35.3	58.2	88.5%
VScan (TMLR'26)	53.9	73.2	55.1	66.9	1298.2	61.9	56.2	59.3	57.3	34.8	58.27	88.7%
HoloV (NeurIPS'25)	53.4	71.5	55.6	69.1	1281.8	62.7	58.2	57.2	56.4	**36.2**	58.4	88.9%
MMTok (ICLR'26)	**57.6**	72.8	55.9	82.2	1450.2	62.5	57.8	**61.7**	56.5	31.7	61.1	93.0%
ApET (CVPR'26)	55.6	72.4	53.5	79.1	1388.4	61.6	56.4	59.4	56.4	34.7	59.8	91.1%
MiCo-S1	56.9	**73.5**	**57.3**	78.8	1375.4	64.1	60.5	60.8	57.0	35.6	61.3	93.3%
MiCo	57.3	73.0	57.0	**82.3**	**1466.2**	**64.6**	**61.0**	61.5	**57.4**	35.9	**62.3**	**94.9%**
LLaVA-NeXT-7B												
Upper Bound — 2880 tokens (no pruning)												
Vanilla	62.5	67.5	60.3	86.8	1511.8	65.8	57.3	69.7	64.7	35.2	64.5	100.0%
Retain 640 tokens (↓77.8%)												
FastV (ECCV'24)	58.9	67.4	58.1	79.5	1412.6	63.1	53.5	64.3	65.1	35.0	61.5	95.4%
SparseVLM (ICML'25)	61.2	67.6	59.7	85.3	1456.8	65.9	58.6	67.2	64.7	35.1	63.8	98.9%
DivPrune (CVPR'25)	61.9	67.8	57.0	86.9	1469.7	65.8	57.3	67.6	65.6	35.4	63.9	99.0%

Continued on next page

Table 20 (continued): Complete benchmark results across the LLaVA series, including MiCo-S1

Method	GQA	SQA-IMG	TextVQA	POPE	MME	MMB-EN	MMB-CN	SEED	AI2D	MMMU	Acc	Rel
VisionZip (CVPR'25)	61.2	68.1	60.0	86.1	1449.0	65.4	58.1	66.8	65.3	35.9	63.9	99.0%
VisPruner (ICCV'25)	61.5	**68.5**	59.8	85.7	1468.7	64.8	57.4	66.3	64.9	35.0	63.7	98.7%
VScan (TMLR'26)	62.4	67.9	58.3	86.9	1473.8	**66.6**	**59.8**	66.7	**66.7**	34.9	**64.22**	**99.5%**
HoloV (NeurIPS'25)	60.5	66.5	54.8	85.6	1476.6	63.9	57.3	66.8	62.8	34.7	62.7	97.1%
MMTok (ICLR'26)	62.2	68.2	59.0	86.9	1466.1	65.4	57.6	67.6	65.5	35.8	64.2	99.4%
ApET (CVPR'26)	**63.4**	68.1	54.2	86.7	**1476.7**	65.5	58.8	68.2	65.9	**36.9**	64.1	99.4%
MiCo-S1	62.0	68.3	**60.2**	86.4	1459.9	65.7	58.3	67.5	64.8	34.3	64.1	99.3%
MiCo	61.9	68.2	58.9	**87.3**	1446.8	65.5	57.7	**68.4**	65.3	35.0	64.1	99.2%
Retain 320 tokens (↓88.9%; LLaVA-1.5: 64 tokens)												
FastV (ECCV'24)	49.8	66.6	52.2	49.5	1099.0	53.4	42.5	52.8	62.2	34.9	51.9	80.4%
SparseVLM (ICML'25)	57.9	67.2	56.5	76.9	1386.1	63.1	56.7	62.1	63.0	34.0	60.7	94.0%
DivPrune (CVPR'25)	61.1	67.7	56.2	84.7	1423.3	63.9	55.7	65.6	**65.8**	35.4	62.7	97.2%
VisionZip (CVPR'25)	58.9	67.5	**58.8**	82.2	1416.6	62.8	55.6	63.4	63.2	35.8	61.9	95.9%
VisPruner (ICCV'25)	58.9	**68.7**	**58.8**	81.5	1406.2	62.9	55.7	63.0	63.5	33.9	61.7	95.6%
VScan (TMLR'26)	59.9	68.4	57.7	83.0	1411.1	63.5	56.4	63.3	63.5	35.4	62.11	96.2%
HoloV (NeurIPS'25)	59.2	67.1	55.5	82.8	**1465.8**	63.8	55.7	64.6	62.2	36.1	62.0	96.1%
MMTok (ICLR'26)	61.16	67.18	56.75	85.86	1418.9	64.0	55.8	66.3	64.2	36.2	62.8	97.4%
ApET (CVPR'26)	61.1	67.1	52.0	84.7	1430.2	62.0	55.8	64.3	63.9	**37.0**	61.9	96.0%
MiCo-S1	60.7	68.4	58.2	84.3	1462.6	64.6	57.0	65.5	63.3	34.9	63.0	97.6%
MiCo	**61.3**	68.5	58.3	**86.4**	1437.0	**65.5**	**57.7**	**66.4**	64.8	35.1	**63.6**	**98.5%**
Retain 160 tokens (↓94.4%; LLaVA-1.5: 32 tokens)												
DivPrune (CVPR'25)	59.3	67.1	54.1	80.0	1356.6	62.9	53.7	62.9	**65.5**	34.4	60.8	94.2%
VisionZip (CVPR'25)	55.2	67.9	56.0	74.9	1324.8	58.2	50.4	58.1	62.8	34.0	58.4	90.5%
VisPruner (ICCV'25)	56.5	**68.5**	55.8	75.0	1332.0	58.8	50.9	58.1	62.3	34.6	58.7	91.0%
VScan (TMLR'26)	56.6	68.4	53.4	78.1	1255.5	59.2	49.1	58.3	62.4	35.0	58.29	90.3%
HoloV (NeurIPS'25)	57.0	67.2	55.4	78.1	1349.4	62.1	54.2	60.7	60.7	34.4	59.7	92.5%
MMTok (ICLR'26)	**59.97**	68.12	54.47	83.79	1342.9	63.3	53.4	**65.0**	62.9	35.6	61.4	95.1%
ApET (CVPR'26)	58.3	67.0	50.0	81.5	1356.3	58.3	51.6	60.4	62.9	**36.0**	59.4	92.0%
MiCo-S1	59.4	67.7	56.5	80.8	1398.5	62.8	55.8	63.7	62.8	34.8	61.4	95.1%
MiCo	59.5	68.2	**56.8**	**83.8**	**1446.1**	**64.1**	**56.6**	64.4	63.5	35.2	**62.4**	**96.7%**
LLaVA-NeXT-13B												
Upper Bound — 2880 tokens (no pruning)												
Vanilla	64.4	73.1	63.2	85.3	1539.5	68.5	61.2	71.6	70.1	35.7	67.0	100.0%
Retain 640 tokens (↓77.8%)												
FastV (ECCV'24)	60.9	71.7	60.7	80.2	1516.7	65.5	59.9	67.4	67.7	35.8	64.6	96.4%
SparseVLM (ICML'25)	62.7	72.5	**62.8**	85.6	1573.6	68.5	**64.0**	69.7	68.1	37.6	67.0	100.0%
DivPrune (CVPR'25)	63.5	72.2	59.2	86.5	1526.1	67.5	62.9	69.4	68.4	37.8	66.4	99.1%
VisionZip (CVPR'25)	62.9	70.8	62.1	85.8	1531.1	67.5	62.6	68.7	68.3	36.9	66.2	98.8%
VisPruner (ICCV'25)	62.6	71.4	62.1	85.2	1561.2	67.4	63.1	68.8	68.2	37.0	66.4	99.1%
VScan (TMLR'26)	62.8	72.2	61.8	85.2	1553.6	67.6	62.7	68.2	69.9	36.6	66.5	99.2%
HoloV (NeurIPS'25)	61.7	70.3	57.3	85.5	1497.7	67.5	61.9	68.7	66.6	37.0	65.1	97.2%
MMTok (ICLR'26)	63.7	71.4	60.8	86.8	1540.9	66.6	62.5	69.7	68.1	**38.3**	66.5	99.2%
ApET (CVPR'26)	64.2	72.5	58.0	86.4	1475.4	67.7	62.5	69.5	**70.9**	36.7	66.2	98.8%
MiCo-S1	63.4	71.4	61.9	86.1	1533.8	67.4	63.2	69.6	68.6	37.2	66.6	99.3%
MiCo	**64.3**	**73.2**	61.9	**87.0**	**1573.9**	**69.0**	63.4	**71.4**	70.2	37.2	**67.6**	**100.9%**
Retain 320 tokens (↓88.9%; LLaVA-1.5: 64 tokens)												
FastV (ECCV'24)	54.6	70.5	55.4	63.6	1279.0	59.8	54.4	59.2	65.0	35.8	58.2	86.9%
SparseVLM (ICML'25)	60.9	70.9	60.0	81.5	1498.6	67.4	**63.5**	65.8	66.8	**37.4**	64.9	96.9%
DivPrune (CVPR'25)	61.8	72.3	57.6	85.2	1473.0	65.9	61.9	67.2	67.7	37.2	65.0	97.1%
VisionZip (CVPR'25)	60.7	70.2	60.7	82.3	1497.4	65.9	62.3	65.3	67.1	37.2	64.6	96.5%
VisPruner (ICCV'25)	60.8	70.1	60.3	81.1	1486.0	65.7	62.5	65.1	67.1	36.4	64.4	96.1%
VScan (TMLR'26)	61.0	**72.4**	59.3	82.0	1496.4	65.2	59.8	64.9	66.9	36.3	64.3	95.9%
HoloV (NeurIPS'25)	60.7	69.9	58.2	83.2	1496.6	66.3	62.2	66.5	66.5	35.8	64.4	96.1%
MMTok (ICLR'26)	62.78	71.49	59.05	86.06	1501.1	65.0	61.7	67.7	67.9	37.2	65.4	97.6%
ApET (CVPR'26)	61.7	71.4	54.7	84.1	1455.0	65.6	60.1	65.5	**68.4**	36.4	64.1	95.6%
MiCo-S1	62.5	71.6	60.1	84.4	1472.9	65.9	63.2	67.7	68.1	36.3	65.3	97.5%
MiCo	**63.3**	71.3	**60.9**	**86.5**	**1543.3**	**67.9**	63.2	**69.5**	68.3	37.0	**66.5**	**99.3%**
Retain 160 tokens (↓94.4%; LLaVA-1.5: 32 tokens)												
DivPrune (CVPR'25)	60	71.4	56.3	81.9	1436.7	65.1	60.9	64.5	67.3	36.6	63.6	94.9%
VisionZip (CVPR'25)	57.78	69.66	58.55	76.81	1402	64.1	60.0	61.2	65.7	37.0	62.1	92.7%
VisPruner (ICCV'25)	58.4	71.2	58.4	76.2	1390.8	64.1	59.6	61.1	65.5	36.2	62.0	92.6%
VScan (TMLR'26)	58.2	70.3	55.0	75.5	1385.1	61.5	53.4	61.5	63.4	34.1	60.2	89.9%
HoloV (NeurIPS'25)	58.5	70	57.7	78.5	1424.5	65.7	61.6	62.5	65.4	36.9	62.8	93.7%
MMTok (ICLR'26)	**62.02**	**72.19**	56.47	**85.52**	1465.7	65.3	60.3	65.5	67.3	37.0	64.5	96.3%
ApET (CVPR'26)	59.1	71.4	52.9	79.6	1342.7	62.1	55.8	62.3	65.6	36.8	61.3	91.4%
MiCo-S1	60.8	71.3	**58.9**	81.9	1460.9	**66.3**	62.4	65.7	67.2	**37.2**	64.5	96.2%
MiCo	62.0	71.8	58.4	85.0	**1488.4**	66.2	**63.1**	**67.6**	**68.0**	36.7	**65.3**	**97.5%**

G.2 QWEN-VL AND INTERNVL SERIES

The second longtable consolidates Qwen2.5-VL-7B, Qwen3-VL-8B, Qwen3.5-9B, and InternVL3-8B. These models share the same eight benchmark columns, followed by Acc and Rel. Qwen3.5 uses the fixed $K = 12$ configuration of Appendix C.5. VScan and SparseVLM are not reported on Qwen3.5: both prune inside the decoder at layers fixed by their released configurations, using those layers' self-attention maps, and in Qwen3.5 these layers fall on Gated DeltaNet (linear-attention) blocks that expose no attention maps. Moving their pruning layers to a full-attention block would change the methods as published, so we do not report them.

Table 21: Complete benchmark results for Qwen-VL and InternVL models, including MiCo-S1.

Method	AI2D	POPE	HallB	MME	MMB-EN	MMB-CN	MMStar	SQA-IMG	Acc	Rel
Qwen2.5-VL-7B										
Vanilla	84.9	87.7	55.9	2302.0	84.8	82.9	65.5	86.8	83.0	100.0%
Retain 256 tokens (80.2% pruned)										
FastV (ECCV'24)	78.4	83.0	49.1	2169.0	80.5	78.8	55.5	83.6	77.2	93.0%
SparseVLM (ICML'25)	77.6	82.9	46.4	2207.5	80.9	79.4	55.5	**87.7**	77.6	93.5%
DivPrune (CVPR'25)	81.2	85.3	46.6	2167.0	81.8	80.9	57.9	84.8	78.4	94.5%
VisionZip (CVPR'25)	76.0	82.1	42.1	2101.2	78.7	78.6	54.1	82.8	74.9	90.3%
VisPruner (ICCV'25)	77.7	84.1	41.8	2084.1	80.7	79.7	56.3	84.6	76.1	91.8%
VScan (TMLR'26)	79.1	**86.6**	49.7	2223.4	80.6	80.2	57.1	84.4	78.6	94.8%
HoloV (NeurIPS'25)	82.5	86.2	47.5	2229.1	81.8	81.0	58.9	85.1	79.3	95.6%
MMTok (ICLR'26)	81.0	85.7	47.9	2245.6	83.1	81.0	59.1	85.5	79.4	95.8%
ApET (CVPR'26)	77.8	82.6	44.4	2174.7	79.0	78.2	53.7	83.7	76.0	91.7%
MiCo-S1	81.7	86.1	48.6	2282.2	82.0	80.5	59.5	85.5	79.8	96.1%
MiCo	**83.8**	86.3	**51.2**	**2291.0**	**83.2**	**81.3**	**60.1**	86.2	**80.8**	**97.4%**
Retain 128 tokens (90.1% pruned)										
FastV (ECCV'24)	69.9	67.7	41.0	1597.0	66.4	68.8	43.7	79.6	64.6	77.9%
SparseVLM (ICML'25)	70.7	72.1	37.7	1815.3	72.9	74.1	45.7	81.9	68.2	82.3%
DivPrune (CVPR'25)	75.9	83.9	44.5	2044.0	79.2	78.6	52.3	82.8	74.9	90.3%
VisionZip (CVPR'25)	74.1	79.9	38.8	2013.4	79.3	76.4	52.7	80.9	72.8	87.8%
VisPruner (ICCV'25)	73.8	80.1	39.7	2007.8	78.7	77.1	52.3	82.1	73.0	88.0%
VScan (TMLR'26)	75.4	84.8	46.4	2024.5	79.2	78.0	53.5	82.9	75.2	90.6%
HoloV (NeurIPS'25)	74.3	80.3	39.3	1987.9	79.1	77.4	51.7	82.3	73.0	88.0%
MMTok (ICLR'26)	76.3	84.2	43.3	2120.7	79.8	77.5	53.7	81.4	75.3	90.8%
ApET (CVPR'26)	70.5	76.7	36.4	1911.3	74.6	73.4	47.7	79.3	69.3	83.5%
MiCo-S1	78.8	84.3	46.3	2176.4	80.6	78.4	54.6	83.5	76.9	92.7%
MiCo	**80.4**	**85.5**	**48.6**	**2252.4**	**80.8**	**79.1**	**56.5**	**84.6**	**78.5**	**94.7%**
Qwen3-VL-8B										
Vanilla	84.1	89.7	56.2	2406.0	86.3	86.1	67.1	94.6	85.6	100.0%
Retain 256 tokens										
FastV (ECCV'24)	69.3	81.7	43.0	1913.0	77.3	76.7	51.1	82.0	72.1	84.3%
SparseVLM (ICML'25)	78.8	86.8	46.5	2304.5	83.8	**83.7**	58.3	91.2	80.5	94.1%
DivPrune (CVPR'25)	80.0	89.4	46.9	2236.0	83.6	82.3	58.3	88.0	80.0	93.6%
VisionZip (CVPR'25)	81.4	89.0	50.4	2334.2	83.8	83.6	60.2	**92.0**	82.1	96.0%
VisPruner (ICCV'25)	79.9	89.2	47.7	2238.0	83.8	82.9	59.3	90.1	80.6	94.2%
VScan (TMLR'26)	80.2	87.8	**51.8**	2272.0	**84.6**	82.6	61.1	91.0	81.6	95.4%
HoloV (NeurIPS'25)	**82.0**	88.9	50.3	2305.2	83.4	83.5	61.2	91.6	82.0	95.9%
MMTok (ICLR'26)	81.6	89.3	49.6	2331.1	84.5	82.6	**61.4**	90.2	82.0	95.8%
ApET (CVPR'26)	78.6	89.5	47.0	2243.1	83.8	81.7	59.4	90.8	80.4	93.9%
MiCo-S1	81.8	89.3	51.3	2338.9	83.6	82.5	61.0	91.0	82.2	96.1%
MiCo	81.6	**89.6**	51.4	**2375.3**	84.3	83.6	59.7	91.4	**82.5**	**96.5%**
Retain 128 tokens										
FastV (ECCV'24)	66.3	57.2	30.3	1312.0	51.6	50.0	38.3	76.4	54.5	63.7%
SparseVLM (ICML'25)	72.2	81.2	40.1	1941.4	78.4	77.8	50.1	85.5	72.8	85.1%
DivPrune (CVPR'25)	74.4	88.3	42.6	2089.0	80.8	79.5	52.2	83.8	75.8	88.6%
VisionZip (CVPR'25)	75.8	86.7	41.5	2142.4	80.8	80.0	53.9	87.2	76.6	89.6%
VisPruner (ICCV'25)	72.4	86.3	40.5	1997.3	79.8	78.1	52.3	84.9	74.3	86.8%
VScan (TMLR'26)	76.2	85.6	44.1	2146.1	81.2	80.1	55.3	86.8	77.1	90.1%
HoloV (NeurIPS'25)	**79.5**	87.2	42.1	2126.0	81.4	79.8	**55.9**	85.7	77.2	90.3%
MMTok (ICLR'26)	77.5	**89.2**	46.0	2223.7	**82.3**	79.0	55.8	87.3	78.5	91.8%
ApET (CVPR'26)	71.5	89.1	42.9	2044.0	79.2	77.7	53.5	87.3	75.4	88.2%
MiCo-S1	77.6	88.4	46.5	**2281.7**	81.8	79.3	55.7	**88.2**	78.9	92.3%
MiCo	78.7	88.6	**48.3**	2268.6	81.5	**81.0**	55.4	88.1	**79.4**	**92.8%**

Continued on next page

Table 21 (continued): Complete benchmark results for Qwen-VL and InternVL models, including MiCo-S1

Method	AI2D	POPE	HallB	MME	MMB-EN	MMB-CN	MMStar	SQA-IMG	Acc	Rel
Qwen3.5-9B										
Vanilla	89.2	90.8	56.8	2381.5	86.4	85.7	69.8	95.2	86.6	100.0%
Retain 256 tokens										
FastV (ECCV'24)	73.9	84.1	42.8	1835.3	75.5	76.0	53.1	85.2	72.8	84.0%
DivPrune (CVPR'25)	86.6	90.4	50.6	2250.2	85.0	84.0	63.7	91.6	83.0	95.9%
VisionZip (CVPR'25)	87.1	90.4	54.0	**2352.7**	85.7	84.7	65.3	**94.1**	84.9	98.0%
VisPruner (ICCV'25)	86.7	90.1	54.4	2286.9	85.9	85.1	62.9	92.9	84.1	97.0%
HoloV (NeurIPS'25)	**87.4**	90.2	54.3	2340.3	85.7	84.6	65.5	93.8	84.8	97.9%
MMTok (ICLR'26)	87.0	90.7	54.7	2334.5	85.1	83.2	64.7	91.9	84.3	97.3%
ApET (CVPR'26)	87.3	**90.8**	53.4	2331.0	85.8	84.6	**67.5**	93.5	84.9	98.1%
MiCo-S1	87.2	90.5	54.9	2335.5	85.1	84.2	63.8	92.7	84.4	97.4%
MiCo	87.2	**90.8**	**55.5**	2350.7	**86.5**	**85.6**	67.1	93.8	**85.5**	**98.7%**
Retain 128 tokens										
FastV (ECCV'24)	67.0	62.7	31.5	1211.4	49.7	49.6	39.0	75.3	54.4	62.8%
DivPrune (CVPR'25)	82.0	89.7	44.6	2160.4	83.8	81.6	57.4	89.2	79.5	91.8%
VisionZip (CVPR'25)	82.1	88.4	46.8	2140.3	83.1	81.8	58.0	90.4	79.7	92.0%
VisPruner (ICCV'25)	81.5	88.1	44.0	2069.2	81.7	79.8	56.2	89.2	78.0	90.1%
HoloV (NeurIPS'25)	82.4	88.7	48.8	2109.5	82.6	82.5	59.2	89.7	79.9	92.3%
MMTok (ICLR'26)	83.7	90.5	47.8	2177.8	82.1	80.0	60.5	89.5	80.4	92.8%
ApET (CVPR'26)	80.2	**90.7**	49.4	2141.8	83.0	81.9	59.7	91.1	80.4	92.8%
MiCo-S1	83.6	89.7	49.8	2193.5	82.0	80.7	58.8	89.7	80.5	92.9%
MiCo	**85.6**	**90.7**	**51.4**	**2306.4**	**85.2**	**83.8**	**61.7**	**91.4**	**83.2**	**96.0%**
InternVL3-8B										
Vanilla	85.1	90.7	49.4	2369.0	85.7	85.1	68.3	97.8	85.1	100.0%
Retain 256 tokens										
FastV (ECCV'24)	80.5	89.1	44.0	**2289.0**	83.6	83.7	61.2	93.3	81.2	95.5%
SparseVLM (ICML'25)	75.4	86.3	36.7	2217.6	81.3	81.8	57.3	89.8	77.4	91.0%
DivPrune (CVPR'25)	80.3	89.8	43.0	2178.0	81.9	80.5	58.9	91.8	79.4	93.3%
VisionZip (CVPR'25)	76.0	87.0	40.0	2148.8	82.0	81.2	55.1	90.7	77.4	91.0%
VisPruner (ICCV'25)	72.3	88.9	37.6	2037.7	81.2	79.5	56.5	90.8	76.1	89.4%
VScan (TMLR'26)	76.7	87.6	41.4	2230.0	82.4	80.5	58.5	93.9	79.1	92.9%
HoloV (NeurIPS'25)	77.9	87.5	41.7	2194.0	82.1	80.8	57.0	91.8	78.6	92.4%
MMTok (ICLR'26)	80.2	90.0	41.1	2268.0	81.8	82.0	58.5	94.0	80.1	94.2%
ApET (CVPR'26)	77.9	90.5	42.5	2228.8	80.4	79.0	58.2	92.0	79.0	92.9%
MiCo-S1	80.1	**90.8**	44.1	2279.1	83.7	83.2	59.7	93.4	81.1	95.4%
MiCo	**82.9**	90.2	**44.4**	2276.1	**84.7**	**84.9**	**62.9**	**95.2**	**82.4**	**96.8%**
Retain 128 tokens										
FastV (ECCV'24)	68.4	78.7	38.9	1807.0	73.7	73.5	47.3	83.8	69.3	81.5%
SparseVLM (ICML'25)	69.7	79.6	32.4	1849.4	75.3	75.1	49.0	83.5	69.6	81.9%
DivPrune (CVPR'25)	74.2	88.8	37.6	2051.0	78.3	75.7	52.0	87.5	74.6	87.7%
VisionZip (CVPR'25)	68.5	81.5	32.7	1864.7	74.7	73.5	49.2	81.7	69.4	81.5%
VisPruner (ICCV'25)	68.5	84.4	33.0	1769.8	72.9	72.8	49.0	82.8	69.0	81.1%
VScan (TMLR'26)	71.9	84.7	37.8	2028.0	78.8	76.9	53.9	89.0	74.3	87.3%
HoloV (NeurIPS'25)	72.4	84.1	37.9	1968.0	78.4	77.4	53.2	86.7	73.6	86.5%
MMTok (ICLR'26)	74.7	88.6	37.9	2128.9	78.3	77.4	52.9	87.8	75.5	88.8%
ApET (CVPR'26)	74.3	**89.7**	38.1	2087.9	76.2	75.9	53.9	87.5	75.0	88.2%
MiCo-S1	73.9	89.3	36.8	2112.5	81.0	79.0	55.5	88.9	76.2	89.6%
MiCo	**77.4**	**89.7**	**40.7**	**2260.5**	**82.8**	**81.4**	**58.2**	**91.5**	**79.3**	**93.3%**

G.3 Refinement-Layer Sweep Tables

The following two long tables report all 204 LLaVA model–budget–layer settings: 15 layers for each 7B model and 19 for each 13B model, across three budgets. Acc_{10} is the equal-weight mean of GQA, SQA-IMG, TextVQA, POPE, MME/20, MMBench-EN, MMBench-CN, SEED, AI2D, and MMMU. These sweeps use POPE average F1, MME Perception, circular MMBench scoring, image-only SEED, and the recorded single-image MMMU protocol. Rel normalizes Acc_{10} by the corresponding unpruned reference. Aggregation preserves available source precision: reused scores may already be rounded, whereas newly measured scores retain evaluator precision. Small differences at the displayed precision should not be interpreted as distinct robust peaks. Pale red rows mark the original selected layers, and score highlighting marks the highest and second-highest distinct displayed values within each sweep.

Table 22: Complete refinement-layer sweeps for LLaVA-1.5 models across the three reported budgets.

K	GQA	SQA IMG	TextVQA	POPE	MME	MMB EN	MMB CN	SEED	AI2D	MMMU	Acc	Rel
LLaVA-1.5-7B $T = 32$												
Vanilla	61.9	69.5	58.2	85.9	1508.8	64.7	58.1	66.0	55.5	35.0	63.0	100.0%
1	56.1	69.0	54.3	80.4	1295.0	58.9	52.2	57.5	52.7	33.3	57.9	91.9%
2	55.6	68.6	54.7	79.7	1298.2	58.4	52.2	57.8	53.3	34.4	58.0	92.0%
3	55.6	68.8	54.9	79.0	1336.5	59.7	53.0	57.9	53.0	34.2	58.3	92.5%
4	55.5	68.8	54.6	78.7	1323.3	59.5	52.7	57.9	53.0	35.2	58.2	92.4%
5	55.7	68.8	54.6	78.7	1331.1	59.4	53.4	57.9	52.9	34.9	58.3	92.5%
6	55.6	68.7	55.0	78.8	1297.6	58.8	53.4	58.2	52.6	34.8	58.1	92.1%
7	55.9	68.7	**55.2**	80.9	1355.0	59.2	54.0	58.2	53.6	34.1	58.8	93.3%
8	56.0	68.6	54.6	81.2	1332.6	59.0	53.7	58.2	**53.7**	34.9	58.7	93.1%
9	56.6	68.9	54.6	82.1	1398.6	59.9	53.8	59.1	53.2	34.9	59.3	94.1%
10	55.9	68.3	54.4	80.4	1357.4	59.9	52.9	59.1	53.1	34.0	58.6	93.0%
11	56.3	68.9	54.6	83.2	1394.7	59.5	53.8	59.8	53.0	35.1	59.4	94.2%
12	**57.1**	68.5	54.6	**84.4**	**1405.3**	59.7	56.1	60.6	53.2	34.9	**59.9**	**95.1%**
13	56.4	68.9	54.6	82.9	1398.2	60.2	**56.3**	60.6	53.6	35.0	59.8	94.9%
14	54.7	**69.1**	54.0	81.6	1395.8	**60.5**	56.0	**60.7**	53.3	**35.4**	59.5	94.4%
15	55.1	69.0	54.3	83.4	1398.1	60.3	56.1	**60.7**	**53.7**	35.0	59.8	94.8%
LLaVA-1.5-7B $T = 64$												
Vanilla	61.9	69.5	58.2	85.9	1508.8	64.7	58.1	66.0	55.5	35.0	63.0	100.0%
1	58.4	69.0	55.8	84.7	1397.1	60.3	54.0	61.1	53.7	34.7	60.1	95.4%
2	58.2	**69.8**	56.3	83.7	1390.1	59.8	54.5	60.7	53.6	34.7	60.1	95.3%
3	57.7	69.7	56.5	82.7	1403.5	60.5	55.4	60.9	53.8	**35.4**	60.3	95.6%
4	57.7	69.7	**56.6**	83.3	1418.8	61.1	55.4	60.9	54.0	34.9	60.5	95.9%
5	57.9	69.0	56.1	83.0	1411.5	**61.6**	55.8	60.9	53.9	34.3	60.3	95.7%
6	57.7	69.7	56.3	83.3	1394.5	61.1	55.2	60.5	54.4	34.4	60.2	95.6%
7	58.1	69.7	56.4	84.7	1426.8	60.4	55.7	61.0	54.1	34.2	60.6	96.1%
8	57.9	69.2	56.1	84.7	1421.5	61.2	56.0	60.9	54.3	34.1	60.6	96.1%
9	57.9	69.0	56.2	85.0	1411.8	60.9	56.2	61.6	54.7	34.6	60.7	96.3%
10	57.9	68.8	56.1	84.5	1420.3	61.3	55.8	61.8	54.4	33.3	60.5	96.0%
11	58.1	69.3	56.4	86.2	1406.5	61.4	56.2	62.2	54.0	34.6	60.9	96.6%
12	**58.8**	68.5	56.5	**86.6**	1399.7	60.4	57.3	**62.6**	54.1	34.6	60.9	96.7%
13	58.5	69.0	56.0	85.9	1406.6	61.1	57.5	**62.6**	54.6	35.0	**61.0**	**96.9%**
14	57.8	68.9	55.2	85.4	**1434.2**	61.0	**57.6**	62.4	54.7	34.8	60.9	96.7%
15	58.2	68.7	56.1	85.9	1420.7	61.4	56.8	**62.6**	**54.8**	34.3	**61.0**	96.8%
LLaVA-1.5-7B $T = 128$												
Vanilla	61.9	69.5	58.2	85.9	1508.8	64.7	58.1	66.0	55.5	35.0	63.0	100.0%
1	59.7	68.9	56.8	86.0	1445.0	62.7	56.5	63.8	54.6	34.9	61.6	97.8%
2	59.5	**69.5**	57.5	85.0	1463.7	62.4	56.4	63.6	55.0	35.1	61.7	97.9%
3	59.6	69.4	57.2	84.8	1460.7	62.5	56.6	63.9	55.0	35.2	61.7	97.9%
4	59.2	69.4	57.5	84.8	1472.3	62.1	56.8	63.9	**55.5**	35.6	61.8	98.1%
5	59.7	69.3	57.5	84.6	1466.9	62.3	57.2	63.4	55.3	35.9	61.9	98.2%
6	59.2	69.1	**57.7**	85.6	1457.3	61.9	56.7	63.5	54.9	35.6	61.7	97.9%
7	60.0	69.3	57.6	85.9	1474.8	62.4	56.5	63.7	54.9	36.3	62.0	98.5%
8	59.5	69.3	**57.7**	86.1	1481.4	62.5	56.9	63.6	54.9	36.6	62.1	98.6%
9	59.9	68.9	57.2	86.2	1476.9	62.5	56.9	64.0	54.8	36.1	62.0	98.4%
10	60.2	68.5	57.2	85.7	1463.2	**63.0**	56.6	64.1	54.7	36.2	61.9	98.3%
11	60.0	68.6	57.1	86.6	1471.9	62.6	57.2	64.4	55.0	**37.0**	62.2	98.7%
12	60.4	68.5	**57.7**	**87.0**	1461.5	**63.0**	**57.6**	64.6	55.2	**37.0**	**62.4**	**99.0%**
13	**60.5**	68.0	57.1	86.8	1477.5	62.9	57.5	**64.8**	55.4	36.1	62.3	98.8%
14	59.6	68.7	56.7	86.3	**1487.3**	62.6	57.4	**64.8**	55.4	36.2	62.2	98.7%
15	60.2	68.8	56.8	86.4	1465.9	62.9	57.3	**64.8**	55.4	36.3	62.2	98.7%
LLaVA-1.5-13B $T = 32$												
Vanilla	63.3	72.8	61.2	86.0	1533.2	68.5	63.5	68.2	60.8	36.4	65.7	100.0%
1	56.7	72.3	56.7	79.7	1370.0	63.1	60.3	60.6	57.2	34.1	60.9	92.7%
2	57.3	72.5	57.2	78.4	1413.2	64.4	60.1	60.8	57.4	35.7	61.4	93.5%
3	57.2	72.6	57.2	78.5	1417.1	65.0	59.6	60.6	56.7	34.7	61.3	93.2%
4	57.1	72.4	57.0	78.0	1430.7	63.9	60.2	60.7	56.9	34.3	61.2	93.1%
5	57.1	72.9	56.8	78.7	1416.5	64.3	60.1	60.7	56.5	35.3	61.3	93.3%
6	56.9	72.8	56.9	79.9	1428.1	64.2	60.1	60.8	57.4	35.0	61.5	93.6%
7	57.2	72.2	57.5	80.1	1436.6	64.4	60.4	61.1	56.9	35.6	61.7	93.9%
8	57.3	73.0	57.0	82.3	**1466.2**	64.6	61.0	61.5	57.4	35.9	62.3	94.9%
9	57.7	**74.0**	57.1	82.3	1418.8	64.9	60.3	61.9	57.0	35.7	62.2	94.6%
10	57.2	72.7	57.4	82.1	1453.5	64.7	60.3	62.2	**57.5**	**36.6**	62.3	94.8%
11	**58.4**	73.0	57.4	82.1	1465.1	65.6	61.7	63.3	57.2	36.0	**62.8**	**95.5%**
12	58.1	73.1	**57.8**	81.8	1444.1	65.3	61.8	63.2	**57.5**	35.6	62.6	95.3%
13	57.8	72.9	57.5	**82.6**	1448.6	65.4	61.2	63.1	57.3	35.7	62.6	95.2%
14	58.0	72.8	57.7	82.4	1452.9	65.1	61.7	63.2	57.4	35.0	62.6	95.2%

Continued on next page

Table 22 (continued): LLaVA-1.5 refinement-layer sweeps

K	GQA	SQA IMG	TextVQA	POPE	MME	MMB EN	MMB CN	SEED	AI2D	MMMU	Acc	Rel
15	57.6	73.0	**57.8**	82.3	1444.7	65.5	61.9	63.3	**57.5**	34.8	62.6	95.2%
16	57.7	72.9	**57.8**	**82.6**	1438.1	**65.7**	61.9	63.3	**57.5**	34.9	62.6	95.3%
17	57.7	73.1	57.7	82.3	1437.5	65.6	**62.0**	63.3	**57.5**	34.6	62.6	95.2%
18	57.4	73.1	57.4	82.4	1437.4	65.4	61.9	**63.4**	57.4	34.7	62.5	95.1%
19	56.5	73.1	56.2	82.4	1444.9	65.5	61.9	63.3	57.4	35.6	62.4	94.9%
LLaVA-1.5-13B $T = 64$												
Vanilla	63.3	72.8	61.2	86.0	1533.2	68.5	63.5	68.2	60.8	36.4	65.7	100.0%
1	58.9	72.8	58.4	84.5	1468.0	66.3	61.0	63.7	57.9	35.0	63.2	96.1%
2	58.7	73.6	58.5	83.0	1467.5	65.9	61.3	63.3	**58.0**	36.2	63.2	96.1%
3	58.7	73.9	58.5	82.9	1455.7	65.6	61.7	63.4	57.7	36.0	63.1	96.0%
4	58.6	**74.2**	58.5	82.7	1474.1	65.7	61.5	63.3	57.3	35.7	63.1	96.0%
5	58.5	73.7	58.6	83.6	1486.6	65.7	61.3	63.5	57.3	36.3	63.3	96.3%
6	58.5	73.7	58.4	84.7	**1498.8**	65.8	61.4	63.2	57.3	36.4	63.4	96.5%
7	58.7	73.2	**58.8**	84.6	1472.2	66.2	61.4	63.5	57.5	36.2	63.4	96.4%
8	58.7	73.2	58.4	**86.4**	1490.8	65.9	61.1	63.9	57.1	36.4	63.6	96.7%
9	58.7	73.6	58.2	86.0	1470.3	66.1	61.7	64.1	57.6	36.4	63.6	96.7%
10	58.7	73.0	58.6	85.5	1449.7	66.0	61.9	64.4	57.9	36.8	63.5	96.6%
11	**59.5**	73.5	58.5	85.7	1478.4	66.2	62.0	**64.9**	57.7	**37.2**	**63.9**	**97.2%**
12	59.3	73.3	58.5	85.8	1464.4	66.7	62.2	64.6	57.3	36.3	63.7	96.9%
13	59.1	73.9	58.3	85.6	1470.1	66.6	62.2	64.6	57.0	36.3	63.7	96.9%
14	59.3	73.7	58.6	85.6	1455.9	66.6	62.3	64.6	57.1	**37.2**	63.8	97.0%
15	59.1	73.9	58.4	85.5	1453.3	**66.8**	**62.7**	64.6	57.1	36.4	63.7	96.9%
16	59.1	73.7	58.4	85.9	1448.3	66.6	62.6	64.5	57.1	36.9	63.7	96.9%
17	58.8	73.7	58.6	85.7	1461.6	66.7	62.3	64.6	57.2	36.3	63.7	96.9%
18	58.7	73.8	58.2	85.6	1446.6	66.6	62.6	64.6	57.1	35.8	63.5	96.6%
19	58.0	73.8	57.3	85.8	1458.3	66.6	62.4	64.7	57.2	35.4	63.4	96.5%
LLaVA-1.5-13B $T = 128$												
Vanilla	63.3	72.8	61.2	86.0	1533.2	68.5	63.5	68.2	60.8	36.4	65.7	100.0%
1	59.8	73.1	59.0	86.1	1515.3	66.6	62.2	65.3	58.6	36.3	64.3	97.8%
2	59.9	**74.4**	59.6	84.8	**1524.0**	67.0	62.5	65.4	57.8	**36.9**	64.5	98.1%
3	59.8	74.3	59.2	84.9	1516.5	66.6	62.5	65.5	57.6	36.6	64.3	97.8%
4	59.8	74.3	59.4	84.9	1511.8	66.4	61.9	65.4	58.2	36.2	64.2	97.7%
5	59.6	73.7	59.7	85.8	1518.7	66.6	62.3	65.6	58.0	36.3	64.4	97.9%
6	59.9	73.5	59.6	86.1	1523.6	67.0	62.6	65.3	57.8	36.7	64.5	98.1%
7	59.9	73.3	**59.9**	86.1	1512.4	67.4	**63.1**	65.5	58.1	36.2	64.5	98.1%
8	60.3	73.7	59.6	86.6	1509.3	67.0	62.4	66.0	58.1	36.4	**64.6**	98.3%
9	60.2	73.8	59.7	**87.2**	1482.1	67.2	62.8	66.1	58.5	36.1	**64.6**	98.2%
10	60.1	73.7	**59.9**	86.6	1505.3	66.7	61.9	66.0	58.6	35.9	64.5	98.1%
11	60.2	73.2	59.8	86.7	1500.5	**67.5**	62.5	66.3	**58.7**	36.2	**64.6**	**98.3%**
12	**60.5**	73.0	59.7	86.5	1505.7	67.1	62.5	66.2	58.5	36.3	**64.6**	98.2%
13	59.9	73.3	59.6	86.7	1494.3	66.9	62.0	66.2	58.3	36.0	64.4	97.9%
14	60.1	73.2	59.5	86.7	1488.9	67.0	62.6	**66.4**	58.3	36.0	64.4	98.0%
15	60.1	73.5	59.4	86.5	1494.1	67.2	62.7	**66.4**	58.4	36.6	**64.6**	98.2%
16	59.8	73.5	59.4	86.6	1491.3	67.3	62.7	**66.4**	58.4	35.7	64.4	98.0%
17	59.9	73.5	59.5	86.5	1491.1	67.3	62.7	**66.4**	58.5	36.2	64.5	98.1%
18	59.7	73.3	59.3	86.5	1490.4	67.3	62.9	**66.4**	58.4	36.0	64.4	98.0%
19	59.5	73.3	58.7	86.5	1495.9	67.3	62.7	**66.4**	58.4	36.0	64.4	97.9%

Table 23: Complete refinement-layer sweeps for LLaVA-NeXT models across the three reported budgets.

K	GQA	SQA IMG	TextVQA	POPE	MME	MMB EN	MMB CN	SEED	AI2D	MMMU	Acc	Rel
LLaVA-NeXT-7B $T = 160$												
Vanilla	62.5	67.5	60.3	86.8	1511.8	65.8	57.3	69.7	64.7	35.2	64.5	100.0%
1	59.4	68.0	54.3	82.3	1379.2	62.4	54.3	63.4	62.8	36.8	61.3	94.9%
2	59.5	67.3	55.9	81.9	1388.0	62.6	54.9	63.4	63.0	**37.0**	61.5	95.3%
3	58.8	67.7	56.3	81.0	1442.7	62.8	55.6	63.5	62.6	35.3	61.6	95.4%
4	59.1	68.5	56.1	81.5	1419.2	62.3	55.1	63.5	62.9	35.3	61.5	95.3%
5	59.2	67.8	56.1	82.0	1420.9	63.0	55.9	63.3	62.4	34.8	61.6	95.4%
6	59.1	67.9	56.5	81.5	1442.7	63.4	55.8	63.7	63.0	34.8	61.8	95.7%
7	59.5	68.2	**56.8**	83.8	1446.1	64.1	56.6	64.4	63.5	35.2	62.4	96.7%
8	59.5	68.6	56.1	84.2	1454.2	64.0	56.6	64.2	63.3	35.7	62.5	96.8%
9	59.9	**68.7**	56.3	83.8	1434.7	63.4	55.9	64.3	63.3	35.7	62.3	96.5%
10	59.6	68.6	55.4	83.2	1432.8	63.7	55.9	64.4	63.4	35.1	62.1	96.2%
11	59.6	67.6	55.6	**85.5**	1445.3	63.9	56.1	64.7	**64.2**	35.2	62.5	96.8%
12	**60.1**	68.4	56.4	85.0	1444.2	64.0	56.4	65.4	63.4	35.2	62.6	**97.1%**
13	**60.1**	68.4	56.0	84.4	1452.1	64.2	56.6	65.3	63.4	35.6	**62.7**	**97.1%**

Continued on next page

Table 23 (continued): LLaVA-NeXT refinement-layer sweeps

K	GQA	SQA IMG	TextVQA	POPE	MME	MMB EN	MMB CN	SEED	AI2D	MMMU	Acc	Rel
14	59.4	68.5	54.2	83.6	1465.3	**64.3**	56.6	65.3	63.2	35.2	62.4	96.6%
15	59.8	68.3	54.2	84.2	**1467.6**	**64.3**	**57.2**	**65.5**	63.5	35.3	62.6	97.0%
LLaVA-NeXT-7B $T = 320$												
Vanilla	62.5	67.5	60.3	86.8	1511.8	65.8	57.3	69.7	64.7	35.2	64.5	100.0%
1	60.8	68.2	55.6	85.4	1407.0	64.1	55.7	65.5	64.1	36.0	62.6	97.0%
2	61.1	67.9	57.4	85.3	1420.2	65.6	57.0	65.7	**64.9**	**36.4**	63.2	98.0%
3	60.4	67.5	57.3	84.6	1413.3	64.4	57.2	65.8	64.5	35.7	62.8	97.3%
4	61.1	68.1	57.5	85.6	1419.0	64.5	57.5	65.8	64.5	35.1	63.1	97.7%
5	60.9	67.3	57.4	85.2	1418.6	65.0	57.0	65.7	64.6	35.3	62.9	97.5%
6	61.0	67.9	57.5	85.1	1437.6	65.6	57.4	65.9	64.5	35.8	63.3	98.0%
7	61.3	68.5	58.3	86.4	1437.0	65.5	57.7	66.4	64.8	35.1	63.6	98.5%
8	61.1	68.8	**58.7**	86.9	1444.6	65.2	57.7	66.4	64.3	35.0	63.6	98.6%
9	61.3	68.8	57.9	86.8	1443.9	65.1	57.5	66.6	64.6	34.6	63.5	98.4%
10	**61.5**	68.4	57.2	85.8	1445.5	64.8	57.3	66.6	64.3	35.0	63.3	98.1%
11	61.3	68.1	58.0	**87.5**	1458.1	65.3	57.2	66.9	64.4	35.0	63.7	98.6%
12	**61.5**	68.4	57.9	86.9	1465.3	**66.0**	57.9	67.4	64.8	35.6	**64.0**	**99.1%**
13	**61.5**	**68.9**	57.8	86.7	1455.9	**66.0**	**58.2**	67.4	64.6	35.1	63.9	99.0%
14	60.9	68.7	56.6	86.1	**1470.4**	65.9	57.6	67.4	64.5	35.0	63.6	98.6%
15	61.2	68.4	56.5	86.5	1470.2	65.7	**58.2**	**67.5**	**64.9**	34.8	63.7	98.7%
LLaVA-NeXT-7B $T = 640$												
Vanilla	62.5	67.5	60.3	86.8	1511.8	65.8	57.3	69.7	64.7	35.2	64.5	100.0%
1	**62.3**	68.3	56.5	87.3	1445.1	64.5	56.5	67.9	65.0	35.0	63.6	98.5%
2	62.1	68.7	58.1	87.1	1455.7	65.3	56.9	68.2	64.7	**35.3**	63.9	99.0%
3	62.0	68.5	58.0	87.1	1440.8	64.9	57.2	68.5	65.1	35.1	63.8	98.9%
4	62.1	68.9	57.8	87.3	1463.9	65.1	58.2	68.4	65.5	**35.3**	64.2	99.4%
5	61.9	68.9	57.9	87.1	1431.7	65.1	57.7	68.4	65.2	34.8	63.9	98.9%
6	62.1	68.4	58.2	87.1	1455.9	65.1	57.7	68.1	65.2	35.0	64.0	99.1%
7	61.9	68.2	58.9	87.3	1446.8	65.5	57.7	68.4	65.3	35.0	64.1	99.2%
8	61.9	68.3	59.2	87.8	1454.8	65.3	57.9	68.3	**65.8**	34.6	64.2	99.4%
9	62.2	68.7	**59.3**	87.9	1436.3	65.4	58.2	68.3	65.6	34.9	64.2	99.5%
10	62.2	69.1	58.2	87.5	1437.1	65.2	57.6	68.4	**65.8**	34.6	64.0	99.2%
11	62.0	68.7	59.0	**88.0**	1465.9	65.6	58.1	68.7	65.5	34.1	64.3	99.6%
12	**62.3**	**69.4**	59.2	87.8	1472.6	65.4	58.2	69.0	65.3	34.0	**64.4**	**99.8%**
13	62.2	68.9	58.9	87.5	1489.4	**65.9**	57.9	69.2	65.5	33.8	**64.4**	**99.8%**
14	62.0	69.0	57.8	87.3	1495.8	65.8	57.7	69.1	65.0	34.0	64.3	99.6%
15	61.9	69.0	57.7	87.3	**1504.0**	**65.9**	**58.4**	**69.3**	65.3	33.7	**64.4**	99.7%
LLaVA-NeXT-13B $T = 160$												
Vanilla	64.4	73.1	63.2	85.3	1539.5	68.5	61.2	71.6	70.1	35.7	67.0	100.0%
1	60.7	72.3	56.9	83.8	1446.2	64.4	60.1	65.0	66.8	37.1	63.9	95.4%
2	60.5	71.0	58.4	82.3	1446.4	64.5	62.3	65.1	67.7	36.8	64.1	95.6%
3	60.4	71.5	58.2	82.1	1453.7	64.9	61.9	65.4	67.7	37.1	64.2	95.8%
4	60.4	71.1	58.2	82.7	1430.8	65.4	62.3	65.5	67.5	37.0	64.2	95.8%
5	60.4	71.7	58.3	82.4	1469.6	65.4	62.5	65.8	67.6	37.3	64.5	96.3%
6	60.8	72.2	58.3	83.1	1492.0	66.3	62.3	66.0	67.8	36.6	64.8	96.7%
7	60.7	**72.4**	57.9	83.7	1488.9	65.4	62.8	66.4	67.5	37.4	64.9	96.8%
8	61.1	72.3	58.2	84.7	1482.5	65.7	62.3	66.5	67.4	36.9	64.9	96.9%
9	61.6	72.0	**58.5**	85.0	1500.6	66.1	63.0	67.3	67.7	**37.6**	**65.4**	**97.6%**
10	61.5	71.7	58.2	**85.4**	**1505.0**	65.7	63.1	67.0	68.0	36.6	65.3	97.4%
11	62.1	71.2	58.4	84.9	1482.5	66.2	62.8	67.5	68.1	36.8	65.2	97.3%
12	61.8	71.5	**58.5**	84.8	1491.9	66.3	62.5	67.4	68.0	37.0	65.3	97.4%
13	61.6	71.6	58.2	84.8	1478.8	**66.6**	**63.2**	67.4	67.9	36.8	65.2	97.3%
14	62.0	71.8	58.4	85.0	1488.4	66.2	63.1	67.6	68.0	36.7	65.3	97.5%
15	**62.2**	71.7	58.2	85.0	1480.9	66.0	62.9	**67.7**	68.0	36.9	65.3	97.4%
16	62.0	71.7	**58.5**	85.1	1486.4	66.0	62.8	**67.7**	68.2	36.7	65.3	97.4%
17	62.1	71.7	**58.5**	85.0	1489.6	66.0	63.0	**67.7**	68.1	36.9	65.3	97.5%
18	62.1	71.5	58.3	84.9	1473.8	66.0	62.8	**67.7**	**68.3**	36.7	65.2	97.3%
19	61.8	71.5	57.5	84.6	1482.6	66.0	62.8	**67.7**	68.2	36.6	65.1	97.1%
LLaVA-NeXT-13B $T = 320$												
Vanilla	64.4	73.1	63.2	85.3	1539.5	68.5	61.2	71.6	70.1	35.7	67.0	100.0%
1	62.3	71.4	59.1	85.4	1483.6	65.5	61.2	67.8	67.7	**38.2**	65.3	97.4%
2	62.5	71.2	60.8	84.1	1506.8	66.6	62.4	68.1	67.8	37.4	65.6	97.9%
3	62.4	71.0	60.6	84.5	1501.0	66.2	62.2	68.0	67.9	37.2	65.5	97.8%
4	62.2	70.8	60.4	84.7	1507.8	66.2	62.0	67.8	67.8	38.1	65.6	97.8%
5	62.4	71.0	60.5	84.5	1507.5	66.5	62.8	68.2	67.5	37.9	65.6	98.0%
6	62.4	71.7	60.5	85.3	1512.3	67.3	63.0	68.5	68.2	37.3	66.0	98.5%
7	62.7	**72.5**	60.6	86.0	1513.7	**67.9**	63.0	68.7	68.2	37.1	66.2	98.9%
8	62.8	71.9	60.1	86.2	1514.3	67.3	62.7	68.8	68.3	38.1	66.2	98.8%
9	**63.3**	72.0	60.6	86.9	1526.7	67.6	63.5	69.2	68.3	37.7	**66.5**	**99.3%**

Continued on next page

Table 23 (continued): LLaVA-NeXT refinement-layer sweeps

K	GQA	SQA IMG	TextVQA	POPE	MME	MMB EN	MMB CN	SEED	AI2D	MMMU	Acc	Rel
10	62.8	72.2	60.4	**87.0**	1541.0	67.7	62.7	69.3	68.5	37.6	**66.5**	**99.3%**
11	63.0	71.6	60.4	86.7	**1551.1**	67.7	**63.7**	69.4	68.2	37.0	**66.5**	**99.3%**
12	63.2	71.4	60.6	86.5	1523.9	67.6	**63.7**	69.2	68.5	36.9	66.4	99.0%
13	62.9	71.5	60.4	86.5	1533.5	67.7	63.5	69.3	68.4	37.1	66.4	99.1%
14	**63.3**	71.3	60.9	86.5	1543.3	**67.9**	63.2	**69.5**	68.3	37.0	**66.5**	**99.3%**
15	63.2	71.8	60.7	86.5	1539.2	67.8	63.4	69.4	68.4	36.9	**66.5**	99.2%
16	63.0	71.4	60.6	86.7	1537.1	67.7	63.1	**69.5**	68.4	36.6	66.4	99.1%
17	63.1	71.7	**61.2**	86.6	1536.6	67.4	63.1	**69.5**	**68.6**	37.2	**66.5**	**99.3%**
18	62.9	71.6	60.4	86.4	1535.8	67.3	63.1	**69.5**	68.5	37.4	66.4	99.1%
19	63.0	71.5	59.8	86.5	1547.4	67.5	63.1	**69.5**	68.5	36.3	66.3	99.0%
					LLaVA-NeXT-13B	$T = 640$						
Vanilla	64.4	73.1	63.2	85.3	1539.5	68.5	61.2	71.6	70.1	35.7	67.0	100.0%
1	63.6	71.7	60.2	86.2	1534.2	66.6	62.9	70.2	68.6	**38.7**	66.5	99.3%
2	63.6	72.4	61.1	85.9	1514.2	67.9	62.5	70.4	69.1	37.6	66.6	99.4%
3	63.4	72.2	61.3	85.6	1536.8	67.3	62.7	70.8	69.3	37.0	66.7	99.5%
4	63.5	72.3	60.7	86.0	1537.3	68.0	63.0	70.3	69.0	37.7	66.7	99.6%
5	63.4	72.3	61.2	85.9	1552.9	67.6	62.7	70.5	68.9	37.7	66.8	99.7%
6	63.5	72.5	61.4	86.4	1543.4	68.6	62.9	70.4	69.1	37.6	67.0	99.9%
7	63.7	72.7	61.4	86.8	1531.1	68.5	63.0	70.7	69.5	37.8	67.1	100.1%
8	63.7	72.5	61.2	86.9	1524.7	68.4	63.3	70.4	69.4	37.6	67.0	99.9%
9	64.0	73.1	61.6	87.1	1547.1	68.3	63.2	71.1	69.8	37.6	67.3	100.4%
10	63.7	73.1	61.7	**87.2**	1545.8	69.0	63.4	71.1	69.8	36.4	67.3	100.4%
11	64.1	**73.3**	61.6	87.1	1543.8	68.8	63.5	71.2	69.9	36.4	67.3	100.5%
12	64.1	**73.3**	**61.9**	86.8	1543.7	68.6	63.4	71.2	69.8	37.1	67.3	100.5%
13	64.1	72.4	61.8	87.0	1567.0	**69.1**	62.7	71.2	69.9	37.0	67.4	100.5%
14	**64.3**	73.2	**61.9**	87.0	**1573.9**	69.0	63.4	71.4	70.2	37.2	**67.6**	**100.9%**
15	64.2	73.1	61.6	87.0	1558.0	69.0	63.4	71.4	70.1	37.6	67.5	100.8%
16	64.1	73.0	61.6	86.9	1567.1	68.7	**63.6**	71.4	70.2	37.4	67.5	100.8%
17	64.2	73.1	61.6	86.8	1558.8	69.0	63.5	**71.5**	70.2	37.7	67.5	100.8%
18	64.0	73.2	61.5	86.8	1566.4	68.8	63.4	**71.5**	70.3	37.4	67.5	100.8%
19	63.9	73.1	60.7	86.8	1565.7	68.8	63.5	**71.5**	**70.4**	37.3	67.4	100.6%

H Supplementary Benchmark Results

H.1 VTC-Bench: Fixed-Reference Groups

VTC-Bench evaluates visual-token compression by separating questions according to their sensitivity to image downsampling (Liao et al., 2026). We use the author-released groups constructed with Qwen2-VL-7B-Instruct. For each reference downsampling factor d, Group A contains samples answered correctly at full resolution but incorrectly after downsampling, while Group B contains samples answered correctly in both settings. These groups are reused unchanged for Qwen2.5-VL-7B.

Tables 24 and 25 report Qwen2.5-VL-7B at $T = 256$ and 128, respectively. Following the original VTC-Bench layout, each cell shows Group A (Group B) on one line and their absolute gap Δ below. Scores and gaps are rounded independently from the unrounded results to one decimal place; gaps are in percentage points. Each table is organized by reference image-area reduction, with the two groups ranked independently. Both tables include all eight task scores, their unweighted mean Acc, and Rel normalized by vanilla for the same model, group, and d. Task scores, Acc, and Rel are percentages; Rel is computed before rounding, and score highlighting ranks distinct displayed values within each T, d setting. MME uses per-question correctness rather than the native summed score, and MMBench uses per-question correctness rather than circular evaluation.

The reference factors $d = 2, 3, 4, 5, 10$ correspond to image-area reductions of $75.00\%, 88.89\%, 93.75\%, 96.00\%, 99.00\%$ when constructing the groups; they do not specify the pruning ratios of our target models. We evaluate the target model on the original images with token budgets $T = 256$ and 128, using greedy decoding and a maximum of 2,048 generated tokens. Empty and length-truncated responses remain in the evaluation denominators.

Table 24: VTC-Bench scores (%) on Qwen2.5-VL-7B at $T = 256$. Cells show Group A (Group B), with the absolute gap Δ in percentage points below.

Method	Chart QA	GQA	MMB CN	MMB EN	MME	MMStar	OCR Bench	POPE	Acc	Rel
Reference image-area reduction: 75.00%										
Vanilla	88.5 (94.3)	65.0 (88.2)	67.4 (94.7)	76.5 (97.9)	70.4 (93.9)	75.4 (86.7)	83.8 (90.4)	81.0 (96.7)	76.0 (92.8)	100.0 (100.0)
	Δ 5.8	Δ 23.2	Δ 27.3	Δ 21.4	Δ 23.5	Δ 11.3	Δ 6.6	Δ 15.6	Δ 16.8	Δ 0.0
FastV	57.8 (83.7)	54.7 (77.3)	58.7 (93.5)	65.9 (95.9)	62.0 (91.8)	55.2 (76.2)	43.0 (64.8)	65.8 (91.5)	57.9 (84.3)	76.1 (90.8)
	Δ 25.9	Δ 22.6	Δ 34.8	Δ 30.0	Δ 29.9	Δ 21.0	Δ 21.8	Δ 25.7	Δ 26.5	Δ 14.7
SparseVLM	26.2 (51.5)	50.4 (78.8)	59.6 (93.7)	71.0 (95.4)	60.6 (92.5)	53.0 (73.0)	19.0 (44.3)	70.2 (90.8)	51.2 (77.5)	67.4 (83.5)
	Δ 25.3	Δ 28.4	Δ 34.2	Δ 24.5	Δ 31.9	Δ 20.0	Δ 25.3	Δ 20.6	Δ 26.3	Δ 16.1
DivPrune	57.4 (78.0)	**60.3** (86.1)	66.5 (93.9)	68.7 (96.6)	66.2 (92.4)	59.6 (79.8)	53.1 (72.3)	75.4 (**95.7**)	63.4 (86.8)	83.4 (93.5)
	Δ 20.6	Δ 25.7	Δ 27.4	Δ 27.9	Δ 26.2	Δ 20.3	Δ 19.2	Δ 20.3	Δ 23.5	Δ 10.1
VisionZip	41.4 (65.9)	57.6 (85.6)	62.2 (93.7)	70.0 (96.5)	62.0 (92.5)	55.7 (77.3)	35.2 (60.4)	64.6 (93.2)	56.1 (83.1)	73.8 (89.5)
	Δ 24.6	Δ 28.0	Δ 31.6	Δ 26.4	Δ 30.5	Δ 21.5	Δ 25.2	Δ 28.7	Δ 27.0	Δ 15.7
VisPruner	46.3 (67.5)	58.1 (85.7)	60.4 (93.6)	68.2 (96.4)	62.0 (92.3)	56.3 (78.0)	34.6 (59.2)	66.7 (93.8)	56.6 (83.3)	74.4 (89.7)
	Δ 21.2	Δ 27.6	Δ 33.2	Δ 28.2	Δ 30.3	Δ 21.7	Δ 24.6	Δ 27.2	Δ 26.8	Δ 15.3
HoloV	40.7 (65.0)	59.1 (85.6)	61.3 (93.8)	67.3 (96.5)	63.4 (92.5)	59.6 (78.2)	33.0 (61.1)	69.4 (94.4)	56.7 (83.4)	74.6 (89.8)
	Δ 24.2	Δ 26.5	Δ 32.5	Δ 29.2	Δ 29.1	Δ 18.6	Δ 28.2	Δ 25.0	Δ 26.7	Δ 15.2
MMTok	56.7 (76.7)	**60.3** (**86.9**)	**67.4** (94.3)	71.4 (97.2)	**67.6** (93.4)	63.4 (79.7)	54.7 (67.0)	71.7 (95.4)	64.2 (86.3)	84.4 (93.0)
	Δ 20.0	Δ 26.5	Δ 26.9	Δ 25.8	Δ 25.8	Δ 16.3	Δ 12.3	Δ 23.7	Δ 22.2	Δ 8.6
ApET	37.1 (60.6)	54.2 (83.5)	59.1 (93.7)	64.5 (95.6)	63.4 (92.0)	52.5 (72.4)	41.9 (60.2)	63.8 (92.4)	54.5 (81.3)	71.8 (87.6)
	Δ 23.5	Δ 29.4	Δ 34.6	Δ 31.1	Δ 28.6	Δ 19.9	Δ 18.3	Δ 28.6	Δ 26.8	Δ 15.8
MiCo	**74.8** (**87.5**)	59.9 (**86.9**)	65.2 (**94.6**)	**72.4** (**97.5**)	**67.6** (**93.7**)	**66.7** (**82.0**)	**65.9** (**75.6**)	**75.6** (95.6)	**68.5** (**89.2**)	**90.1** (**96.1**)
	Δ 12.7	Δ 27.0	Δ 29.4	Δ 25.1	Δ 26.1	Δ 15.3	Δ 9.7	Δ 19.9	Δ 20.7	Δ 5.9
Reference image-area reduction: 88.89%										
Vanilla	91.9 (95.6)	71.5 (88.5)	77.2 (94.8)	86.7 (97.9)	85.7 (93.6)	77.9 (87.3)	87.2 (90.2)	82.6 (97.2)	82.6 (93.1)	100.0 (100.0)
	Δ 3.7	Δ 17.0	Δ 17.6	Δ 11.2	Δ 7.9	Δ 9.4	Δ 3.0	Δ 14.6	Δ 10.6	Δ 0.0
FastV	74.5 (85.8)	54.7 (79.2)	66.9 (94.2)	76.6 (96.3)	75.2 (91.9)	60.9 (76.7)	50.7 (66.6)	69.2 (92.4)	66.1 (85.4)	80.0 (91.7)
	Δ 11.3	Δ 24.4	Δ 27.2	Δ 19.6	Δ 16.7	Δ 15.8	Δ 15.9	Δ 23.1	Δ 19.3	Δ 11.6
SparseVLM	38.7 (63.1)	52.9 (80.6)	68.3 (94.3)	75.9 (96.2)	76.7 (92.4)	56.1 (74.6)	24.0 (49.1)	71.2 (91.7)	58.0 (80.2)	70.2 (86.2)
	Δ 24.4	Δ 27.7	Δ 26.0	Δ 20.3	Δ 15.7	Δ 18.5	Δ 25.1	Δ 20.5	Δ 22.3	Δ 16.0
DivPrune	69.0 (83.8)	66.2 (86.8)	73.4 (94.4)	80.2 (96.7)	78.9 (92.4)	67.2 (79.4)	58.5 (74.9)	**78.8** (**96.3**)	71.5 (88.1)	86.6 (94.6)
	Δ 14.8	Δ 20.7	Δ 21.0	Δ 16.6	Δ 13.4	Δ 12.2	Δ 16.4	Δ 17.5	Δ 16.6	Δ 8.0
VisionZip	54.0 (75.9)	64.5 (86.2)	70.2 (94.2)	81.4 (96.5)	81.2 (92.1)	62.0 (77.6)	36.8 (67.7)	65.0 (94.6)	64.4 (85.6)	78.0 (91.9)
	Δ 21.9	Δ 21.7	Δ 24.1	Δ 15.1	Δ 10.9	Δ 15.6	Δ 30.9	Δ 29.6	Δ 21.2	Δ 13.9
VisPruner	57.3 (75.9)	64.3 (86.5)	69.6 (94.1)	80.7 (96.5)	77.4 (92.2)	62.0 (78.7)	38.0 (65.1)	67.6 (95.0)	64.6 (85.5)	78.3 (91.8)
	Δ 18.6	Δ 22.1	Δ 24.4	Δ 15.8	Δ 14.7	Δ 16.7	Δ 27.1	Δ 27.4	Δ 20.9	Δ 13.5
HoloV	53.6 (73.7)	65.0 (86.4)	71.2 (94.1)	77.6 (96.8)	77.4 (92.5)	63.5 (79.2)	36.5 (68.1)	72.6 (95.2)	64.7 (85.7)	78.3 (92.1)
	Δ 20.0	Δ 21.4	Δ 22.8	Δ 19.2	Δ 15.0	Δ 15.7	Δ 31.6	Δ 22.6	Δ 21.0	Δ 13.7
MMTok	67.7 (83.1)	**66.6** (**87.6**)	**75.8** (94.6)	81.4 (97.4)	**82.0** (93.3)	64.9 (81.5)	54.3 (71.5)	76.8 (95.9)	71.2 (88.1)	86.2 (94.6)
	Δ 15.4	Δ 21.0	Δ 18.8	Δ 16.0	Δ 11.3	Δ 16.5	Δ 17.2	Δ 19.2	Δ 16.9	Δ 8.4
ApET	48.2 (72.6)	60.3 (84.5)	71.0 (93.9)	73.9 (96.2)	81.2 (91.7)	56.5 (73.5)	43.9 (64.9)	67.3 (93.3)	62.8 (83.8)	76.0 (90.0)
	Δ 24.4	Δ 24.2	Δ 22.9	Δ 22.4	Δ 10.5	Δ 17.1	Δ 21.0	Δ 26.1	Δ 21.1	Δ 14.0
MiCo	**81.8** (**91.6**)	66.3 (**87.6**)	73.9 (**95.0**)	**82.7** (**97.6**)	**82.0** (**93.6**)	**70.1** (**82.7**)	**69.4** (**76.4**)	**78.8** (96.2)	**75.6** (**90.1**)	**91.6** (**96.7**)
	Δ 9.8	Δ 21.3	Δ 21.0	Δ 14.9	Δ 11.6	Δ 12.6	Δ 6.9	Δ 17.3	Δ 14.4	Δ 5.1
Reference image-area reduction: 93.75%										
Vanilla	93.0 (92.7)	75.0 (88.6)	82.2 (94.6)	89.1 (97.9)	86.5 (93.8)	79.4 (87.3)	87.7 (90.4)	84.2 (97.6)	84.6 (92.9)	100.0 (100.0)
	Δ 0.3	Δ 13.6	Δ 12.4	Δ 8.9	Δ 7.4	Δ 7.9	Δ 2.7	Δ 13.5	Δ 8.2	Δ 0.0
FastV	76.8 (83.6)	59.4 (79.4)	72.9 (94.2)	80.0 (96.5)	79.2 (92.1)	61.5 (78.0)	51.7 (69.1)	67.0 (93.8)	68.6 (85.8)	81.0 (92.4)
	Δ 6.8	Δ 20.0	Δ 21.3	Δ 16.4	Δ 12.9	Δ 16.4	Δ 17.4	Δ 26.7	Δ 17.2	Δ 11.4
SparseVLM	41.0 (74.5)	55.2 (81.9)	74.6 (94.2)	79.5 (96.4)	79.2 (92.7)	60.3 (73.9)	27.0 (51.4)	69.4 (93.0)	60.8 (82.2)	71.8 (88.6)
	Δ 33.5	Δ 26.7	Δ 19.6	Δ 16.9	Δ 13.5	Δ 13.6	Δ 24.4	Δ 23.5	Δ 21.5	Δ 16.7
DivPrune	71.5 (83.6)	69.8 (87.1)	77.9 (94.4)	82.0 (97.1)	81.6 (92.6)	66.5 (81.1)	59.7 (77.1)	**80.4** (**96.8**)	73.7 (88.7)	87.1 (95.5)
	Δ 12.0	Δ 17.3	Δ 16.5	Δ 15.1	Δ 11.0	Δ 14.6	Δ 17.4	Δ 16.4	Δ 15.1	Δ 8.5
VisionZip	56.9 (81.1)	67.6 (86.8)	75.4 (94.3)	82.7 (96.9)	79.2 (92.8)	64.0 (78.0)	41.2 (69.6)	66.2 (95.7)	66.6 (86.9)	78.8 (93.6)
	Δ 24.2	Δ 19.2	Δ 18.9	Δ 14.2	Δ 13.6	Δ 14.0	Δ 28.4	Δ 29.6	Δ 20.3	Δ 14.8
VisPruner	59.9 (79.4)	67.3 (87.1)	74.2 (94.2)	82.0 (96.9)	79.2 (92.6)	64.3 (78.9)	41.2 (67.5)	68.7 (96.1)	67.1 (86.6)	79.3 (93.2)
	Δ 19.5	Δ 19.9	Δ 20.1	Δ 14.9	Δ 13.3	Δ 14.6	Δ 26.3	Δ 27.4	Δ 19.5	Δ 14.0
HoloV	56.2 (79.0)	67.8 (87.0)	76.7 (94.0)	80.8 (97.1)	78.3 (93.0)	65.8 (79.3)	40.8 (70.4)	71.9 (96.4)	67.3 (87.0)	79.5 (93.7)
	Δ 22.8	Δ 19.3	Δ 17.4	Δ 16.3	Δ 14.7	Δ 13.5	Δ 29.6	Δ 24.5	Δ 19.8	Δ 14.2
MMTok	69.9 (85.7)	70.2 (**87.9**)	**80.2** (94.6)	83.7 (97.6)	81.6 (93.7)	66.8 (82.1)	56.4 (73.0)	78.7 (96.5)	73.4 (88.9)	86.8 (95.7)
	Δ 15.8	Δ 17.7	Δ 14.4	Δ 14.0	Δ 12.1	Δ 15.3	Δ 16.6	Δ 17.7	Δ 15.4	Δ 8.9
ApET	51.7 (76.6)	64.7 (84.8)	74.8 (94.1)	78.1 (96.4)	81.6 (92.0)	56.9 (75.0)	45.0 (68.3)	67.3 (94.5)	65.0 (85.2)	76.8 (91.8)
	Δ 24.8	Δ 20.1	Δ 19.3	Δ 18.3	Δ 10.4	Δ 18.1	Δ 23.3	Δ 27.2	Δ 20.2	Δ 14.9
MiCo	**83.6** (**90.6**)	**70.5** (87.8)	78.7 (**95.0**)	**85.0** (**97.8**)	**84.1** (**93.8**)	**72.0** (**82.8**)	**69.0** (**78.4**)	79.6 (**96.8**)	**77.8** (**90.4**)	**91.9** (**97.3**)
	Δ 6.9	Δ 17.3	Δ 16.2	Δ 12.8	Δ 9.7	Δ 10.8	Δ 9.5	Δ 17.2	Δ 12.6	Δ 5.4

Continued on next page

Table 24 (continued): Qwen2.5-VL-7B, $T = 256$, Group A (Group B)

Method	Chart QA	GQA	MMB CN	MMB EN	MME	MMStar	OCR Bench	POPE	Acc	Rel
Reference image-area reduction: 96.00%										
Vanilla	93.0 (92.4)	77.2 (88.4)	83.7 (94.9)	90.6 (98.0)	85.4 (94.0)	81.2 (86.6)	88.8 (89.3)	85.7 (97.6)	85.7 (92.7)	100.0 (100.0)
	Δ 0.6	Δ 11.2	Δ 11.2	Δ 7.4	Δ 8.6	Δ 5.4	Δ 0.6	Δ 11.8	Δ 7.0	Δ 0.0
FastV	77.0 (82.5)	60.9 (79.9)	77.3 (94.4)	82.6 (96.7)	79.9 (92.1)	65.9 (76.0)	53.6 (70.2)	66.7 (94.3)	70.5 (85.8)	82.3 (92.5)
	Δ 5.5	Δ 19.0	Δ 17.1	Δ 14.1	Δ 12.2	Δ 10.1	Δ 16.6	Δ 27.6	Δ 15.3	Δ 10.3
SparseVLM	41.4 (74.9)	58.5 (82.0)	78.8 (94.3)	83.4 (96.4)	81.3 (92.6)	62.7 (73.2)	28.9 (54.4)	69.5 (93.3)	63.1 (82.6)	73.6 (89.2)
	Δ 33.5	Δ 23.5	Δ 15.5	Δ 13.0	Δ 11.3	Δ 10.6	Δ 25.5	Δ 23.8	Δ 19.6	Δ 15.6
DivPrune	72.0 (81.7)	72.0 (87.1)	79.8 (94.9)	84.0 (97.4)	81.3 (92.7)	68.4 (80.9)	61.8 (78.0)	79.7 (**97.2**)	74.9 (88.7)	87.4 (95.8)
	Δ 9.8	Δ 15.1	Δ 15.1	Δ 13.4	Δ 11.5	Δ 12.5	Δ 16.1	Δ 17.5	Δ 13.9	Δ 8.4
VisionZip	57.2 (81.0)	70.3 (86.7)	77.6 (94.8)	84.7 (97.2)	80.4 (92.7)	65.1 (78.3)	45.2 (70.2)	68.0 (95.8)	68.6 (87.1)	80.0 (94.0)
	Δ 23.8	Δ 16.4	Δ 17.2	Δ 12.5	Δ 12.4	Δ 13.2	Δ 25.0	Δ 27.8	Δ 18.5	Δ 14.0
VisPruner	60.4 (77.9)	69.9 (87.1)	77.0 (94.7)	83.5 (97.3)	78.5 (92.7)	65.4 (79.4)	44.6 (68.6)	69.7 (96.4)	68.6 (86.8)	80.1 (93.6)
	Δ 17.6	Δ 17.2	Δ 17.8	Δ 13.8	Δ 14.2	Δ 14.0	Δ 24.0	Δ 26.7	Δ 18.2	Δ 13.6
HoloV	56.6 (78.3)	70.7 (86.8)	79.0 (94.5)	82.8 (97.5)	79.0 (93.0)	67.0 (79.6)	44.0 (72.5)	72.6 (96.6)	69.0 (87.3)	80.5 (94.3)
	Δ 21.7	Δ 16.2	Δ 15.5	Δ 14.7	Δ 14.0	Δ 12.6	Δ 28.5	Δ 24.0	Δ 18.4	Δ 13.8
MMTok	70.6 (82.1)	**73.3** (87.6)	**81.8** (95.0)	86.2 (97.8)	81.7 (93.8)	69.5 (81.3)	57.8 (74.8)	79.5 (96.6)	75.1 (88.6)	87.6 (95.6)
	Δ 11.5	Δ 14.3	Δ 13.2	Δ 11.6	Δ 12.1	Δ 11.8	Δ 16.9	Δ 17.1	Δ 13.6	Δ 8.1
ApET	52.2 (75.7)	66.2 (85.2)	78.4 (94.4)	80.7 (96.8)	80.8 (92.2)	56.7 (76.9)	47.8 (69.6)	68.0 (94.8)	66.3 (85.7)	77.4 (92.5)
	Δ 23.5	Δ 19.0	Δ 16.0	Δ 16.1	Δ 11.4	Δ 20.2	Δ 21.8	Δ 26.8	Δ 19.3	Δ 15.0
MiCo	**84.0** (**88.6**)	72.7 (**87.8**)	81.0 (**95.3**)	**87.5** (**97.9**)	**83.1** (**94.0**)	**74.7** (**81.7**)	**70.3** (**78.6**)	**79.9** (97.0)	**79.2** (**90.1**)	**92.4** (**97.3**)
	Δ 4.6	Δ 15.1	Δ 14.3	Δ 10.4	Δ 10.9	Δ 7.1	Δ 8.4	Δ 17.1	Δ 11.0	Δ 4.9
Reference image-area reduction: 99.00%										
Vanilla	92.8 (93.6)	80.9 (88.0)	87.3 (95.1)	93.3 (98.0)	89.4 (93.9)	81.3 (87.4)	87.4 (97.6)	92.3 (96.8)	88.1 (93.8)	100.0 (100.0)
	Δ 0.8	Δ 7.1	Δ 7.8	Δ 4.7	Δ 4.6	Δ 6.1	Δ 10.1	Δ 4.5	Δ 5.7	Δ 0.0
FastV	76.7 (84.6)	67.1 (79.2)	81.7 (95.0)	86.9 (97.1)	83.8 (92.4)	67.0 (76.5)	55.1 (87.0)	81.2 (92.7)	74.9 (88.1)	85.1 (93.9)
	Δ 7.9	Δ 12.1	Δ 13.3	Δ 10.2	Δ 8.6	Δ 9.6	Δ 31.9	Δ 11.5	Δ 13.1	Δ 8.8
SparseVLM	41.2 (75.6)	64.9 (81.9)	83.2 (94.8)	87.1 (96.8)	84.3 (92.9)	61.9 (75.8)	32.0 (75.6)	80.7 (92.3)	66.9 (85.7)	76.0 (91.4)
	Δ 34.3	Δ 17.0	Δ 11.6	Δ 9.7	Δ 8.6	Δ 13.9	Δ 43.6	Δ 11.6	Δ 18.8	Δ 15.4
DivPrune	71.9 (82.0)	77.0 (86.5)	84.2 (95.3)	88.4 (97.6)	85.6 (92.8)	70.4 (80.7)	63.6 (92.7)	**89.1** (96.1)	78.8 (90.5)	89.4 (96.4)
	Δ 10.0	Δ 9.5	Δ 11.1	Δ 9.2	Δ 7.2	Δ 10.3	Δ 29.1	Δ 7.0	Δ 11.7	Δ 7.0
VisionZip	57.8 (76.7)	76.0 (85.9)	82.4 (95.4)	88.7 (97.4)	85.1 (92.8)	65.6 (80.0)	49.3 (85.4)	81.6 (94.6)	73.3 (88.5)	83.2 (94.4)
	Δ 18.9	Δ 9.9	Δ 13.0	Δ 8.8	Δ 7.7	Δ 14.4	Δ 36.1	Δ 13.0	Δ 15.2	Δ 11.2
VisPruner	60.2 (78.6)	75.6 (86.4)	82.0 (95.2)	87.8 (97.6)	84.8 (92.6)	65.8 (81.2)	48.2 (84.6)	83.1 (95.0)	73.5 (88.9)	83.4 (94.8)
	Δ 18.3	Δ 10.8	Δ 13.2	Δ 9.8	Δ 7.8	Δ 15.4	Δ 36.3	Δ 11.8	Δ 15.4	Δ 11.4
HoloV	56.9 (76.3)	76.1 (86.1)	83.0 (95.1)	88.1 (97.5)	85.4 (92.9)	68.6 (80.0)	48.4 (91.1)	84.4 (95.5)	73.8 (89.3)	83.8 (95.2)
	Δ 19.5	Δ 10.0	Δ 12.1	Δ 9.4	Δ 7.5	Δ 11.4	Δ 42.7	Δ 11.1	Δ 15.5	Δ 11.4
MMTok	70.4 (83.1)	**77.6** (87.3)	85.2 (95.5)	90.3 (97.8)	85.6 (**94.0**)	70.4 (82.2)	60.2 (87.0)	87.3 (96.0)	78.4 (90.4)	89.0 (96.3)
	Δ 12.6	Δ 9.7	Δ 10.3	Δ 7.6	Δ 8.4	Δ 11.8	Δ 26.8	Δ 8.6	Δ 12.0	Δ 7.3
ApET	52.3 (74.8)	72.4 (84.5)	81.9 (95.2)	85.3 (97.3)	84.8 (92.4)	59.8 (76.8)	50.7 (86.2)	81.7 (93.4)	71.1 (87.6)	80.7 (93.4)
	Δ 22.5	Δ 12.1	Δ 13.3	Δ 12.0	Δ 7.5	Δ 17.0	Δ 35.4	Δ 11.7	Δ 16.4	Δ 12.6
MiCo	**83.9** (**89.5**)	77.1 (**87.6**)	**85.5** (**95.6**)	**91.2** (**97.9**)	**87.5** (**94.0**)	**74.1** (**83.5**)	**69.9** (**93.5**)	88.1 (**96.3**)	**82.2** (**92.2**)	**93.3** (**98.3**)
	Δ 5.6	Δ 10.5	Δ 10.1	Δ 6.7	Δ 6.5	Δ 9.3	Δ 23.6	Δ 8.1	Δ 10.1	Δ 5.0

Table 25: VTC-Bench scores (%) on Qwen2.5-VL-7B at $T = 128$. Cells show Group A (Group B), with the absolute gap Δ in percentage points below.

Method	Chart QA	GQA	MMB CN	MMB EN	MME	MMStar	OCR Bench	POPE	Acc	Rel
Reference image-area reduction: 75.00%										
Vanilla	88.5 (94.3)	65.0 (88.2)	67.4 (94.7)	76.5 (97.9)	70.4 (93.9)	75.4 (86.7)	83.8 (90.4)	81.0 (96.7)	76.0 (92.8)	100.0 (100.0)
	Δ 5.8	Δ 23.2	Δ 27.3	Δ 21.4	Δ 23.5	Δ 11.3	Δ 6.6	Δ 15.6	Δ 16.8	Δ 0.0
FastV	17.7 (40.1)	37.3 (60.8)	43.5 (82.3)	49.3 (82.5)	50.7 (77.8)	34.4 (55.9)	15.1 (26.9)	49.6 (69.8)	37.2 (62.0)	48.9 (66.8)
	Δ 22.4	Δ 23.5	Δ 38.8	Δ 33.2	Δ 27.1	Δ 21.5	Δ 11.8	Δ 20.2	Δ 24.8	Δ 17.8
SparseVLM	10.0 (24.5)	40.8 (68.5)	50.9 (89.5)	58.5 (89.7)	57.7 (84.7)	39.3 (61.4)	11.7 (23.2)	54.0 (80.3)	40.4 (65.2)	53.1 (70.2)
	Δ 14.4	Δ 27.6	Δ 38.6	Δ 31.2	Δ 26.9	Δ 22.0	Δ 11.5	Δ 26.4	Δ 24.8	Δ 17.1
DivPrune	35.4 (62.2)	54.1 (82.9)	**59.1** (92.8)	63.6 (94.7)	**67.6** (90.1)	52.5 (74.0)	31.3 (58.3)	66.7 (94.2)	53.8 (81.2)	70.7 (87.4)
	Δ 26.8	Δ 28.9	Δ 33.7	Δ 31.1	Δ 22.5	Δ 21.6	Δ 27.0	Δ 27.6	Δ 27.4	Δ 16.7
VisionZip	24.3 (45.4)	50.9 (81.3)	53.9 (92.5)	62.7 (94.8)	62.0 (90.2)	46.4 (70.7)	17.3 (49.4)	55.6 (91.1)	46.6 (76.9)	61.4 (82.8)
	Δ 21.0	Δ 30.4	Δ 38.6	Δ 32.1	Δ 28.2	Δ 24.2	Δ 32.0	Δ 35.5	Δ 30.3	Δ 21.5
VisPruner	23.2 (44.5)	51.0 (81.6)	54.8 (92.1)	**65.9** (94.3)	63.4 (89.8)	46.4 (68.5)	22.9 (48.2)	53.5 (90.3)	47.7 (76.2)	62.7 (82.0)
	Δ 21.2	Δ 30.6	Δ 37.3	Δ 28.4	Δ 26.5	Δ 22.1	Δ 25.3	Δ 36.7	Δ 28.5	Δ 19.3
HoloV	24.1 (46.1)	52.8 (81.7)	56.1 (92.7)	63.6 (94.4)	60.6 (90.3)	48.6 (69.8)	19.0 (48.2)	54.6 (90.7)	47.4 (76.7)	62.4 (82.6)
	Δ 22.0	Δ 28.9	Δ 36.6	Δ 30.8	Δ 29.8	Δ 21.1	Δ 29.3	Δ 36.1	Δ 29.3	Δ 20.3

Continued on next page

Table 25 (continued): Qwen2.5-VL-7B, $T = 128$, Group A (Group B)

Method	Chart QA	GQA	MMB CN	MMB EN	MME	MMStar	OCR Bench	POPE	Acc	Rel
MMTok	33.5 (55.9)	**58.1** (**85.3**)	55.2 (93.2)	65.0 (95.7)	66.2 (91.9)	54.1 (74.4)	33.5 (54.1)	**72.5** (**94.3**)	54.8 (80.6)	72.0 (86.8)
	Δ 22.4	Δ 27.2	Δ 38.0	Δ 30.8	Δ 25.7	Δ 20.3	Δ 20.6	Δ 21.8	Δ 25.8	Δ 14.8
ApET	17.1 (36.7)	47.1 (77.8)	51.7 (90.7)	56.7 (92.1)	63.4 (88.5)	42.1 (62.0)	14.5 (42.5)	51.0 (87.5)	43.0 (72.2)	56.5 (77.8)
	Δ 19.6	Δ 30.7	Δ 39.0	Δ 35.4	Δ 25.1	Δ 19.9	Δ 28.0	Δ 36.4	Δ 29.3	Δ 21.3
MiCo	**48.4** (**72.9**)	57.4 (84.3)	56.1 (**93.5**)	**65.9** (**96.3**)	63.4 (**92.5**)	**60.1** (**77.3**)	**39.1** (**63.1**)	70.6 (93.7)	**57.6** (**84.2**)	**75.8** (**90.7**)
	Δ 24.5	Δ 26.9	Δ 37.4	Δ 30.4	Δ 29.1	Δ 17.1	Δ 24.0	Δ 23.1	Δ 26.6	Δ 14.9
Reference image-area reduction: 88.89%										
Vanilla	91.9 (95.6)	71.5 (88.5)	77.2 (94.8)	86.7 (97.9)	85.7 (93.6)	77.9 (87.3)	87.2 (90.2)	82.6 (97.2)	82.6 (93.1)	100.0 (100.0)
	Δ 3.7	Δ 17.0	Δ 17.6	Δ 11.2	Δ 7.9	Δ 9.4	Δ 3.0	Δ 14.6	Δ 10.6	Δ 0.0
FastV	26.4 (56.1)	38.4 (62.5)	49.5 (83.4)	57.0 (83.4)	65.4 (77.7)	36.2 (58.4)	17.8 (28.9)	40.3 (71.8)	41.4 (65.3)	50.1 (70.1)
	Δ 29.7	Δ 24.1	Δ 33.9	Δ 26.4	Δ 12.3	Δ 22.2	Δ 11.1	Δ 31.5	Δ 23.9	Δ 20.0
SparseVLM	12.3 (43.1)	42.2 (70.5)	54.8 (90.8)	61.8 (91.0)	67.7 (84.9)	41.7 (63.7)	11.0 (27.7)	48.9 (82.2)	42.5 (69.2)	51.5 (74.3)
	Δ 30.8	Δ 28.3	Δ 36.0	Δ 29.2	Δ 17.2	Δ 22.0	Δ 16.7	Δ 33.3	Δ 26.7	Δ 22.8
DivPrune	48.6 (74.5)	59.1 (84.1)	**66.9** (93.5)	75.1 (95.1)	73.7 (90.4)	56.1 (75.7)	36.8 (63.4)	73.1 (94.8)	61.2 (83.9)	74.1 (90.1)
	Δ 26.0	Δ 25.1	Δ 26.5	Δ 19.9	Δ 16.7	Δ 19.6	Δ 26.6	Δ 21.7	Δ 22.8	Δ 16.0
VisionZip	32.5 (60.5)	55.8 (82.7)	64.2 (93.1)	74.1 (95.2)	76.7 (90.1)	50.9 (72.3)	21.1 (57.4)	55.5 (92.9)	53.8 (80.5)	65.2 (86.4)
	Δ 28.1	Δ 26.9	Δ 28.8	Δ 21.1	Δ 13.4	Δ 21.4	Δ 36.4	Δ 37.4	Δ 26.7	Δ 21.2
VisPruner	31.2 (60.3)	56.4 (82.8)	63.4 (92.8)	73.9 (94.9)	78.9 (89.6)	48.0 (71.3)	23.7 (56.2)	53.5 (92.0)	53.6 (80.0)	64.9 (85.9)
	Δ 29.1	Δ 26.5	Δ 29.3	Δ 21.0	Δ 10.7	Δ 23.3	Δ 32.4	Δ 38.5	Δ 26.4	Δ 20.9
HoloV	33.1 (60.5)	57.6 (82.9)	65.3 (93.3)	70.6 (95.2)	75.9 (90.2)	50.9 (72.0)	22.0 (56.0)	54.3 (92.5)	53.7 (80.3)	65.0 (86.2)
	Δ 27.4	Δ 25.3	Δ 28.0	Δ 24.6	Δ 14.3	Δ 21.0	Δ 34.0	Δ 38.2	Δ 26.6	Δ 21.2
MMTok	43.3 (69.2)	61.7 (**86.7**)	65.6 (93.7)	74.6 (96.3)	**82.0** (91.6)	57.6 (75.8)	31.8 (62.3)	**74.6** (**95.1**)	61.4 (83.8)	74.3 (90.0)
	Δ 25.9	Δ 24.9	Δ 28.1	Δ 21.6	Δ 9.7	Δ 18.3	Δ 30.6	Δ 20.5	Δ 22.5	Δ 15.7
ApET	23.9 (52.6)	50.2 (79.6)	57.8 (91.8)	61.8 (93.5)	75.2 (88.5)	38.7 (66.7)	17.8 (49.6)	50.0 (89.3)	46.9 (76.5)	56.8 (82.1)
	Δ 28.7	Δ 29.5	Δ 34.0	Δ 31.7	Δ 13.3	Δ 27.9	Δ 31.8	Δ 39.3	Δ 29.5	Δ 25.3
MiCo	**62.4** (**79.5**)	**62.1** (85.4)	65.9 (**94.0**)	**76.6** (**96.7**)	77.4 (**92.5**)	**61.3** (**79.4**)	**43.9** (**67.7**)	72.9 (94.6)	**65.3** (**86.2**)	**79.1** (**92.6**)
	Δ 17.1	Δ 23.3	Δ 28.2	Δ 20.0	Δ 15.0	Δ 18.1	Δ 23.7	Δ 21.7	Δ 20.9	Δ 13.5
Reference image-area reduction: 93.75%										
Vanilla	93.0 (92.7)	75.0 (88.6)	82.2 (94.6)	89.1 (97.9)	86.5 (93.8)	79.4 (87.3)	87.7 (90.4)	84.2 (97.6)	84.6 (92.9)	100.0 (100.0)
	Δ 0.3	Δ 13.6	Δ 12.4	Δ 8.9	Δ 7.4	Δ 7.9	Δ 2.7	Δ 13.5	Δ 8.2	Δ 0.0
FastV	29.1 (70.6)	38.8 (64.2)	51.4 (84.4)	58.2 (84.3)	64.3 (78.3)	36.0 (60.8)	18.7 (30.4)	34.4 (74.3)	41.4 (68.4)	48.9 (73.6)
	Δ 41.5	Δ 25.3	Δ 32.9	Δ 26.1	Δ 14.0	Δ 24.8	Δ 11.7	Δ 39.9	Δ 27.0	Δ 24.8
SparseVLM	13.6 (67.1)	43.4 (72.1)	58.3 (91.7)	64.5 (91.7)	69.1 (85.4)	42.8 (65.3)	12.8 (29.4)	44.9 (84.4)	43.7 (73.4)	51.6 (79.0)
	Δ 53.5	Δ 28.7	Δ 33.4	Δ 27.3	Δ 16.3	Δ 22.5	Δ 16.6	Δ 39.5	Δ 29.7	Δ 27.4
DivPrune	52.0 (80.4)	63.7 (84.4)	72.3 (93.6)	76.4 (95.6)	74.9 (90.9)	59.7 (75.4)	40.5 (65.2)	72.1 (**96.0**)	64.0 (85.2)	75.6 (91.7)
	Δ 28.4	Δ 20.7	Δ 21.3	Δ 19.2	Δ 16.0	Δ 15.7	Δ 24.7	Δ 23.9	Δ 21.2	Δ 16.2
VisionZip	35.3 (72.7)	59.5 (83.4)	69.0 (93.4)	77.2 (95.5)	75.4 (90.8)	51.7 (74.1)	24.2 (62.1)	58.5 (94.1)	56.3 (83.3)	66.6 (89.6)
	Δ 37.5	Δ 23.9	Δ 24.4	Δ 18.3	Δ 15.4	Δ 22.4	Δ 37.9	Δ 35.6	Δ 26.9	Δ 23.1
VisPruner	34.5 (70.6)	59.7 (83.7)	69.2 (92.9)	76.0 (95.4)	77.3 (90.2)	49.5 (72.7)	26.8 (60.0)	54.6 (93.6)	56.0 (82.4)	66.1 (88.7)
	Δ 36.1	Δ 23.9	Δ 23.7	Δ 19.4	Δ 12.9	Δ 23.2	Δ 33.2	Δ 39.1	Δ 26.4	Δ 22.6
HoloV	35.9 (72.0)	60.4 (83.8)	69.6 (93.7)	74.1 (95.6)	76.8 (90.7)	52.3 (73.3)	24.9 (60.3)	58.4 (93.5)	56.6 (82.9)	66.8 (89.2)
	Δ 36.1	Δ 23.5	Δ 24.0	Δ 21.5	Δ 13.9	Δ 21.0	Δ 35.4	Δ 35.1	Δ 26.3	Δ 22.4
MMTok	46.5 (76.6)	**66.4** (**86.9**)	71.9 (93.7)	77.9 (96.6)	81.2 (92.1)	60.6 (75.8)	35.3 (65.2)	**74.7** (**96.0**)	64.3 (85.4)	76.0 (91.9)
	Δ 30.0	Δ 20.5	Δ 21.8	Δ 18.6	Δ 10.9	Δ 15.2	Δ 29.9	Δ 21.3	Δ 21.0	Δ 15.9
ApET	26.5 (67.1)	53.2 (80.8)	62.4 (92.4)	65.8 (94.0)	75.8 (88.9)	41.8 (67.6)	20.4 (53.8)	49.1 (91.3)	49.4 (79.5)	58.4 (85.6)
	Δ 40.7	Δ 27.5	Δ 30.0	Δ 28.2	Δ 13.1	Δ 25.8	Δ 33.4	Δ 42.2	Δ 30.1	Δ 27.2
MiCo	**65.0** (**81.5**)	66.1 (85.7)	**72.9** (**93.9**)	**80.4** (**96.8**)	**82.6** (**92.5**)	**64.0** (**79.5**)	**48.1** (**68.3**)	72.9 (95.6)	**69.0** (**86.7**)	**81.5** (**93.4**)
	Δ 16.5	Δ 19.7	Δ 21.0	Δ 16.4	Δ 9.9	Δ 15.5	Δ 20.2	Δ 22.7	Δ 17.7	Δ 11.8
Reference image-area reduction: 96.00%										
Vanilla	93.0 (92.4)	77.2 (88.4)	83.7 (94.9)	90.6 (98.0)	85.4 (94.0)	81.2 (86.6)	88.8 (89.3)	85.7 (97.6)	85.7 (92.7)	100.0 (100.0)
	Δ 0.6	Δ 11.2	Δ 11.2	Δ 7.4	Δ 8.6	Δ 5.4	Δ 0.6	Δ 11.8	Δ 7.0	Δ 0.0
FastV	29.3 (73.0)	39.7 (65.2)	56.4 (85.0)	61.1 (85.0)	60.3 (78.8)	36.2 (62.8)	20.1 (31.1)	31.9 (75.3)	41.9 (69.5)	48.9 (75.0)
	Δ 43.7	Δ 25.5	Δ 28.6	Δ 23.9	Δ 18.6	Δ 26.6	Δ 11.0	Δ 43.4	Δ 27.7	Δ 26.2
SparseVLM	13.8 (70.7)	43.8 (73.5)	65.4 (91.9)	69.2 (92.1)	66.7 (85.8)	44.7 (65.8)	13.7 (32.0)	44.4 (85.1)	45.2 (74.6)	52.7 (80.5)
	Δ 56.9	Δ 29.8	Δ 26.5	Δ 22.8	Δ 19.1	Δ 21.1	Δ 18.4	Δ 40.7	Δ 29.4	Δ 27.8
DivPrune	52.9 (76.8)	65.9 (84.6)	75.4 (94.1)	78.0 (96.2)	75.8 (90.9)	61.3 (75.6)	43.2 (67.0)	72.5 (**96.3**)	65.6 (85.2)	76.6 (91.9)
	Δ 23.9	Δ 18.7	Δ 18.7	Δ 18.2	Δ 15.1	Δ 14.3	Δ 23.8	Δ 23.8	Δ 19.6	Δ 15.4
VisionZip	35.8 (72.6)	63.3 (83.1)	73.5 (93.7)	79.2 (96.0)	78.1 (90.5)	51.5 (76.2)	29.3 (63.1)	60.2 (94.3)	58.9 (83.7)	68.7 (90.3)
	Δ 36.9	Δ 19.9	Δ 20.2	Δ 16.8	Δ 12.5	Δ 24.7	Δ 33.8	Δ 34.1	Δ 24.8	Δ 21.7
VisPruner	35.0 (70.7)	63.0 (83.6)	72.4 (93.5)	78.3 (95.8)	79.0 (90.1)	49.9 (74.5)	30.9 (61.5)	55.5 (94.0)	58.0 (83.0)	67.7 (89.6)
	Δ 35.8	Δ 20.6	Δ 21.1	Δ 17.5	Δ 11.1	Δ 24.7	Δ 30.6	Δ 38.5	Δ 25.0	Δ 21.9
HoloV	36.5 (71.1)	63.4 (83.9)	73.5 (94.1)	76.7 (96.1)	79.5 (90.5)	52.3 (75.2)	29.5 (61.5)	57.5 (94.2)	58.6 (83.3)	68.4 (89.9)
	Δ 34.6	Δ 20.5	Δ 20.6	Δ 19.5	Δ 11.0	Δ 22.8	Δ 32.0	Δ 36.7	Δ 24.7	Δ 21.5
MMTok	47.2 (74.9)	**69.1** (**86.9**)	75.5 (94.1)	80.7 (**96.9**)	80.4 (92.3)	62.1 (76.0)	37.8 (68.6)	**77.0** (95.9)	66.2 (85.7)	77.3 (92.5)
	Δ 27.7	Δ 17.7	Δ 18.6	Δ 16.2	Δ 11.9	Δ 13.9	Δ 30.9	Δ 18.9	Δ 19.5	Δ 15.2

Continued on next page

Table 25 (continued): Qwen2.5-VL-7B, $T = 128$, Group A (Group B)

Method	Chart QA	GQA	MMB CN	MMB EN	MME	MMStar	OCR Bench	POPE	Acc	Rel
ApET	26.6 (69.6)	56.6 (80.8)	68.5 (92.6)	69.7 (94.6)	75.8 (89.0)	42.2 (69.6)	23.7 (56.6)	50.7 (91.7)	51.7 (80.6)	60.4 (87.0)
	Δ 42.9	Δ 24.2	Δ 24.2	Δ 24.9	Δ 13.2	Δ 27.4	Δ 32.9	Δ 41.0	Δ 28.8	Δ 26.6
MiCo	**65.4** (**79.8**)	68.0 (86.0)	**76.3** (**94.3**)	**83.7** (**96.9**)	**81.7** (**92.7**)	**65.4** (**79.8**)	**50.4** (**69.6**)	73.5 (95.8)	**70.6** (**86.9**)	**82.3** (**93.7**)
	Δ 14.4	Δ 18.0	Δ 18.0	Δ 13.3	Δ 10.9	Δ 14.4	Δ 19.2	Δ 22.3	Δ 16.3	Δ 11.4
Reference image-area reduction: 99.00%										
Vanilla	92.8 (93.6)	80.9 (88.0)	87.3 (95.1)	93.3 (98.0)	89.4 (93.9)	81.3 (87.4)	87.4 (97.6)	92.3 (96.8)	88.1 (93.8)	100.0 (100.0)
	Δ 0.8	Δ 7.1	Δ 7.8	Δ 4.7	Δ 4.6	Δ 6.1	Δ 10.1	Δ 4.5	Δ 5.7	Δ 0.0
FastV	29.4 (71.8)	42.1 (68.3)	65.9 (85.3)	67.3 (86.0)	64.4 (79.6)	37.4 (65.9)	21.2 (41.5)	50.0 (74.2)	47.2 (71.6)	53.6 (76.3)
	Δ 42.4	Δ 26.2	Δ 19.4	Δ 18.6	Δ 15.3	Δ 28.5	Δ 20.3	Δ 24.2	Δ 24.4	Δ 22.7
SparseVLM	13.8 (70.3)	50.1 (74.7)	72.0 (93.0)	74.7 (93.2)	71.8 (86.4)	44.8 (69.1)	16.1 (46.3)	61.9 (83.9)	50.7 (77.1)	57.5 (82.2)
	Δ 56.5	Δ 24.5	Δ 21.0	Δ 18.5	Δ 14.6	Δ 24.3	Δ 30.3	Δ 21.9	Δ 26.5	Δ 24.7
DivPrune	52.6 (78.6)	72.1 (83.9)	79.8 (**95.0**)	83.8 (96.6)	81.4 (91.1)	61.9 (77.3)	46.6 (83.7)	84.9 (94.9)	70.4 (87.6)	79.9 (93.4)
	Δ 25.9	Δ 11.8	Δ 15.1	Δ 12.9	Δ 9.7	Δ 15.4	Δ 37.1	Δ 10.0	Δ 17.2	Δ 13.5
VisionZip	36.3 (68.8)	70.2 (82.1)	77.9 (94.8)	84.4 (96.3)	82.7 (90.7)	54.0 (77.5)	35.2 (81.3)	76.6 (92.8)	64.7 (85.5)	73.4 (91.2)
	Δ 32.5	Δ 12.0	Δ 16.9	Δ 11.9	Δ 8.0	Δ 23.5	Δ 46.1	Δ 16.2	Δ 20.9	Δ 17.8
VisPruner	35.5 (66.9)	70.1 (82.6)	77.3 (94.6)	83.8 (96.2)	82.2 (90.4)	52.4 (75.8)	35.4 (82.9)	73.1 (92.6)	63.7 (85.3)	72.3 (90.9)
	Δ 31.4	Δ 12.5	Δ 17.3	Δ 12.5	Δ 8.3	Δ 23.4	Δ 47.5	Δ 19.6	Δ 21.6	Δ 18.6
HoloV	36.7 (69.9)	70.5 (82.8)	78.8 (94.9)	83.4 (96.3)	83.5 (90.6)	54.3 (76.8)	34.5 (82.1)	75.1 (92.6)	64.6 (85.8)	73.3 (91.4)
	Δ 33.3	Δ 12.3	Δ 16.1	Δ 12.9	Δ 7.1	Δ 22.5	Δ 47.6	Δ 17.5	Δ 21.2	Δ 18.1
MMTok	47.3 (74.1)	**74.4** (**86.6**)	80.0 (**95.0**)	85.3 (**97.4**)	84.3 (92.5)	62.8 (77.5)	43.1 (85.4)	**85.5** (**95.3**)	70.3 (88.0)	79.8 (93.8)
	Δ 26.8	Δ 12.2	Δ 14.9	Δ 12.2	Δ 8.2	Δ 14.7	Δ 42.2	Δ 9.8	Δ 17.6	Δ 13.9
ApET	27.1 (66.2)	64.9 (79.7)	74.0 (93.9)	76.7 (95.5)	79.0 (89.5)	44.8 (71.4)	29.1 (76.4)	70.8 (89.7)	58.3 (82.8)	66.2 (88.2)
	Δ 39.1	Δ 14.8	Δ 19.8	Δ 18.8	Δ 10.5	Δ 26.6	Δ 47.3	Δ 18.9	Δ 24.5	Δ 22.1
MiCo	**65.2** (**81.2**)	74.0 (85.3)	**81.4** (94.8)	**88.2** (97.0)	**85.4** (**92.9**)	**67.0** (**80.5**)	**52.6** (**86.2**)	84.5 (94.7)	**74.8** (**89.1**)	**84.9** (**95.0**)
	Δ 16.0	Δ 11.3	Δ 13.4	Δ 8.8	Δ 7.5	Δ 13.5	Δ 33.5	Δ 10.2	Δ 14.3	Δ 10.1

H.2 Generalized Referring Expression Comprehension

Table 26 extends the evaluation to gRefCOCO (He et al., 2023), which includes expressions referring to single, multiple, or no target objects. We report $\mathrm{Pr@}(F_1 = 1, \mathrm{GIoU} \geq 0.5)$ on val, testA, and testB, computed with the released GREC code using a generalized-IoU matching threshold (a variant of the original $\mathrm{IoU} \geq 0.5$ definition, see Appendix A.2). MiCo leads the displayed pruners in all six model–budget–split settings on Qwen2.5 and all four testA/testB settings on Qwen3-VL. At $T = 128$ on Qwen3-VL, it reaches 36.3/40.0 on testA/testB, versus VisPruner's 31.8/34.7; FastV remains strongest on Qwen3-VL val at both budgets. Parenthesized values report retention relative to the matching Vanilla score.

Table 26: GREC scores (%) on gRefCOCO.

	Qwen2.5-VL-7B			Qwen3-VL-8B		
Method	val	testA	testB	val	testA	testB
Vanilla	66.8 (100.0%)	47.3 (100.0%)	51.9 (100.0%)	49.1 (100.0%)	47.6 (100.0%)	47.8 (100.0%)
Retain 256 tokens (80.2% pruned)						
FastV (ECCV'24)	61.9 (92.8%)	27.4 (57.9%)	34.6 (66.7%)	**55.3** (112.7%)	38.0 (79.9%)	39.6 (82.8%)
VisPruner (ICCV'25)	64.0 (95.8%)	35.8 (75.8%)	40.8 (78.5%)	52.0 (106.0%)	41.7 (87.5%)	42.9 (89.6%)
ApET (CVPR'26)	62.3 (93.3%)	27.6 (58.3%)	33.0 (63.6%)	50.0 (101.9%)	31.4 (66.0%)	35.1 (73.4%)
MiCo	**64.9** (97.2%)	**38.8** (82.1%)	**44.1** (84.8%)	52.7 (107.4%)	**43.8** (92.0%)	**45.1** (94.3%)
Retain 128 tokens (90.1% pruned)						
FastV (ECCV'24)	62.4 (93.5%)	23.2 (49.0%)	29.1 (56.0%)	**60.1** (122.4%)	26.5 (55.6%)	30.3 (63.3%)
VisPruner (ICCV'25)	62.2 (93.2%)	29.3 (62.1%)	33.2 (63.9%)	54.3 (110.5%)	31.8 (66.8%)	34.7 (72.6%)
ApET (CVPR'26)	61.8 (92.5%)	24.5 (51.9%)	29.6 (56.9%)	52.4 (106.8%)	26.4 (55.4%)	30.1 (62.8%)
MiCo	**62.5** (93.6%)	**31.0** (65.6%)	**35.2** (67.8%)	53.7 (109.4%)	**36.3** (76.2%)	**40.0** (83.6%)

H.3 Scene Text, Chart, and Document Understanding

Table 27 gives the complete TextVQA, ChartQA, and DocVQA comparison; the main text reports a compact view with FastV, VisPruner, and ApET.

Table 27: Text (TextVQA), Chart (ChartQA), and Doc (DocVQA) scores (%) on Qwen2.5-VL-7B and Qwen3-VL-8B at $T = 256/128$.

	Qwen2.5-VL-7B						Qwen3-VL-8B					
	$T = 256$			$T = 128$			$T = 256$			$T = 128$		
Method	Text	Chart	Doc	Text	Chart	Doc	Text	Chart	Doc	Text	Chart	Doc
Vanilla	85.06	87.28	94.86	85.06	87.28	94.86	84.28	83.04	95.72	84.28	83.04	95.72
FastV	78.44	68.28	38.22	46.45	33.20	14.54	62.18	28.64	14.79	32.88	16.24	9.83
SparseVLM	64.23	40.08	12.50	39.48	19.24	9.50	66.72	11.08	17.05	41.29	20.56	11.17
DivPrune	76.55	70.44	51.70	66.69	52.16	35.56	74.61	59.32	47.91	64.72	43.76	34.14
VisionZip	61.87	55.56	27.44	49.50	36.64	20.01	74.68	69.76	49.68	57.88	47.00	25.10
VisPruner	64.02	56.08	27.66	49.40	36.68	19.69	71.96	63.52	39.63	54.58	39.64	23.00
HoloV	63.19	54.68	28.19	49.05	36.48	19.19	74.63	**71.40**	50.40	57.27	**51.20**	27.73
MMTok	75.29	67.68	39.31	61.23	45.08	22.25	76.93	70.84	**66.95**	68.11	48.68	45.77
ApET	69.17	50.60	23.12	45.82	31.16	14.35	59.54	54.52	39.25	41.27	35.44	25.59
MiCo	**80.96**	**81.16**	**57.44**	**71.74**	**70.68**	**38.58**	**79.69**	66.80	64.98	**73.44**	50.12	**46.35**

Metrics. TextVQA_VAL: VQA consensus score over 5,000 questions; ChartQA_TEST: relaxed accuracy over 2,500 questions; DocVQA_VAL: ANLS over 5,349 questions. Each model's unpruned reference is repeated under both budgets.

H.4 OCRBench

Table 28 reports the complete OCRBench comparison on four image models; this benchmark is not tabulated in the main text.

Table 28: OCRBench scores (0–100, higher is better) on LLaVA-NeXT-7B/13B, Qwen2.5-VL-7B, and InternVL3-8B at token budget T.

	LLaVA-NeXT-7B			LLaVA-NeXT-13B			Qwen2.5-VL-7B		InternVL3-8B	
Method / T	640	320	160	640	320	160	256	128	256	128
Vanilla		50.7			51.1			87.7		88.1
FastV	37.0	11.9	–	38.9	18.9	–	70.6	39.6	55.0	25.0
DivPrune	38.2	30.1	23.8	39.5	31.7	26.2	75.1	66.7	44.9	32.5
VisPruner	**43.6**	34.2	25.8	47.8	37.9	31.6	71.5	65.5	26.9	6.9
MMTok	40.8	30.4	24.9	42.3	34.1	26.3	71.1	62.7	51.7	35.3
ApET	35.8	29.0	19.5	37.9	29.6	23.4	67.6	55.5	49.0	36.6
MiCo	**43.6**	**37.3**	**30.3**	**49.8**	**46.0**	**37.7**	**77.7**	**68.3**	**65.3**	**49.7**

Note. On LLaVA-NeXT, VisPruner and ApET follow their authors' per-crop budgets of 128/64/32 tokens for $T = 640/320/160$, so their realized layer-average token counts are not exactly matched to the other methods.

I Limitations

Proxy and hyperparameter choices. MiCo estimates information, reliability, and relevance with simple statistics that are readily available in the backbone, and uses one refinement layer and one Stage-1 pool size per architecture. These choices work well in practice; more refined estimators and input-adaptive settings are interesting directions for future work.

Selection cost. The coverage selector adds a small amount of computation for pairwise similarities. This overhead is modest in our experiments and could be reduced further with approximate similarity computation for very long inputs.

Evaluation scope. We focus on standard image and video benchmarks with short-answer prompts. Broader settings such as long-form generation and interactive use are left for future work.