Why 1 + 1 < 1 in Visual Token Pruning: Beyond Naïve Integration via Multi-Objective Balanced Covering

Yangfu Li[1,2] **Hongjian Zhan**[1,2] **Tianyi Chen**[3] **Qi Liu**[1,2] **Yue Lu**[1,2]
[1] Shanghai Key Laboratory of Multidimensional Information Processing
[2] School of Communications and Electronic Engineering, East China Normal University
[3] School of Mathematical Sciences, Shanghai Jiao Tong University
{yfli_cee, qiliu}@stu.ecnu.edu.cn, guleimurray@sjtu.edu.cn
hjzhan@cee.ecnu.edu.cn, ylu@cs.ecnu.edu.cn

Abstract

Existing visual token pruning methods target prompt alignment and visual preservation with static strategies, overlooking the varying relative importance of these objectives across tasks, which leads to inconsistent performance. To address this, we derive the first closed-form error bound for visual token pruning based on the Hausdorff distance, uniformly characterizing the contributions of both objectives. Moreover, leveraging ϵ-covering theory, we reveal an intrinsic trade-off between these objectives and quantify their optimal attainment levels under a fixed budget. To practically handle this trade-off, we propose Multi-Objective Balanced Covering (MoB), which reformulates visual token pruning as a bi-objective covering problem. In this framework, the attainment trade-off reduces to budget allocation via greedy radius trading. MoB offers a provable performance bound and linear scalability with respect to the number of input visual tokens, enabling adaptation to challenging pruning scenarios. Extensive experiments show that MoB preserves 96.4% of performance for LLaVA-1.5-7B using only 11.1% of the original visual tokens and accelerates LLaVA-Next-7B by 1.3-1.5× with negligible performance loss. Additionally, evaluations on Qwen2-VL and Video-LLaVA confirm that MoB integrates seamlessly into advanced MLLMs and diverse vision-language tasks.

1 Introduction

Multimodal large language models (MLLMs) have shown impressive performance across a variety of vision-language tasks, including visual understanding [27, 24, 17], visual question answering [37, 14, 34], and visual-language reasoning [8, 44, 42]. Since visual data exhibits much higher spatial redundancy than language, MLLMs are typically required to encode visual inputs as numerous tokens, resulting in substantial computational overhead.

To address this issue, visual token pruning methods are proposed to accelerate MLLMs by selecting representative subsets of visual tokens. Most pruning methods focus on two distinct objectives: Visual Preservation (VP) [5, 7, 56, 43], which retains tokens by minimizing redundancy or maximizing visual salience, and Prompt Alignment (PA) [55, 48, 45], which selects tokens most relevant to the prompt. Recently, several multi-objective approaches [28, 48, 39] have been proposed to integrate VP and PA through various complex strategies. Counterintuitively, these methods do not exhibit dominant superiority compared to single-objective approaches, as shown in Figure 1(a). This observation naturally raises a question: *Does integrating different objectives offer fundamental advantages?*

Inspired by this question, we formulate preservation using the *Hausdorff distance* between the original and pruned token sets and derive the first closed-form error bound for visual token pruning (Lemma 1). This bound depends on VP and PA, while it is also affected by a prompt-visual coupling, measured by

Preprint. Under review.

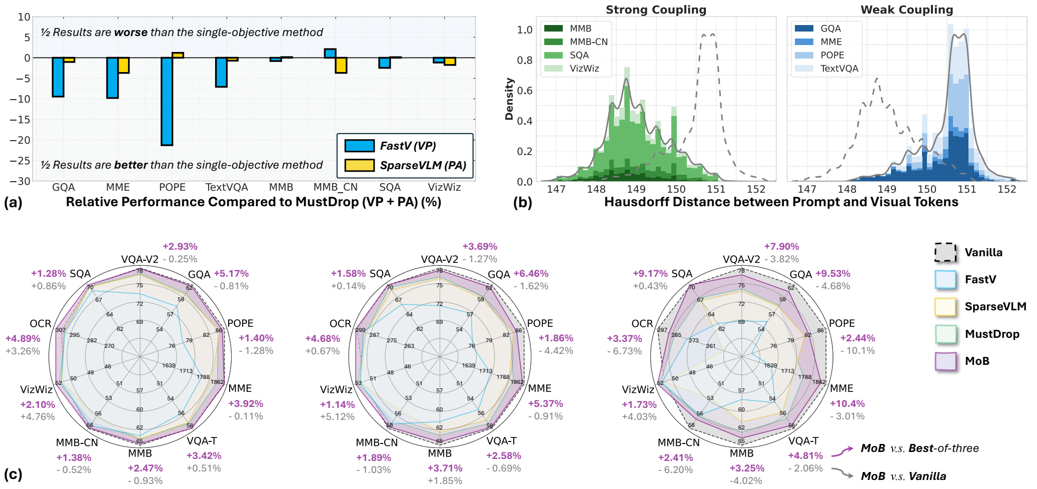


Figure 1: (a) Comparison of single- vs. bi-objective pruning methods on LLaVA-1.5-7B at a 66.7% pruning rate; (b) distribution of the prompt-visual coupling, revealing two distinct patterns across various tasks: weak coupling (large distance) and strong coupling (small distance); (c) radar charts of LLaVA-1.5-7B with visual tokens reduced from 576 to 192, 128, and 64 (*left-to-right*), demonstrating the consistent improvements of MoB across 10 well-recognized benchmarks.

the Hausdorff distance between prompt and visual tokens. Notably, we identify two patterns of this coupling across popular benchmarks, as presented in Figure 1(b): weak coupling with large distance (*e.g.*, TextVQA, POPE) and strong coupling with small distance (*e.g.*, MMB, VizWiz). Our further analysis reveals that the effectiveness of the pruning objectives varies under distinct coupling patterns (Lemma 2). However, existing multi-objective methods overlook this variation and integrate VP and PA via constant strategies, yielding inconsistent improvements over single-objective baselines.

To quantify the effect of prompt-visual coupling, we reexamine visual token pruning from a geometric covering perspective. In this view, the retained tokens can be thought of as the union of two disjoint covers for prompt and visual tokens, where each objective corresponds to a Hausdorff covering radius, and the prompt-visual coupling is represented by the inter-cover diameter. By analyzing the geometric relationship between the radii and the diameter, we reveal an intrinsic trade-off between the two objectives (Theorem 1), which *identifies the optimal attainment level of each objective to achieve the performance ceiling under a fixed pruning budget and prompt-visual coupling.*

For a practical solution to this trade-off, we propose Multi-objective Balanced Covering (MoB), a training-free visual token pruning method with provable performance guarantees and multilinear complexity (Theorem 2). MoB partitions the retained tokens into two disjoint subsets for PA and VP, employing greedy radius-trading strategies to reduce the trade-off in objective attainment to a budget allocation problem. This allows MoB to achieve the optimal balance under each coupling pattern by selecting appropriate subset sizes. As shown in Figure 1(c), MoB consistently outperforms both single-objective and multi-objective baselines by a clear margin at identical pruning rates. Besides, MoB accelerates LLaVA-Next-7B by 1.3-1.5× with negligible performance loss. Ablation studies further validate our theoretical analysis. Our key contributions are summarized as follows:

- To our knowledge, we present the first closed-form error bound for visual token pruning and its practical relaxation, characterizing the contributions of the two objectives to preservation quality.
- We quantify the trade-off between the objectives and identify their optimal attainment level under a fixed budget and prompt-visual coupling, offering valuable insights into visual token pruning.
- We propose Multi-objective Balanced Covering (MoB) for training-free visual token pruning, which reduces the trade-off of objective attainment to a budget allocation problem via two greedy radius-trading strategies, yielding both a provable performance guarantee and multilinear scalability.
- Extensive experiments across 14 public benchmarks demonstrate the superiority of MoB. For instance, it retains 96.4% and 97.9% performance for LLaVA-1.5-7B and Video-LLaVA-7B with an 88.9% reduction ratio, outperforming the second-best method by 2.7% and 1.6%, respectively. MoB can also be readily incorporated into advanced MLLMs, such as LLaVA-Next and Qwen2-VL.

2 Background

2.1 Related Work

Multimodal Large Language Model (MLLM). MLLMs [27, 18, 57, 25] have achieved remarkable progress in vision-language reasoning, owing to their robust cross-modality modeling via attention mechanisms [40, 31]. However, the spatial redundancy inherent in visual signals typically leads to a large number of input tokens [22, 19, 26, 41], particularly in high-resolution images and multi-frame videos (*e.g.*, 2048 tokens in Video-LLaVA [24]). This issue exacerbates the quadratic scaling problem of attention mechanisms, posing significant computational challenges. Moreover, to further enhance the visual capability by incorporating high-quality details, advanced MLLMs are now designed to support higher resolution images [21, 10, 9, 3], thereby necessitating the processing of even more visual tokens (*e.g.*, 2880 tokens in LLaVA-NEXT [26]). In these scenarios, effectively selecting representative visual tokens becomes a critical requirement for the real-world application of MLLMs.

Visual Token Pruning. Due to the spatial redundancy, inputs to MLLMs contain numerous less informative visual tokens. Visual token pruning accelerates MLLMs by selectively retaining only the most critical tokens during inference. Existing methods typically focus on either visual preservation (VP) [5, 35, 7, 49, 54, 29, 43] or prompt alignment (PA) [55, 48, 45]. VP-driven methods, such as ToMe [5] and LLaVA-PruMerge [35], reduce redundancy by merging similar tokens, while FastV [7] and FasterVLM [54] select tokens based on visual salience. PA-driven approaches like SparseVLM [55] rely on cross-modal attention to identify prompt-relevant tokens. More recently, MustDrop [28] integrates VP and PA through a multi-stage pruning pipeline, reporting notable improvements. Despite these advances, existing methods largely overlook the varying relative importance of VP and PA across different scenarios. In this paper, we formally characterize the contribution of each objective under a fixed pruning budget, and propose an algorithm that balances these objectives per scenario, yielding consistent improvements across diverse pruning conditions.

2.2 Preliminaries

Pipeline of MLLM. MLLMs perform vision-language reasoning by jointly processing multimodal inputs in a shared representation space. Formally, given visual tokens $\mathcal{V}^{(1)}$ extracted from the visual inputs and prompt tokens $\mathcal{P}^{(1)}$ encoded from user prompts, the multimodal input is defined as

$$\mathcal{X}^{(1)} \;=\; \mathcal{V}^{(1)} \sqcup \mathcal{P}^{(1)}, \quad \mathcal{V}^{(1)} = \{v_1^{(1)}, \ldots, v_N^{(1)}\},\; \mathcal{P}^{(1)} = \{p_1^{(1)}, \ldots, p_L^{(1)}\} \subseteq \mathbb{R}^d,$$

where N and L denote the numbers of visual and prompt tokens, respectively. We regard both $\mathcal{V}^{(1)}$ and $\mathcal{P}^{(1)}$ as compact sets on d-dimensional Euclidean space $(\mathbb{R}^d, \|\cdot\|)$. The input $\mathcal{X}^{(1)}$ is then fed into a language model $\mathcal{F}_{[1,I]}$ with I transformer block, and the final output is given by

$$y = \mathcal{F}_{[1,I]}\big(\mathcal{X}^{(1)}\big) \quad \text{where} \quad \mathcal{F}_{[1,I]} = f_I \circ f_{I-1} \circ \ldots \circ f_1,$$

In particular, each f_ℓ follows the standard Transformer (*e.g.*, multi-head self-attention [40], layer normalization [2, 47]). The intermediate feature for any layer $\ell \in \{2, \ldots, I\}$ is defined as

$$\mathcal{X}^{(\ell)} \;:=\; \mathcal{F}_{[1,\ell-1]}\big(\mathcal{X}^{(1)}\big) \;=\; \mathcal{V}^{(\ell)} \,\sqcup\, \mathcal{P}^{(\ell)}, \qquad \mathcal{F}_{[1,\ell-1]} \;:=\; f_{\ell-1} \circ \ldots \circ f_1,$$

with $\mathcal{V}^{(\ell)}$ and $\mathcal{P}^{(\ell)}$ representing the visual and prompt tokens after $\ell-1$ layers, respectively.

Visual Token Pruning. To accelerate MLLMs with minimal performance loss, visual token pruning selectively removes less-informative visual tokens at chosen intermediate layers of the language model $\mathcal{F}_{[1,I]}$. Specifically, for any chosen layer f_ℓ, $\ell \in \{2, \ldots, I\}$, pruning algorithms first select a subset $\mathcal{S}^{(\ell)} \subseteq \mathcal{V}^{(\ell)}$ of size K (*i.e.*, pruning budget) and form the pruned input $\mathcal{X}_{\mathrm{s}}^{(\ell)} = \mathcal{S}^{(\ell)} \sqcup \mathcal{P}^{(\ell)}$. The corresponding output before and after pruning are then defined as

$$y = \mathcal{F}_{[\ell,I]}\big(\mathcal{X}^{(\ell)}\big), \quad y_{\mathrm{s}} = \mathcal{F}_{[\ell,I]}\big(\mathcal{X}_{\mathrm{s}}^{(\ell)}\big) \quad \text{where} \quad \mathcal{F}_{[\ell,I]} := f_I \circ \cdots \circ f_\ell.$$

Finally, the objective of visual token pruning is formulated as

$$\mathcal{S}^{(\ell)\,*} = \mathrm{argmin}_{\mathcal{S}^{(\ell)} \subseteq \mathcal{V}^{(\ell)},\ |\mathcal{S}^{(\ell)}| = K}\, \|y - y_{\mathrm{s}}\|_2.$$

Notation. For brevity we omit the layer index (ℓ) and simply write $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P}$ and $\mathcal{X}_s = \mathcal{S} \sqcup \mathcal{P}$ to denote the input and its pruned counterpart at an arbitrary layer f_ℓ. We use $\mathcal{F}$ to denote any composition mapping of the full model $\mathcal{F}_{[1,I]}$. Finally, we let $\|\cdot\|$ denote the Euclidean norm.

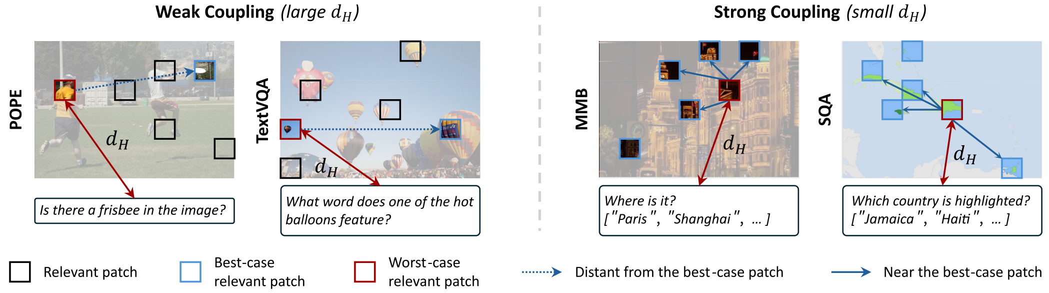

Weak Coupling (large d_H)
Strong Coupling (small d_H)
POPE
TextVQA
MMB
SQA
d_H
d_H
d_H
d_H
Is there a frisbee in the image?
What word does one of the hot balloons feature?
Where is it? ["Paris", "Shanghai", …]
Which country is highlighted? ["Jamaica", "Haiti", …]
Relevant patch
Best-case relevant patch
Worst-case relevant patch
Distant from the best-case patch
Near the best-case patch

Figure 2: Illustration of prompt-visual coupling with two distinct patterns: In fine-grained tasks (*e.g.* POPE), only a few patches are critical, so the worst-case patch lies far from best-case ones, resulting in a large Hausdorff distance and making prompt alignment valuable. In coarse-grained tasks (*e.g.* MMB), many relevant patches contain the answer cues; thus, the worst-case patch remains close to best-case ones, yielding a small Hausdorff distance and making visual preservation more efficient.

3 Methodology

3.1 Revisiting Visual Token Pruning: Insights into Prompt-Visual Coupling

As shown in Fig. 1(a), multi-objective pruning methods fail to achieve the expected improvements, and objective-specific methods exhibit inconsistent performance across benchmarks. These observations motivate us to reexamine the problem of visual token pruning. We begin by introducing Assumption 1, which quantifies pruning performance in terms of the preservation of the original token set.

Assumption 1 (Lipschitz Continuity w.r.t. the Hausdorff Distance)**.** *Assume every partial composition $\mathcal{F}$ (from layer ℓ to I) of the language model is Lipschitz continuous w.r.t. the Hausdorff distance with constant $C_\ell \geq 1$. Formally, for any intermediate token sets $\mathcal{X}, \mathcal{X}_\mathrm{s} \subset \mathbb{R}^d$,*

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_\mathrm{s})\| \;\leq\; C_\ell \, d_H(\mathcal{X}, \mathcal{X}_\mathrm{s}),$$

where d_H is the Hausdorff distance induced by the Euclidean norm:

$$d_H(\mathcal{X}, \mathcal{X}_\mathrm{s}) := \max\Big\{ \sup_{x\in\mathcal{X}} \inf_{x_\mathrm{s}\in\mathcal{X}_\mathrm{s}} \|x - x_\mathrm{s}\|,\; \sup_{x_\mathrm{s}\in\mathcal{X}_\mathrm{s}} \inf_{x\in\mathcal{X}} \|x - x_\mathrm{s}\| \Big\}. \tag{1}$$

Subsequently, we measure the preservation of the original token set $\mathcal{X}$ using three pairwise distances among visual tokens $\mathcal{V}$, retained tokens $\mathcal{S}$, and prompt tokens $\mathcal{P}$, thereby establishing a unified performance bound for various visual token pruning algorithms, as presented in Lemma 1.

Lemma 1 (An Error Bound for Visual Token Pruning)**.** *Under Assumption 1, given any token set with its pruned counterpart $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P},\;\; \mathcal{X}_\mathrm{s} = \mathcal{S} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$, the pruning error bound is given by:*

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_\mathrm{s})\| \;\leq\; C_\ell \;\max\Big\{ \min\big\{d_H(\mathcal{S},\mathcal{V}),\; d_H(\mathcal{V},\mathcal{P})\big\},\; \min\big\{d_H(\mathcal{S},\mathcal{V}),\; d_H(\mathcal{S},\mathcal{P})\big\}\Big\}.$$

Remark. *Here $d_H(\mathcal{S},\mathcal{P})$ and $d_H(\mathcal{S},\mathcal{V})$ describe the prompt alignment and visual preservation, while $d_H(\mathcal{V},\mathcal{P})$ is an inherent term that describes the prompt-visual coupling of input data.*

Proof in Appendix E.1. By Lemma 1, in practical settings where $|\mathcal{S}| \ll |\mathcal{V}|$, pruning performance is governed by a non-trivial interaction among visual preservation, prompt alignment, and prompt-visual coupling. However, existing multi-objective methods typically overlook the coupling term $d_H(\mathcal{V},\mathcal{P})$ and statically combine the two objectives across tasks, limiting their effectiveness. Our empirical evidence across popular benchmarks validates two distinct patterns of $d_H(\mathcal{V},\mathcal{P})$, each favoring different pruning objectives, as shown in Figure 2. To further explicate the effect of prompt-visual coupling, we introduce Assumption 2 and propose a practical relaxed error bound in Lemma 3.

Assumption 2 (Prompt-Visual Coupling Bound)**.** *We assume the input visual data and prompts are not entirely unrelated; hence, there exists a constant $\eta > 0$ for any intermediate token set $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$ such that $d_H(\mathcal{V},\mathcal{P}) \leq \eta$, ensuring the reasonability of vision-language reasoning.*

Lemma 2 (A Relaxed Error Bound under Practical Budgets)**.** *Under Assumptions 1 and 2, let $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P},\;\; \mathcal{X}_\mathrm{s} = \mathcal{S} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$ with $|\mathcal{S}| = K \ll N$. Partition the retained token set $\mathcal{S}$ into two disjoint subsets: $\mathcal{S} = \mathcal{S}_\mathrm{p} \sqcup \mathcal{S}_\mathrm{v}$, devoted to prompt alignment $d_H(\mathcal{S}_\mathrm{p},\mathcal{P})$ and visual preservation $d_H(\mathcal{S}_\mathrm{v},\mathcal{V})$, respectively. Then, the pruning error bound reduces to*

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_\mathrm{s})\| \;\leq\; C_\ell \;\max\big\{d_H(\mathcal{S}_\mathrm{p},\mathcal{P}),\; d_H(\mathcal{S}_\mathrm{v},\mathcal{V})\big\} + C_\ell\, \eta.$$

Proof in Appendix E.2. As Lemma 2 indicates, under weak coupling (large η), most visual regions are distant from prompt tokens in the semantic space. Consequently, if $\mathcal{S}_{\mathrm{p}}$ misses the critical patches, $d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P})$ dominates the pruning error, making the selection of $\mathcal{S}_{\mathrm{p}}$ *i.e.*, prompt alignment, more significant. Conversely, under strong coupling (small η), $d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P})$ tends to decrease in tandem with $d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V})$, reducing the marginal benefit of prompt alignment. To further guide pruning methods design, we next quantify this trade-off governed by η through an ϵ-covering argument.

3.2 Quantifying Prompt-Visual Trade-Off: A Geometric Covering Perspective

We first introduce some geometric metrics in Definition 1, recasting each objective term $d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P})$ and $d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V})$ as covering radii and the coupling term $d_H(\mathcal{V}, \mathcal{P})$ as an inter-cover diameter. Next, we relate each recasted objective to its token budget $|\mathcal{S}_{\mathrm{p}}|, |\mathcal{S}_{\mathrm{v}}|$ via covering regularity in Lemma 3. Finally, by loading the budget constraint and applying the triangle inequality between radii and diameter, we derive a quantitative trade-off jointly governed by K and η in Theorem 1.

Definition 1 (ϵ-cover, Covering Number, and Covering Regularity)**.** *Let* $(\mathbb{R}^d, \|\cdot\|)$ *be the* d*-dimensional Euclidean space and let* $\mathcal{X} \subseteq \mathbb{R}^d$ *be a compact set.*

(a) ***ϵ-cover.*** *if there exists a finite set* $\mathcal{C} = \{c_1, \ldots, c_M\} \subset \mathbb{R}^d$*, an* ϵ-cover *of* $\mathcal{X}$ *is given by*

$$\mathcal{X} \subseteq \bigcup_{c \in \mathcal{C}} B(c, \epsilon), \qquad B(c, \epsilon) \coloneqq \{x \in \mathbb{R}^d : \|x - c\| \ \le \ \epsilon\},$$

where $\mathcal{C}$ *is the collection of covering centers, and* ϵ *is the covering radius.*

(b) ***Covering number.*** *The minimum cardinality of* $\mathcal{C}$ *is the* covering number *of* $\mathcal{X}$ *at raduis* ϵ*:*

$$\mathcal{N}(\mathcal{X}, \epsilon) \ \coloneqq \ \min\Big\{M \in \mathbb{N} \ : \ \exists \mathcal{C} \subset \mathbb{R}^d, \ |\mathcal{C}| = M, \ \mathcal{X} \subseteq \bigcup_{c \in \mathcal{C}} B(c, \epsilon)\Big\}.$$

(c) ***Covering regularity.*** *We say that* $\mathcal{X}$ *satisfies* d-dimensional covering regularity *if there exist constants* $0 < A \ \le \ B$ *and* $\epsilon_0 > 0$ *such that*

$$A\,\epsilon^{-d} \ \le \ \mathcal{N}(\mathcal{X}, \epsilon) \ \le \ B\,\epsilon^{-d}, \qquad \forall\, \epsilon \in (0, \epsilon_0].$$

Based on Definition 1(a) (b), $\mathcal{S}_{\mathrm{p}}, \ \mathcal{S}_{\mathrm{v}} \subseteq \mathcal{V}$ can be thought of as two collections of centers such that

$$\mathcal{P} \subseteq \bigcup_{i=1}^{K_{\mathrm{p}}} B(s_{\mathrm{p}}^{(i)}, \epsilon_{\mathrm{p}}), \ \ \mathcal{V} \subseteq \bigcup_{j=1}^{K_{\mathrm{v}}} B(s_{\mathrm{v}}^{(j)}, \epsilon_{\mathrm{v}}),$$

where the radii are given by $\epsilon_{\mathrm{p}} \coloneqq d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}), \ \ \epsilon_{\mathrm{v}} \coloneqq d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V})$, and the covering numbers satisfy $\mathcal{N}(\mathcal{P}, \epsilon_{\mathrm{p}}) \ \le \ |\mathcal{S}_{\mathrm{p}}|, \ \ \mathcal{N}(\mathcal{V}, \epsilon_{\mathrm{v}}) \ \le \ |\mathcal{S}_{\mathrm{v}}|$. Thereby, we derive a lower bound of the required budget, *i.e.*, $|\mathcal{S}_{\mathrm{p}}|, |\mathcal{S}_{\mathrm{v}}|$, to improve each objective, *i.e.*, $\epsilon_{\mathrm{p}}, \epsilon_{\mathrm{v}}$, based on d_{eff}-dimensional covering regularity.

Lemma 3 (Covering Number Bounds)**.** *Gievn* $\mathcal{P}, \mathcal{V} \subset \mathbb{R}^d$ *with an effective dimension* d_{eff}*. Suppose their* δ*-dilations* $\mathcal{V}_\delta \coloneqq \bigcup_{v \in \mathcal{V}} B(v, \delta)$*,* $\mathcal{P}_\delta \coloneqq \bigcup_{p \in \mathcal{P}} B(p, \delta)$ *(*$\delta \ll \eta$*) satisfy* d_{eff}*-dimensional covering regularity; thus, there exist constants* $b > a > 0$*,* $b' > a' > 0$ *and* $\epsilon_0 > \delta$ *such that*

$$a\,\epsilon_{\mathrm{p}}^{-d_{\mathrm{eff}}} \le \mathcal{N}(\mathcal{P}, \epsilon_{\mathrm{p}}) \le b\,\epsilon_{\mathrm{p}}^{-d_{\mathrm{eff}}}, \qquad a'\,\epsilon_{\mathrm{v}}^{-d_{\mathrm{eff}}} \le \mathcal{N}(\mathcal{V}, \epsilon_{\mathrm{v}}) \le b'\,\epsilon_{\mathrm{v}}^{-d_{\mathrm{eff}}}, \qquad \forall\, \epsilon_{\mathrm{p}}, \epsilon_{\mathrm{v}} \in (\delta, \epsilon_0],$$

Remark. *Previous work suggests that both visual and language embeddings concentrate on a low-dimensional manifold, so the effective covering dimension satisfies the typical relation* $d_{\mathrm{eff}} \ll d$*.*

Proof in Appendix E.3. Lemma 3 demonstrates that once the radius (*i.e.*, the objective) falls below ϵ_0, any further improvement of it demands a $\Theta(\epsilon^{-d_{\mathrm{eff}}})$ increase in the number of selected token.

By loading Lemma 3 into the budget constraint: $|\mathcal{S}_{\mathrm{p}}| + |\mathcal{S}_{\mathrm{v}}| = K$, and applying a two-step triangle inequality between the covering radii $\epsilon_{\mathrm{p}}, \epsilon_{\mathrm{v}}$ and the inter-cover diameter η, we establish a K-η-bound in Theorem 1(b), which quantifies the trade-off governed by the budget and prompt-visual coupling.

Theorem 1 (Trade-off between Prompt Alignment and Visual Preservation)**.** *Under Assumption 2 and the covering-regularity hypothesis of Lemma 3 with constants* $a, a', d_{\mathrm{eff}} > 0$*, there exist a radius-scaling factor* $z > 1$ *such that* $\eta/z > \delta$ *and* $K < \mathcal{N}(\mathcal{P}, \eta/z) + \mathcal{N}(\mathcal{V}, \eta/z)$*, for every pruning results* $\mathcal{S} = (\mathcal{S}_{\mathrm{p}} \sqcup \mathcal{S}_{\mathrm{v}}) \subseteq \mathcal{V}$ *with budget* K *satisfying*

$$\max\big\{D_1 K^{-2/d_{\mathrm{eff}}}, \ D_2\,\eta^2\big\} \ \le \ d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P})\, d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}),$$

where $D_1 \coloneqq (a\,a')^{1/d_{\mathrm{eff}}}\, 4^{1/d_{\mathrm{eff}}} > 0$*,* $D_2 \coloneqq 1/z^2 > 0$*.*

Remark (Optimal Attainment Level)**.** *The term* $D_1\, K^{-2/d_{\mathrm{eff}}}$ *is completely determined by the pruning budget, while* $D_2\,\eta^2$ *quantifies the effect of prompt-visual coupling. The optimal attainment level per objective is given by* $\epsilon^* = \max\{\eta/z, \ \sqrt{D_1}\, K^{-1/d_{\mathrm{eff}}}\}$*. Any attempt to reduce one objective below* ϵ^* *forces the other above* ϵ^**, thereby increasing the overall pruning error.*

Remark (Effect of Budget and Coupling Strength)**.** *As K decreases, z correspondingly shrinks (D_2 growing as a power function), ultimately making $D_2\,\eta^2$ dominate the bound; while as K increases, both of the terms reduce, thereby diminishing the trade-off and tightening the overall error bound.*

Proof in Appendix E.4. Theorem 1 characterizes the optimal attainment level for each objective under a fixed pruning budget and prompt-visual coupling. However, it is actually very challenging to dynamically determine the attainment level per objective during the pruning process. To address this, we propose Multi-objective Balanced Covering, which leverages the monotonic relationship between covering radii and numbers to reduce the trade-off of attainment to a budget-allocation problem.

3.3 Multi-Objective Balanced Covering: From Trade-Off to Budget Allocation

Motivated by the insights in §3.2, Multi-objective Balanced Covering (MoB) recasts visual token pruning as bi-objective covering. Specifically, given a token set $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$ with a budget K, the retained token set $\mathcal{S}$ is defined as the union of a prompt center set $\mathcal{S}_\mathrm{p}$ and a visual center set $\mathcal{S}_\mathrm{v}$:

$$\mathcal{S} = \mathcal{S}_\mathrm{p} \sqcup \mathcal{S}_\mathrm{v} \subseteq \mathcal{V} \subseteq \mathbb{R}^d \quad \text{where} \quad \mathcal{P} \subset \textstyle\bigcup_{i=1}^{K_\mathrm{p}} B(s_\mathrm{p}^{(i)}, \epsilon_\mathrm{p}),\;\; \mathcal{V} \subset \textstyle\bigcup_{j=1}^{K-K_\mathrm{p}} B(s_\mathrm{v}^{(j)}, \epsilon_\mathrm{v}).$$

MoB then selects the cover centers (*i.e.*, retained tokens) by minimizing the overall maximum radius:

$$(\mathcal{S}_\mathrm{p}^*,\, \mathcal{S}_\mathrm{v}^*) = \operatorname*{arg\,min}_{\mathcal{S}_\mathrm{p} \sqcup \mathcal{S}_\mathrm{v} \subseteq \mathcal{V},\; |\mathcal{S}_\mathrm{p}| = K_\mathrm{p},\; |\mathcal{S}_\mathrm{v}| = K - K_\mathrm{p}} \max\{\epsilon_\mathrm{p}(\mathcal{S}_\mathrm{p}), \epsilon_\mathrm{v}(\mathcal{S}_\mathrm{v})\}.$$

In practice, MoB solves this problem approximately by two sequential greedy covering procedures: selection of prompt center set $\mathcal{S}_\mathrm{p}$ with budget K_p, and selection of visual center set $\mathcal{S}_\mathrm{v}$ with the remaining budget $K - K_\mathrm{p}$. By the covering number bounds given in Lemma 3, we have

$$K_\mathrm{p} = \Theta(\epsilon_\mathrm{p}^{-d_\mathrm{eff}}), \quad K - K_\mathrm{p} = \Theta(\epsilon_\mathrm{v}^{-d_\mathrm{eff}}),$$

where d_eff is the effective dimension of $\mathcal{V}$, $\mathcal{P}$. Accordingly, by selecting the unique budget K_p (*i.e.*, fixing the remaining budget $K - K_\mathrm{p}$) under each coupling pattern, MoB ensures $\epsilon_\mathrm{p}, \epsilon_\mathrm{v} = \Omega\big(\max\{\eta/z,\ \sqrt{D_1}\,K^{-1/d_\mathrm{eff}}\}\big)$, thus yielding provable performance guarantees across scenarios.

Normalization. For efficiency, MoB applies L2 normalization to each $x \in \mathcal{X}$ so that $\|x\| = 1$. Hence, for any token pair $x_1, x_2 \in \mathcal{X}$, the Euclidean distance can be induced by their cosine similarity:

$$\|x_1 - x_2\| = \sqrt{2 - 2\cos(x_1, x_2)}.$$

Selection of Prompt Center Set $\mathcal{S}_\mathrm{p}$. Since all $s_\mathrm{p} \in \mathcal{V}$ lie outside $\mathcal{P}$, a typical solution for minimizing the radius ϵ_p is *Nearest-Neighbor covering* (NN covering) [13], which uniformly allocates the nearest $s_\mathrm{p} \in \mathcal{V}$ for each prompt token. However, the contribution of each prompt token is inequivalent, especially under weak prompt-visual coupling; thus, equal allocation risks missing the "best-case tokens." To remedy this, we introduce a k-fold NN covering procedure. Formally, let $L = |\mathcal{P}|$ and $k > 1$ be a hyperparameter; we first utilize a temporary budget of kL to form a candidate set.

$$\mathcal{S}'_\mathrm{p} = \textstyle\bigcup_{p \in \mathcal{P}} \arg\operatorname{topk}_{s \in \mathcal{V}}\big(\cos(s, p),\, k\big), \quad |\mathcal{S}'_\mathrm{p}| \geq K_\mathrm{p},$$

thereby over-sampling the k nearest visual tokens for each prompt token. Subsequently, we refine the candidate set by selecting the final K_p centers that maximize their worst-case alignment with $\mathcal{P}$:

$$\mathcal{S}_\mathrm{p} = \arg\operatorname{topk}_{s \in \mathcal{S}'_\mathrm{p}}\big(\max_{p \in \mathcal{P}} \cos(s, p),\, K_\mathrm{p}\big).$$

By concentrating the limited budget on those visual tokens most strongly aligned with the key prompt tokens, this strategy ensures a better preservation of the critical regions in the visual input. We determine the appropriate k by ablation to avoid the oversampling of a few salient prompt tokens.

Selection of Visual Center Set $\mathcal{S}_\mathrm{v}$. Unlike the prompt center selection, each visual center s_v lies in $\mathcal{V}$. Thereby, we employ *Farthest Point Sampling* (FPS) [33] on the remaining tokens, *i.e.*, $\mathcal{V} \setminus \mathcal{S}$, to select the visual centers, which makes the visual centers $\mathcal{S}_\mathrm{v}$ well-spread over $\mathcal{V}$, minimizing the covering radius ϵ_v. Concretely, FPS operates by iteratively selecting the token farthest (*i.e.*, the most different) from the current centers $\mathcal{S}$, where the distance is given by

$$\mathrm{dist}_\mathrm{FPS}(s_\mathrm{v}, \mathcal{S}) = \min_{s \in \mathcal{S}}(1 - \cos(s_\mathrm{v}, s)), \quad \forall s_\mathrm{v} \in \mathcal{V} \setminus \mathcal{S}.$$

Subsequently, we initialize the visual centers with the empty set, *i.e.*, $\mathcal{S}_\mathrm{v}^{(1)} := \varnothing$. We then successively add the farthest visual token to the current centers $\mathcal{S}_\mathrm{v}^{(i)} \sqcup \mathcal{S}_\mathrm{p}$ until it contains a total of K elements. Hence, the visual centers at the subsequent iteration, $\mathcal{S}_\mathrm{v}^{(i+1)}$, is given by:

$$\mathcal{S}_\mathrm{v}^{(i+1)} = \mathcal{S}_\mathrm{v}^{(i)} \sqcup \arg\max_{s_\mathrm{v} \in \mathcal{V} \setminus \left(\mathcal{S}_\mathrm{v}^{(i)} \sqcup \mathcal{S}_\mathrm{p}\right)} \mathrm{dist}_\mathrm{FPS}(s_\mathrm{v},\, \mathcal{S}_\mathrm{v}^{(i)} \sqcup \mathcal{S}_\mathrm{p}), \quad \text{for } i \in [1, \ldots, K - K_\mathrm{p}].$$

More details of the proposed MoB algorithm are provided in Appendix B.

Method	Objectives	Strong Coupling				Weak Coupling						Avg.
		MMB	MMB_{CN}	SQA	VizWiz	GQA	MME	POPE	VQA^{T}	VQA^{V2}	OCR	
LLaVA-1.5-7B		*w/o Pruning,* $N = 576$; *Token Reduction Rate* = **0.0%**										
Vanilla [25]	-	64.7	58.1	69.5	50.0	61.9	1862	85.9	58.2	78.5	297	100%
LLaVA-1.5-7B		*Pruning budget* $K = 192$; *Token Reduction Rate* = **66.7%**										
FastV (ECCV'24) [7]	VP	61.2	57.0	67.3	50.8	52.7	1612	64.8	52.5	67.1	291	91.2%
SparseVLM (ICML'25) [55]	PA	62.5	53.7	69.1	50.5	57.6	1721	83.6	56.1	75.6	292	96.3%
MustDrop (24.11) [28]	PA VP	62.3	55.8	69.2	51.4	58.2	1787	82.6	56.5	76.0	289	97.2%
DART (25.02) [43]	VP	63.6	57.0	69.8	51.2	60.0	1856	82.8	57.4	76.7	296	98.8%
MoB (Ours)	PA VP	64.1	57.8	70.1	52.5	61.4	1860	84.8	58.5	78.3	307	100.6%
LLaVA-1.5-7B		*Pruning budget* $K = 128$; *Token Reduction Rate* = **77.8%**										
FastV (ECCV'24)	VP	56.1	56.4	60.2	51.3	49.6	1490	59.6	50.6	61.8	285	86.4%
SparseVLM (ICML'25)	PA	60.0	51.1	67.1	51.4	56.0	1696	80.5	54.9	73.8	280	93.8%
MustDrop (24.11)	PA VP	61.1	55.2	68.5	52.1	56.9	1745	78.7	56.3	74.6	281	95.6%
DART (25.02)	VP	63.2	57.5	69.1	51.7	58.7	1840	80.1	56.4	75.9	296	98.0%
MoB (Our)	PA VP	63.5	57.5	69.6	52.7	60.9	1845	82.1	57.8	77.5	299	99.4%
LLaVA-1.5-7B		*Pruning budget* $K = 64$; *Token Reduction Rate* = **88.9%**										
FastV (ECCV'24)	VP	48.0	52.7	51.1	50.8	46.1	1256	48.0	47.8	55.0	245	77.3%
SparseVLM (ICML'25)	PA	56.2	46.1	62.2	50.1	52.7	1505	75.1	51.8	68.2	180	84.6%
MustDrop (24.11)	PA VP	60.0	53.1	63.4	51.2	53.1	1612	68.0	54.2	69.3	267	90.1%
DART (25.02)	VP	60.6	53.2	69.8	51.6	55.9	1765	73.9	54.4	72.4	270	93.7%
MoB (Our)	PA VP	62.1	54.5	69.8	52.1	59.0	1806	77.2	57.0	75.5	277	96.4%
LLaVA-Next-7B		*w/o Pruning,* $N = 2880$; *Token Reduction Rate* = **0.0%**										
Vanilla [26]	-	67.4	60.6	70.1	57.6	64.2	1851	86.5	64.9	81.8	517	100%
LLaVA-Next-7B		*Pruning budget* $K = 320$; *Token Reduction Rate* = **88.9%**										
FastV (ECCV'24)	VP	61.6	51.9	62.8	53.1	55.9	1661	71.7	55.7	71.9	374	86.4%
SparseVLM (ICML'25)	PA	60.6	54.5	66.1	52.0	56.1	1533	82.4	58.4	71.5	270	85.9%
MustDrop (24.11)	PA VP	62.8	55.1	68.0	54.0	57.3	1641	82.1	59.9	73.7	382	90.4%
FasterVLM (24.12) [54]	VP	61.6	53.5	66.5	52.6	56.9	1701	83.6	56.5	74.0	401	89.8%
DART (25.02)	VP	65.3	58.2	68.4	56.1	61.7	1710	84.1	58.7	79.1	406	93.9%
MoB (Our)	PA VP	65.8	58.9	68.7	57.0	62.6	1760	84.4	60.2	80.1	418	95.4%

Table 1: Partial comparative experiments on image understanding with the LLaVA-7B Series, where $K_{\text{p}} \in \{\frac{3K}{8}, \frac{K}{4}, \frac{K}{4}\}$, $k = \frac{3K_{\text{p}}}{40}$ for strong-coupling and $K_{\text{p}} \in \{\frac{K}{2}, \frac{7K}{16}, \frac{5K}{12}\}$, $k = \frac{K_{\text{p}}}{8}$ for weak-coupling benchmarks, corresponding to token reduction rates in $\{88.9\%, 77.8\%, 66.7\%\}$; the pruning layer index $\ell = 2$. See Appendix D for the full results.

Theorem 2 (Performance Guarantee)**.** *Under Assumption 1 and the covering-regularity of Lemma 3 with constants* $a, a', d_{\text{eff}} > 0$ *and* $b > a$, $b' > a'$, *for any budget split* $(K_{\text{p}},\ K - K_{\text{p}})$, *covering fold* k, *and token set* $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$ *with* $|\mathcal{V}| = N$, $|\mathcal{P}| = L$, *and* $d_H(\mathcal{V}, \mathcal{P}) \leq \eta$, *the following hold:*

(a) ***Performance bound:*** *The Performance degradation caused by MoB is upper bounded by*

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\text{MoB}(\mathcal{X}))\| \;\leq\; C_\ell \;\max\Big\{\alpha(\eta, k, L)\,(K_{\text{p}})^{-1/d_{\text{eff}}},\; \beta\,(K - K_{\text{p}})^{-1/d_{\text{eff}}}\Big\} \;+\; C_\ell\,\eta,$$

where $\alpha(\eta, k, L) \;=\; \eta\,\big(b\,k\,L/a\big)^{1/d_{\text{eff}}}, \quad \beta \;=\; 2(b')^{1/d_{\text{eff}}}.$

(b) ***Multilinear complexity:*** *The complexity of MoB is given by* $T_{\text{MoB}} = \mathcal{O}(N\,(L + K)\,d)$.

Remark (Coupling Trade-off)**.** *Under weak coupling (large* $\alpha(\eta, k, L)$*), minimizing the bound requires a larger* K_{p}*. Conversely, under strong coupling (small* $\alpha(\eta, k, L)$*), the alignment term decays rapidly, favoring visual preservation (increasing* $K - K_{\text{p}}$*). Specially, under perfect coupling (*$\eta = 0$*), the bound simplifies to* $\|\Delta y\| \;\leq\; C_\ell\,\beta\,(K - K_{\text{p}})^{-1/d_{\text{eff}}}$,i.e., *MoB reduces to pure visual preservation.*

Remark (Budget Scaling)**.** *As the total budget* K *increases, the preservation term* $\beta\,(K - K_{\text{p}})^{-1/d_{\text{eff}}}$ *decays, requiring a corresponding increase in* K_{p} *(and thus a reduction in the alignment term) to rebalance the trade-off and further lower the overall error bound.*

Remark (Scalability)**.** *MoB exhibits a multilinear scalability with respect to visual tokens* N*, prompt tokens* L*, and retained tokens* K *(especially* $K, L \ll N$*), making it readily adaptable to more challenging scenarios, such as advanced MLLMs with higher-resolution inputs or multi-frame video.*

Proof in Appendix E.5.

4 Experimental Results

Experiment Setting. We perform a comprehensive evaluation of the proposed MoB and several representative methods on two visual tasks: image understanding and visual understanding, together with an efficiency analysis. Our experiments employ four popular MLLMs and include a total of 14 widely recognized benchmarks. For further details regarding the benchmarks, models, baselines, and implement details please refer to Appendix C.

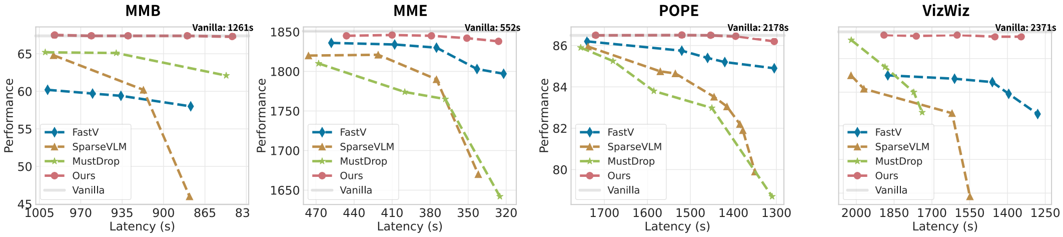


Figure 3: Performance-Latency trade-off comparisons across four benchmarks on LLaVA-Next-7B.

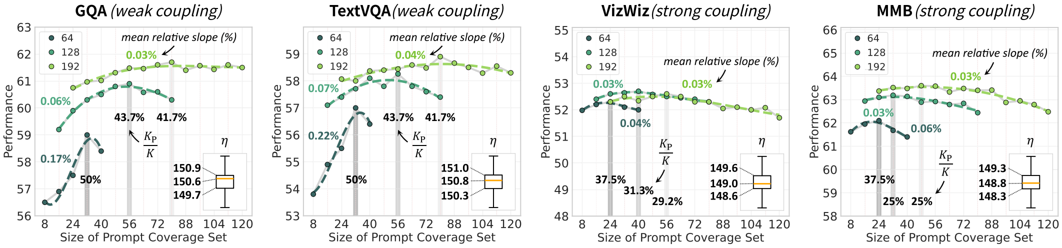


Figure 4: Comprehensive ablation on the budget configuration $\langle K_{\mathrm{p}}, K\rangle$ across four benchmarks with distinct prompt-visual coupling η on LLaVA-1.5-7B, where $K = \{64, 128, 192\}$; the *mean relative slope (%)* is given by $\frac{100}{x_n - x_1}\sum_{i=1}^{n-1}\frac{y_{i+1}-y_i}{y_i}$, quantifying the trade-off intensity; the ratio $\frac{K_{\mathrm{p}}}{K}$ reflects the cost-effectiveness of prompt alignment, and the box plot presents the distribution of η.

Image Understanding. Tables 2 and 4 report the evaluation results across a variety of image-understanding tasks on LLaVA series and Qwen2-VL, respectively. We observe that (a) single-objective baselines exhibit complementary strengths under different coupling patterns, whereas MoB consistently outperforms all baselines, demonstrating the benefit of balanced objectives; (b) the superiority of MoB becomes even more significant under aggressive token reduction. Specifically, the improvement of MoB over the best baseline in average scores increases from 1.8% at a 66.7% token reduction to 2.7% at an 88.8% reduction on LLaVA-1.5-7B; (c) MoB matches the performance of the vanilla LLaVA-1.5-7B with only 33.3% of visual tokens, which may be attributed to the mitigation of hallucinations caused by redundant tokens; and (d) MoB scales seamlessly to advanced models, preserving 95.2% performance on Qwen2-VL-7B using only 22.2% of visual tokens.

Method	GQA	MME	POPE	VQA^{T}	MMB	SQA	Avg.
Qwen2-VL-7B	*w/o Pruning*; *Token Reduction Rate* = **0.0%**						
Vanilla [41]	62.2	2317	86.1	82.1	80.5	84.7	100%
Qwen2-VL-7B	*Token Reduction Rate* = **66.7%**						
FastV	58.0	2130	82.1	77.3	76.1	80.0	94.0%
DART	60.2	2245	83.9	80.5	78.9	81.4	97.0%
MoB (Our)	61.8	2268	84.7	81.1	79.5	82.3	98.4%
Qwen2-VL-7B	*Token Reduction Rate* = **77.8%**						
FastV	56.7	2031	79.2	72.0	74.1	78.3	91.0%
DART	58.5	2175	82.1	75.3	77.3	79.6	94.3%
MoB (Our)	59.4	2203	82.8	75.8	78.1	80.4	95.2%
Qwen2-VL-7B	*Token Reduction Rate* = **88.9%**						
FastV	51.9	1962	76.1	60.3	70.1	75.8	84.4%
DART	55.5	2052	77.9	61.8	72.0	77.6	87.4%
MoB (Our)	56.5	2094	78.5	62.7	72.8	78.4	88.6%

Table 2: Comparative experiments on image understanding with Qwen2-VL-7B.

Method	TGIF	MSVD	MSRV	ActNet	Avg.
Video-LLaVA-7B	*Token Reduction Rate* = **0.0%**				
Vanilla [24]	47.1	69.8	56.7	43.1	100%
Video-LLaVA-7B	*Token Reduction Rate* = **93.4%**				
FastV (ECCV'24)	23.1	38.0	19.3	30.6	52.1%
SparseVLM (ICML'25)	44.7	68.2	31.0	42.6	86.5%
VisionZip (24.12) [48]	42.4	63.5	52.1	43.0	93.2%
TwigVLM (25.03) [36]	44.7	68.3	54.6	41.5	96.3%
MoB (Our)	45.3	68.8	55.2	42.8	97.9%

Table 3: Comparative experiments on video understanding with Video-LLaVA-7B.

Video Understanding. Experimental results presented in Table 3 demonstrate that MoB is general and can be readily extended to more challenging video scenarios without incurring additional cost. Specifically, MoB preserves 97.9% of average performance for Video-LLaVA-7B using only 6.6% of visual tokens, which sets new records in most VideoQA benchmarks, achieving 1.6% and 4.7% improvements over TwigVLM and VisionZip, respectively. These results validate the generalization ability of MoB.

Efficiency Analysis. We present the performance-latency trade-off measured on an NVIDIA A800-80GB GPU in Figure 3. The results show that (a) MoB achieves a strong performance-latency trade-off, delivering a 1.3-1.5× speed-up for LLaVA-NEXT-7B with negligible performance loss; (b) due to ignoring the K-η trade-off, the multi-stage method MustDrop is outperformed by single-objective methods FastV and SparseVLM on MME and POPE, and suffers significant performance drops as token budgets shrink (*i.e.*, latency decreases). In contrast, MoB consistently maintains a robust trade-off across all benchmarks, surpassing all the baselines by a clear margin; (c) MoB does not rely on attention scores to identify important tokens, making it compatible with flash attention and more efficient than attention-based methods such as SparseVLM and FastV.

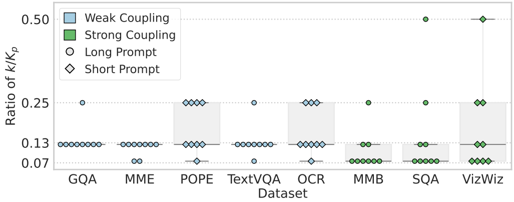


Figure 5: Ablation on the ratio of k/K_p.

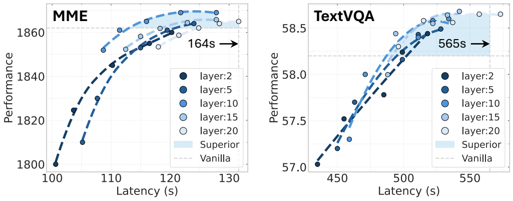


Figure 6: Ablation on the pruning layer.

5 Ablation and Discussion

Impact of $\langle K, \eta, K_\text{p}, \rangle$. We study the impact of K, η, and K_p on pruning performance across four benchmarks: GQA and TextVQA (weak coupling); VizWiz and MMB (strong coupling). As shown in Figure 4, the results can be interpreted by Theorem 1 and Theorem 2(a), respectively.

A. Theorem 1 Perspective: When K is large, *e.g.*, $K = 192$, the trade-off is governed by $D_1 K^{-2/d_\text{eff}}$, hence the trade-off intensity remains nearly identical across benchmarks. Conversely, When K is small, especially $K = 64$, in weak-coupling benchmarks, the trade-off turns to be governed by $D_2\eta^2$; thus, the trade-off intensity is obviously more pronounced in GQA and TextVQA than that in VizWiz and MMB. These observations exactly confirm the validity of Theorem 1.

B. Theorem 2(a) Perspective. (a) Under weak coupling, the alignment term $\alpha(\eta, k, L)(K_\text{p})^{-1/d_\text{eff}}$ is amplified, which requires a larger K_p to suppress the overall error. However, across benchmarks sharing the same coupling pattern, the optimal K_p values exhibit only minor variation. (b) Increasing the total budget K pushes the optimal K_p upward to rebalance the two bound terms. Since the prompt length L is fixed, adding more tokens yields diminishing returns for prompt alignment, which is reflected in the declining ratio K_p/K. These validate the performance bound in Theorem 2(a). Remarkably, the experimental results suggest that simply determining the optimal K_p for each of the two coupling patterns suffices to guarantee effective generalization across all scenarios.

Impact of Covering Fold k. We chose the covering fold k by examining the normalized ratio k/K_p across eight benchmarks and nine budget configurations. As shown in Figure 5, (a) weak-coupling benchmarks generally require a larger k to ensure critical region coverage, whereas strong-coupling settings suffice with a smaller k; (b) benchmarks with longer prompts impose a lower cap on k to preserve sampling diversity and avoid redundant selection of salient tokens. Notably, weak-coupling benchmarks with long prompts (*e.g.*, GQA, TextVQA) exhibit a narrowly clustered optimal k/K_p range, reflecting their strict requirement to cover key tokens without excessive redundancy.

Impact of Pruning Layer. As shown in Figure 6, (a) models with visual token pruning consistently achieve a more favorable performance-efficiency trade-off than the vanilla model on both benchmarks. (b) Pruning in deeper layers provides more significant benefits for the weak-coupling TextVQA than strong-coupling MME. We attribute this to stronger cross-modal interactions in deeper MLLM layers, which facilitate identification of answer-relevant tokens under weak coupling, whereas pruning in shallow layers disrupts these interactions and incurs greater performance degradation.

6 Conclusion

In this paper, we present a comprehensive analysis of visual token pruning, deriving the first closed-form error bound with a practical relaxation. Leveraging ϵ-covering theory, we quantify the intrinsic trade-off between the fundamental pruning objectives, *i.e.*, visual preservation and prompt alignment, and identify their optimal attainment levels under a fixed pruning budget. Building on these insights, we introduce MoB, a training-free algorithm for visual token pruning. Based on greedy radius trading, MoB ensures the near-optimal attainment per objective via budget allocation, offering a provable performance bound and multilinear scalability. Experimental results indicate that MoB matches the performance (100.6%) of LLaVA-1.5-7B with only 33.3% of visual tokens and can be seamlessly integrated into advanced MLLMs, such as LLaVA-Next-7B and Qwen2-VL-7B. Our work advances the understanding of visual token pruning and offers valuable insights for future MLLM compression.

Limitations. Our theoretical guarantees rely on assumption 1, which is generally satisfied in practice but may not hold for all MLLMs. Besides, MoB applies a preliminary search to select the proper K_p, which potentially introduces extra tuning overhead in practical applications. Future work will focus on developing an adaptive K_p selection mechanism driven by online estimation of the coupling η.

References

[1] Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models in resource-constrained environments. *arXiv preprint arXiv:2408.10945*, 2024.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Kaichen Zhang* Fanyi Pu* Xinrun Du Yuhao Dong Haotian Liu Yuanhan Zhang Ge Zhang Chunyuan Li Bo Li*, Peiyuan Zhang* and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024.

[5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2022.

[6] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[7] 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*, pages 19–35. Springer, 2024.

[8] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36:70115–70140, 2023.

[9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[10] 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*, pages 24185–24198, 2024.

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[12] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[13] Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.

[14] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264, 2024.

[15] 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*, pages 6700–6709, 2019.

[16] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[17] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

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

[19] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[21] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26763–26773, 2024.

[22] Youwei Liang, GE Chongjian, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Evit: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

[23] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.

[24] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024.

[25] 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*, pages 26296–26306, 2024.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[28] Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024.

[29] Xuyang Liu, Ziming Wang, Yuhang Han, Yingyao Wang, Jiale Yuan, Jun Song, Bo Zheng, Linfeng Zhang, Siteng Huang, and Honggang Chen. Compression with global guidance: Towards training-free high-resolution mllms acceleration. *arXiv preprint arXiv:2501.05179*, 2025.

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

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[32] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[33] Carsten Moenning and Neil A Dodgson. Fast marching farthest point sampling. Technical report, University of Cambridge, Computer Laboratory, 2003.

[34] Yingzhe Peng, Xinting Hu, Jiawei Peng, Xin Geng, Xu Yang, et al. Live: Learnable in-context vector for visual question answering. *Advances in Neural Information Processing Systems*, 37:9773–9800, 2024.

[35] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[36] Zhenwei Shao, Mingyang Wang, Zhou Yu, Wenwen Pan, Yan Yang, Tao Wei, Hongyuan Zhang, Ning Mao, Wei Chen, and Jun Yu. Growing a twig to accelerate large vision-language models. *arXiv preprint arXiv:2503.14075*, 2025.

[37] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 14974–14983, 2023.

[38] 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*, pages 8317–8326, 2019.

[39] Xudong Tan, Peng Ye, Chongjun Tu, Jianjian Cao, Yaoxin Yang, Lin Zhang, Dongzhan Zhou, and Tao Chen. Tokencarve: Information-preserving visual token compression in multimodal large language models. *arXiv preprint arXiv:2503.10501*, 2025.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[42] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

[43] 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. *arXiv preprint arXiv:2502.11494*, 2025.

[44] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024.

[45] 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.

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

[47] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

[48] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024.

[49] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *arXiv preprint arXiv:2409.10197*, 2024.

[50] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[51] Yuanhan Zhang Bo Li Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023.

[52] Zheng Zhan, Yushu Wu, Zhenglun Kong, Changdi Yang, Yifan Gong, Xuan Shen, Xue Lin, Pu Zhao, and Yanzhi Wang. Rethinking token reduction for state space models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1686–1697, 2024.

[53] 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, 2024.

[54] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*, 2024.

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

[56] Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal llms via token merging and pruning. *arXiv preprint arXiv:2412.03248*, 2024.

[57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Appendix

In the appendix, we provide additional information as listed below:

- §A provides the broader impacts of MoB
- §B provides the algorithm details and pseudocode of MoB
- §C provides the overview of the data, models, baselines and implementation details.
- §D provides the additional experimental results.
- §E provides the omitted technical details.

A Societal Impacts and Limitations

The proposed MoB yields substantial acceleration of MLLMs with negligible performance loss, thereby enabling high-resolution vision-language models to operate on resource-constrained platforms such as edge devices and mobile systems while supporting low-latency applications—including assistive technologies for the visually impaired, autonomous navigation, and AR/VR. Besides, MoB potentially benefits other redundancy-heavy domains (*e.g.*, point clouds and multi-sensor fusion), guiding efficient token-level compression beyond vision. Limitations arise from our reliance on Assumption 1 and lemma 3 (Lipschitz continuity and covering regularity); in embedding spaces that violate metric properties or exhibit highly irregular token distributions, the provable performance bounds may no longer hold.

B Algorithm

Algorithm 1 Multi-Objective Balanced Covering (MoB)

```
Require: Visual token $\mathbf{V} \in \mathbb{R}^{\mathtt{N}\times\mathtt{d}}$, Prompt token $\mathbf{P} \in \mathbb{R}^{\mathtt{L}\times\mathtt{d}}$, Budget $\mathtt{K}_\mathtt{p}, \mathtt{K}_\mathtt{v}$, Covering fold $\mathtt{k}$
Ensure: Index list for select tokens $\mathbf{S} \in \mathbb{N}^{\mathtt{K}_\mathtt{p}+\mathtt{K}_\mathtt{v}}$
Normalize all token embeddings to unit $\ell_2$ norm: $\mathbf{V} \leftarrow \mathbf{V}/\|\mathbf{V}\|_{2,\text{row}}, \quad \mathbf{P} \leftarrow \mathbf{P}/\|\mathbf{P}\|_{2,\text{row}}$

Step 1. Select Prompt Centers via Nearest-Neighbor Covering
Compute cosine-similarity matrix via $\mathbf{P}\,\mathbf{V}^\top$: $\mathbf{M} \leftarrow \mathbf{P}\,\mathbf{V}^\top$    ▷ $\mathbf{M} \in \mathbb{R}^{\mathtt{L}\times\mathtt{N}}$.
Retrieve $k$ nearest token indices per prompt:
    $\mathbf{C}_{\mathtt{idx}} \leftarrow \mathtt{ArgTopK}(\mathbf{M},\, \mathtt{k},\, \mathtt{axis}=1), \quad \mathbf{C}_{\mathtt{sim}} \leftarrow \mathtt{TopK}(\mathbf{M},\, \mathtt{k},\, \mathtt{axis}=1)$
    ▷ $\mathbf{C}_{\mathtt{idx}}, \mathbf{C}_{\mathtt{sim}} \in \mathbb{R}^{\mathtt{L}\times\mathtt{k}}$ collects index and similarity of $k$ closest centers per prompt token.
# Deduplicate candidate indices
Flatten index and similarity arrays: $\mathbf{C}_{\mathtt{idx}} \leftarrow \mathtt{Flatten}(\mathbf{C}_{\mathtt{idx}}), \quad \mathbf{C}_{\mathtt{sim}} \leftarrow \mathtt{Flatten}(\mathbf{C}_{\mathtt{sim}})$    ▷ $\mathbf{C}_{\mathtt{idx}} \in \mathbb{N}^{\mathtt{Lk}}$, $\mathbf{C}_{\mathtt{sim}} \in \mathbb{R}^{\mathtt{Lk}}$
Remove duplicate indices, preserving associated similarities:
    $\langle \mathbf{C}^*_{\mathtt{idx}}, \mathbf{C}^*_{\mathtt{sim}} \rangle \leftarrow \mathtt{UniqueIndices}(\mathbf{C}_{\mathtt{idx}},\, \mathbf{C}_{\mathtt{sim}})$
    ▷ $\mathtt{K}_\mathtt{p} \leq |\mathbf{C}^*_{\mathtt{idx}}| \leq \mathtt{Lk}$
Identify top-$K_\text{p}$ prompt centers by similarity: $\mathbf{i}_\text{p} \leftarrow \mathtt{ArgTopK}(\mathbf{C}^*_{\mathtt{sim}},\, \mathtt{K}_\mathtt{p})$
Form the prompt-center index list: $\mathbf{S}_\text{p} \leftarrow \mathbf{C}^*_{\mathtt{idx}}[\mathbf{i}_\text{p}]$    ▷ $\mathbf{S}_\text{p} \in \mathbb{N}^{\mathtt{K}_\mathtt{p}}$

Step 2. Select Visual Centers via Farthest-Point Sampling
Initialize selected centers: $\mathbf{S} \leftarrow \mathbf{S}_\text{p}$
# Initialize token-to-prompt minimum distances
Compute pairwise minimum distances between all tokens and selected prompt centers:
    $\mathbf{d} \leftarrow \mathbf{1}_{\mathtt{N}\times\mathtt{K}_\mathtt{p}} - \mathbf{V}\,\mathbf{V}[\mathbf{S}_\text{p}]^\top, \quad \mathbf{d} \leftarrow \mathtt{Min}(\mathbf{d},\, \mathtt{axis}=1)$
    ▷ Selected centers have zero distance in $\mathbf{d} \in \mathbb{R}^N$.
# Farthest-Point Sampling
for $t = 1$ to $K_\text{v}$ do
    Select the token farthest from current centers: $\mathtt{i}^* \leftarrow \mathtt{ArgMax}(\mathbf{d}), \quad \mathbf{S} \leftarrow \mathtt{Concat}(\mathbf{S},\, \mathtt{i}^*)$    ▷ Selected tokens are excluded (distance = 0) from further sampling.
    Compute cosine distances to the newly selected token: $\mathbf{d}_\Delta \leftarrow \mathbf{1}_\mathtt{N} - \mathbf{V}\,\mathbf{V}[\mathtt{i}^*]^\top$
    Update each token's minimum distance: $\mathbf{d} \leftarrow \mathtt{ElementwiseMin}(\mathbf{d},\, \mathbf{d}_\Delta)$    ▷ Distance of newly selected token $\mathtt{i}^*$ set to zero in $\mathbf{d}$.
end for
return S
```

Algorithm 2 Compute Prompt-Visual Coupling

```
Require: Visual embeddings $\mathbf{V} \in \mathbb{R}^{\mathtt{n}_\mathtt{v}\times\mathtt{d}}$, Prompt embeddings $\mathbf{P} \in \mathbb{R}^{\mathtt{n}_\mathtt{p}\times\mathtt{d}}$
Ensure: Hausdorff distance $\mathtt{h}(\mathbf{V}, \mathbf{P})$
Step 1. Compute Pairwise Euclidean Distances
Compute distance matrix via cdist: $\mathbf{D} \leftarrow \text{cdist}(\mathbf{V},\, \mathbf{P},\, \mathtt{p}=2)$    ▷ $\mathbf{D} \in \mathbb{R}^{\mathtt{n}_\mathtt{v}\times\mathtt{n}_\mathtt{p}}$
Step 2. Directed Hausdorff Distances
Visual-to-prompt directed distance:
    $\mathtt{d}_{\mathtt{v}\to\mathtt{p}},\, \_ \leftarrow \min(\mathbf{D},\, \text{axis}=2) \quad , \quad \mathtt{h}_{\mathtt{v}\to\mathtt{p}} \leftarrow \max(\mathtt{d}_{\mathtt{v}\to\mathtt{p}})$
Prompt-to-visual directed distance:
    $\mathtt{d}_{\mathtt{p}\to\mathtt{v}},\, \_ \leftarrow \min(\mathbf{D},\, \text{axis}=1) \quad , \quad \mathtt{h}_{\mathtt{p}\to\mathtt{v}} \leftarrow \max(\mathtt{d}_{\mathtt{p}\to\mathtt{v}})$
Step 3. Final Hausdorff Distance
return $\max\big(\mathtt{h}_{\mathtt{v}\to\mathtt{p}},\, \mathtt{h}_{\mathtt{p}\to\mathtt{v}}\big)$
```

C Experiment Details

C.1 Benchmarks

Our experiments evaluate the vision-language reasoning abilities of multimodal large language models using a comprehensive suite of widely recognized benchmarks. For image understanding tasks, we assess performance on ten public benchmarks: GQA, MMBench (MMB) and MMBench-CN (MMB_{CN}), MME, POPE, VizWiz, ScienceQA (SQA), VQA^{V2}, TextVQA (VQA^{T}), and OCRBench (OCR). For video understanding tasks, we conduct experiments on four popular benchmarks: TGIF-QA (TGIF), MSVD-QA (MSVD), MSRVTT-QA (MSRV), and ActivityNet-QA (ActNet). The following section provides a concise overview of these benchmarks:

GQA [15] leverages scene graphs, questions, and images to evaluate visual scene understanding and reasoning. By incorporating detailed spatial relationships and object-level attributes, it poses significant challenges for models to perform accurate visual reasoning in complex environments.

MMBench [51] introduces a hierarchical evaluation framework where model capabilities are dissected into three levels. Level-1 focuses on basic perception and reasoning; Level-2 subdivides these abilities into six distinct sub-skills; and Level-3 further refines the evaluation into 20 specific dimensions. Its Chinese counterpart, **MMBench-CN**, adopts a similar structure.

MME [23] rigorously tests perceptual and cognitive abilities across 14 sub-tasks. By employing carefully crafted instruction-answer pairs and succinct instructions, MME minimizes data leakage and provides a robust, fair assessment of a model's multifaceted performance.

POPE [20] targets the evaluation of object hallucination by posing binary questions about object presence in images. It quantifies hallucination levels using metrics, *e.g.*, accuracy, recall, precision, and F1 score, offering a precise and focused measure of model reliability.

VizWiz [12] is a visual question answering benchmark derived from interactions with blind users. Comprising over $31,000$ image-question pairs with 10 human-annotated answers per query, it encapsulates the challenges of low-quality image capture and conversational spoken queries, thereby emphasizing real-world visual understanding.

ScienceQA [32] spans multiple scientific domains by organizing questions into 26 topics, 127 categories, and 379 skills. This hierarchical categorization provides a diverse and rigorous testbed for evaluating multimodal understanding, multi-step reasoning, and interpretability across natural, language, and social sciences.

VQA^{V2} [11] challenges models with open-ended questions based on $265,016$ images that depict a variety of real-world scenes. Each question is paired with 10 human-annotated answers, facilitating a thorough evaluation of a model's capacity to interpret and respond to diverse visual queries.

TextVQA [38] focuses on the integration of text within visual content. It evaluates a model's proficiency in reading and reasoning about textual information embedded in images, thereby requiring a balanced understanding of both visual and linguistic cues.

OCRBench [30] is a comprehensive benchmark for evaluating the OCR capabilities of multi-modal language models across five key tasks: text recognition, scene text-centric and document-oriented VQA, key information extraction, and handwritten mathematical expression recognition.

TGIF-QA [16] adapts the visual question answering task to the video domain by focusing on GIFs. With 165K question-answer pairs, it incorporates tasks, *e.g.*, counting repetitions, identifying repeating actions, detecting state transitions, and frame-specific question answering, thereby demanding detailed spatio-temporal analysis.

MSVD-QA [46] builds upon the MSVD dataset by pairing $1,970$ video clips with approximately 50.5K QA pairs. Questions are categorized into five distinct types, *e.g.*, what, who, how, when, and where, making it a versatile tool for evaluating video understanding.

MSRVTT-QA [6] features 10K video clips and 243K QA pairs designed to test the integration of visual and temporal information. Its structure, which parallels that of MSVD-QA through the inclusion of five question types, further enriches the evaluation landscape for video-based tasks.

ActivityNet-QA [50] provides 58K human-annotated question-answer pairs drawn from 5.8K videos. Its focus on questions related to motion, spatial relationships, and temporal dynamics necessitates long-term spatio-temporal reasoning, thus serving as a benchmark for advanced video understanding.

C.2 Multi-modal Large Language Models

We evaluate MoB using various open-source multimodal large language models (MLLMs). For image understanding tasks, experiments are conducted on the LLaVA series, including LLaVA-1.5-7B and LLaVA-Next-7B, as well as the Qwen-VL series, such as Qwen2-VL-7B. Specifically, LLaVA-Next and Qwen2-VL are utilized to validate performance on high-resolution images, *i.e.*, those with a larger number of visual tokens. For video understanding tasks, we employ Video-LLaVA-7B as the baseline model, following the settings reported in its original paper to ensure a fair comparison.

LLaVA-1.5-7B [25] is a robust vision-language model built on the LLaVA framework. It processes images resized to 224×224 and tokenizes them into roughly 572 visual tokens using a patch-based vision encoder. This design balances fine-grained visual representation with computational efficiency, making it effective for diverse multimodal tasks.

LLaVA-Next-7B [26] extends the LLaVA-1.5 by incorporating refined training strategies and data curation. It supports higher-resolution inputs (up to 448×448), yielding up to 2880 visual tokens. These enhancements improve its visual reasoning capabilities and enable more precise alignment between visual content and language but also incur significantly increased computational cost.

Qwen2-VL-7B [41] augments the Qwen2 language model with visual input capabilities. This model leverages cross-modal pretraining to seamlessly merge vision and language, demonstrating strong performance in complex visual question answering and comprehensive scene understanding.

Video-LLaVA-7B [24] extends the LLaVA framework into the temporal domain by processing video inputs. It is designed to capture both spatial and temporal dynamics, enabling effective video comprehension and video-based question answering with coherent and context-aware responses.

C.3 Baselines

To validate the superiority of the proposed MoB, we construct a robust baseline that integrates a comprehensive set of representative existing methods, which encompass single-stage methods with both two distinct objectives and several multi-stage methods.

ToMe [5] employs a lightweight token-matching scheme to merge visually similar tokens across transformer layers, thereby reducing computation without additional training. Its simple yet effective design makes it well suited for real-time applications.

FastV [7] leverages attention maps in the early layers to identify and prune non-critical tokens, significantly reducing initial computational overhead. This focus on early-stage reduction allows the model to operate more efficiently while maintaining performance.

SparseVLM [55] ranks tokens based on cross-modal attention to assess image-prompt relevance and adopts adaptive sparsity ratios to retain key information. It further incorporates a token recycling mechanism to balance the trade-off between efficiency and accuracy.

HiRED [1] allocates token budgets across image partitions by using CLS token attention and then selects the most informative tokens within each partition. This spatially aware approach ensures balanced reduction while preserving contextual details.

LLaVA-PruMerge [35] combines pruning and merging strategies by dynamically removing less important tokens using sparse CLS-visual attention. It then clusters the retained tokens based on key similarity, ensuring that crucial visual features remain intact.

PyramidDrop [45] adopts a progressive token-dropping strategy across different model stages, resulting in a pyramid-like token structure. This method carefully balances the reduction of tokens with the preservation of performance as the processing advances.

MustDrop [28] integrates several token-reduction strategies including spatial merging, text-guided pruning, and output-aware cache policies. Its multi-faceted approach efficiently reduces token counts across various stages of the model.

VisionZip [48] first selects dominant tokens that capture the majority of an image's information and then merges the remaining tokens based on semantic similarity. This approach dramatically reduce token redundancy while accelerating inference and maintaining robust performance.

FasterVLM [54] evaluates token importance using CLS attention in the encoder and prunes tokens before they interact with the language model. This preemptive reduction streamlines the overall process and enhances model efficiency.

GlobalCom^2 [29] employs a hierarchical strategy by coordinating thumbnail tokens to allocate adaptive retention ratios for high-resolution crops. This approach successfully preserves local details while providing effective global context reduction.

DART [43] leverages token duplication to guide its pruning process instead of relying solely on attention scores. By selecting a small set of pivot tokens and retaining only those with minimal redundancy, DART achieves significant acceleration in a training-free manner.

TokenCarve [39] implements a two-stage, training-free compression framework that preserves critical visual information during aggressive token reduction. It first prunes low-information tokens using an information-preservation guided selection and then merges the remaining tokens based on similarity to minimize accuracy loss.

TwigVLM [36] accelerates large vision-language models by appending a lightweight twig block to an early layer of a frozen base VLM. It utilizes twig-guided token pruning coupled with self-speculative decoding to boost generation speed while retaining high accuracy even under aggressive token reduction.

C.4 Implement Details

To ensure a fair comparison, we do not meticulously search the optimal hyperparameters of MoB (*i.e.*, the prompt covering cardinality K_p and the covering fold k) for each benchmark; besides, we apply the same configurations to all involved MLLMs. Specifically, for image understanding, we set

$$K_\mathrm{p} \in \left\{\tfrac{3K}{8}, \tfrac{K}{4}, \tfrac{11K}{24}\right\}, \quad k = \tfrac{3K_\mathrm{p}}{40}$$

under strong coupling, and

$$K_\mathrm{p} \in \left\{\tfrac{K}{2}, \tfrac{7K}{16}, \tfrac{5K}{12}\right\}, \quad k = \tfrac{K_\mathrm{p}}{8}$$

under weak coupling, corresponding to token reduction rates of $\{88.9\%, 77.8\%, 66.7\%\}$.

As for video understanding, we set $K_\mathrm{p} = \frac{3K}{8}$, $k = \frac{3K_\mathrm{p}}{40}$ for MSVD, MSRV, and ActNet; and set $K_\mathrm{p} = \frac{K}{2}$, $k = \frac{K_\mathrm{p}}{8}$ for TGIF. The pruning layer index is fixed at $\ell = 2$ for both image and video tasks. All baselines use their default settings.

To ensure reproducibility, we cross-validated our experimental results using the publicly available MLLMs evaluation tool *lmms-eval* (v0.3.0) [53, 4], with the random seed set to 1234. All experiments were conducted on $4\times$ Nvidia A800-80GB GPUs paired with $2\times$ Intel Xeon® Gold 6348 CPUs. The implementation was carried out in Python 3.10 using PyTorch 2.1.2 and CUDA 11.8.

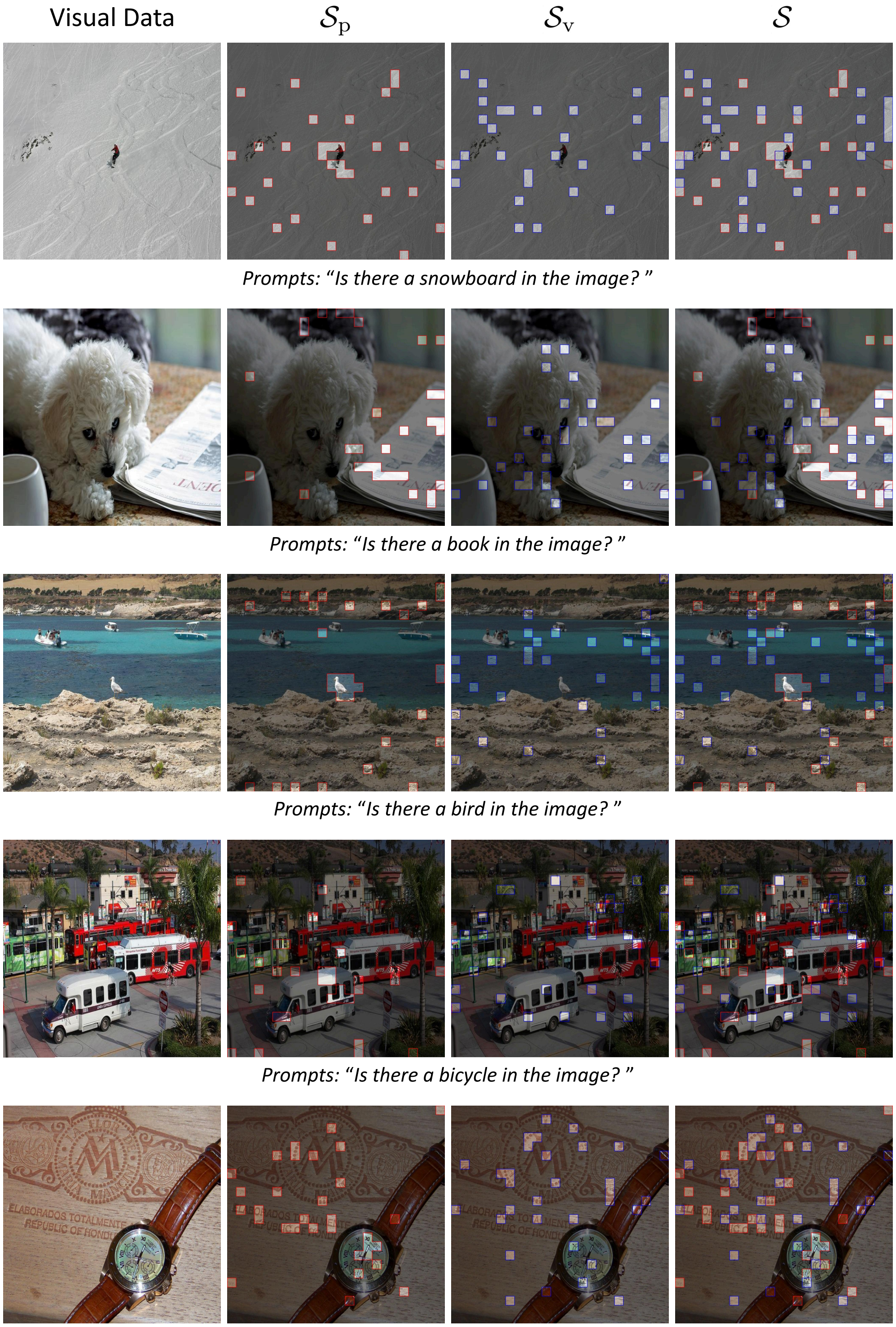
D Additional Experimental Results

D.1 Quantitative Comparison

Method	Objectives	Strong Coupling				Weak Coupling						Avg.
		MMB	MMB_{CN}	SQA	VizWiz	GQA	MME	POPE	VQA^{T}	VQA^{V2}	OCR	
LLaVA-1.5-7B		*w/o Pruning,* $N = 576$; *Token Reduction Rate* = **0.0%**										
Vanilla [25]	-	64.7	58.1	69.5	50.0	61.9	1862	85.9	58.2	78.5	297	100%
LLaVA-1.5-7B		*Pruning budget* $K = 192$; *Token Reduction Rate* = **66.7%**										
ToMe (ICLR'23) [5]	VP	60.5	-	65.2	-	54.3	1563	72.4	52.1	68.0	-	88.5%
FastV (ECCV'24) [7]	VP	61.2	57.0	67.3	50.8	52.7	1612	64.8	52.5	67.1	291	91.2%
HiRED (AAAI'25) [1]	VP	62.8	54.7	68.4	50.1	58.7	1737	82.8	47.4	74.9	190	91.5%
LLaVA-PruMerge (24.05) [35]	VP	59.6	52.9	67.9	50.1	54.3	1632	71.3	54.3	70.6	253	90.8%
SparseVLM (ICML'25) [55]	PA	62.5	53.7	69.1	50.5	57.6	1721	83.6	56.1	75.6	292	96.3%
PyramidDrop (CVPR'25) [45]	PA	63.3	56.8	68.8	51.1	57.1	1797	82.3	56.1	75.1	290	96.7%
FiCoCo-V (EMNLP'24) [52]	VP	62.3	55.3	67.8	51.0	58.5	1732	82.5	55.7	74.4	-	96.1%
MustDrop (24.11) [28]	PA VP	62.3	55.8	69.2	51.4	58.2	1787	82.6	56.5	76.0	289	97.2%
VisionZip (24.12) [48]	VP	63.0	-	68.9	-	59.3	1783	85.3	57.3	76.8	-	97.7%
DART (25.02) [43]	VP	63.6	57.0	69.8	51.2	60.0	1856	82.8	57.4	76.7	296	98.8%
TokenCarve (25.03) [39]	PA VP	63.0	-	69.1	50.9	-	1830	84.9	58.4	78.0	-	99.3%
TwigVLM (25.03) [36]	PA	64.0	-	68.8	-	61.2	1848	87.2	58.0	78.1	-	99.5%
MoB (Ours)	PA VP	64.1	57.8	70.1	52.5	61.4	1860	84.8	58.5	78.3	307	100.6%
LLaVA-1.5-7B		*Pruning budget* $K = 128$; *Token Reduction Rate* = **77.8%**										
ToMe (ICLR'23)	VP	53.3	-	59.6	-	52.4	1343	62.8	49.1	63.0	-	80.4%
FastV (ECCV'24)	VP	56.1	56.4	60.2	51.3	49.6	1490	59.6	50.6	61.8	285	86.4%
HiRED (AAAI'25)	VP	61.5	53.6	68.1	51.3	57.2	1710	79.8	46.1	73.4	191	90.2%
LLaVA-PruMerge (24.05)	VP	58.1	51.7	67.1	50.3	53.3	1554	67.2	54.3	68.8	248	88.8%
SparseVLM (ICML'25)	PA	60.0	51.1	67.1	51.4	56.0	1696	80.5	54.9	73.8	280	93.8%
PyramidDrop (CVPR'25)	PA	61.6	56.6	68.3	51.0	56.0	1761	82.3	55.1	72.9	287	95.1%
FiCoCo-V (EMNLP'24)	VP	61.1	54.3	68.3	49.4	57.6	1711	82.2	55.6	73.1	-	94.9%
MustDrop (24.11)	PA VP	61.1	55.2	68.5	52.1	56.9	1745	78.7	56.3	74.6	281	95.6%
VisionZip (24.12)	VP	62.0	-	68.9	-	57.6	1762	83.2	56.8	75.6	-	96.2%
DART (25.02)	VP	63.2	57.5	69.1	51.7	58.7	1840	80.1	56.4	75.9	296	98.0%
TokenCarve (25.03)	PA VP	62.7	-	68.9	51.0	-	1829	84.5	58.1	77.3	-	99.0%
TwigVLM (25.03)	PA	63.5	-	69.5	-	60.6	1818	86.6	57.8	77.9	-	99.0%
MoB (Our)	PA VP	63.5	57.5	69.6	52.7	60.9	1845	82.1	57.8	77.5	299	99.4%
LLaVA-1.5-7B		*Pruning budget* $K = 64$; *Token Reduction Rate* = **88.9%**										
ToMe (ICLR'23)	VP	43.7	-	50.0	-	48.6	1138	52.5	45.3	57.1	-	70.1%
FastV (ECCV'24)	VP	48.0	52.7	51.1	50.8	46.1	1256	48.0	47.8	55.0	245	77.3%
HiRED (AAAI'25)	VP	60.2	51.4	68.2	50.2	54.6	1599	73.6	44.2	69.7	191	87.0%
LLaVA-PruMerge (24.05)	VP	55.3	49.1	68.1	50.1	51.9	1549	65.3	54.0	67.4	250	87.4%
SparseVLM (ICML'25)	PA	56.2	46.1	62.2	50.1	52.7	1505	75.1	51.8	68.2	180	84.6%
PyramidDrop (CVPR'25)	PA	58.8	50.5	68.6	50.7	41.9	1561	55.9	45.9	69.2	250	78.1%
FiCoCo-V (EMNLP'24)	VP	60.3	53.0	68.1	49.8	52.4	1591	76.0	53.6	71.3	-	91.5%
MustDrop (24.11)	PA VP	60.0	53.1	63.4	51.2	53.1	1612	68.0	54.2	69.3	267	90.1%
VisionZip (24.12)	VP	60.1	-	69.0	-	55.1	1690	77.0	55.5	72.4	-	92.8%
DART (25.02)	VP	60.6	53.2	69.8	51.6	55.9	1765	73.9	54.4	72.4	270	93.7%
TokenCarve (25.03)	PA VP	62.0	-	69.7	51.4	-	1754	79.9	57.0	74.8	-	97.0%
TwigVLM (25.03)	PA	60.4	-	70.0	-	58.8	1760	82.7	55.8	75.6	-	96.1%
MoB (Our)	PA VP	62.1	54.5	69.8	52.1	59.0	1806	77.2	57.0	75.5	277	96.4%
LLaVA-Next-7B		*w/o Pruning,* $N = 2880$; *Token Reduction Rate* = **0.0%**										
Vanilla [26]	-	67.4	60.6	70.1	57.6	64.2	1851	86.5	64.9	81.8	517	100%
LLaVA-Next-7B		*Pruning budget* $K = 320$; *Token Reduction Rate* = **88.9%**										
FastV (ECCV'24)	VP	61.6	51.9	62.8	53.1	55.9	1661	71.7	55.7	71.9	374	86.4%
HiRED (AAAI'25)	VP	64.2	55.9	66.7	54.2	59.3	1690	83.3	58.8	75.7	404	91.8%
LLaVA-PruMerge (24.05)	VP	61.3	55.3	66.4	54.0	53.6	1534	60.8	50.6	69.7	146	79.9%
SparseVLM (ICML'25)	PA	60.6	54.5	66.1	52.0	56.1	1533	82.4	58.4	71.5	270	85.9%
PyramidDrop (CVPR'25)	PA	63.4	56.2	67.5	54.1	56.4	1663	77.6	54.4	73.5	259	86.8%
MustDrop (24.11)	PA VP	62.8	55.1	68.0	54.0	57.3	1641	82.1	59.9	73.7	382	90.4%
VisionZip (24.12)	VP	63.1	-	67.3	-	59.3	1702	-	58.9	76.2	-	93.0%
FasterVLM (24.12) [54]	VP	61.6	53.5	66.5	52.6	56.9	1701	83.6	56.5	74.0	401	89.8%
GlobalCom^2(25.01) [29]	VP	61.8	53.4	67.4	54.6	57.1	1698	83.8	57.2	76.7	375	90.3%
DART (25.02)	VP	65.3	58.2	68.4	56.1	61.7	1710	84.1	58.7	79.1	406	93.9%
TwigVLM (25.03)	PA	65.0	-	68.7	-	62.2	1758	-	57.4	79.7	-	95.4%
MoB (Our)	PA VP	65.8	58.9	68.7	57.0	62.6	1760	84.4	60.2	80.1	418	95.4%

Table 4: Full results on image understanding with the LLaVA-7B Series, where $K_{\text{p}} \in \{\frac{3K}{8}, \frac{K}{4}, \frac{K}{4}\}$, $k = \frac{3K_{\text{p}}}{40}$ for strong-coupling and $K_{\text{p}} \in \{\frac{K}{2}, \frac{7K}{16}, \frac{5K}{12}\}$, $k = \frac{K_{\text{p}}}{8}$ for weak-coupling benchmarks, corresponding to token reduction rates in $\{88.9\%, 77.8\%, 66.7\%\}$; the pruning layer index $\ell = 2$.

D.2 Visualization



Prompts: "Is there a snowboard in the image? "

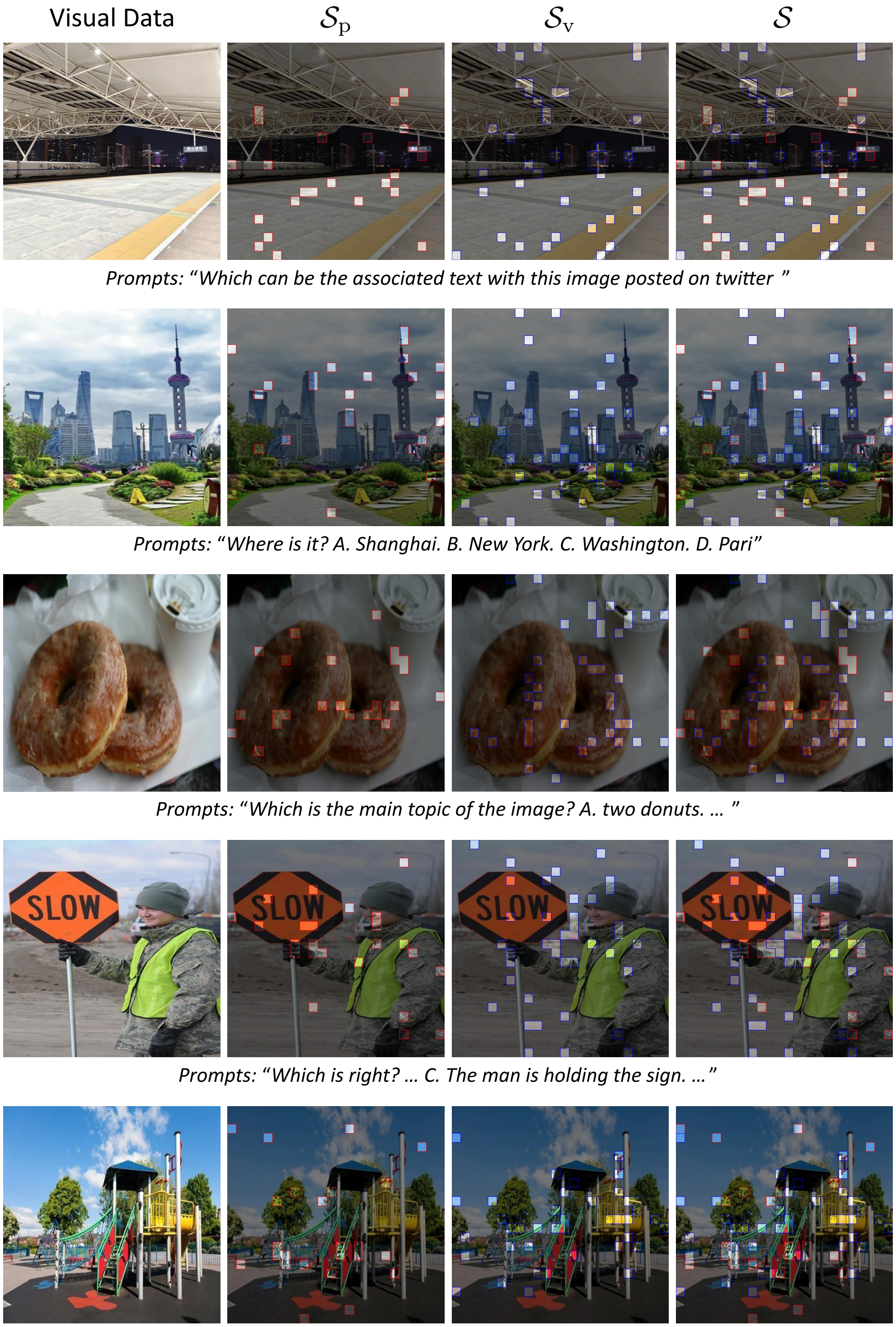
Prompts: "Is there a book in the image? "

Prompts: "Is there a bird in the image? "

Prompts: "Is there a bicycle in the image? "

Prompts: "What brand of watch is this? "

Figure 7: Visualization of the selected prompt and visual centers under weak coupling.



Prompts: "Which can be the associated text with this image posted on twitter "

Prompts: "Where is it? A. Shanghai. B. New York. C. Washington. D. Pari"

Prompts: "Which is the main topic of the image? A. two donuts. … "

Prompts: "Which is right? … C. The man is holding the sign. …"

Prompts: "What type of environment is depicted in the picture? … B. Children's playground. …"

Figure 8: Visualization of the selected prompt and visual centers under strong coupling.

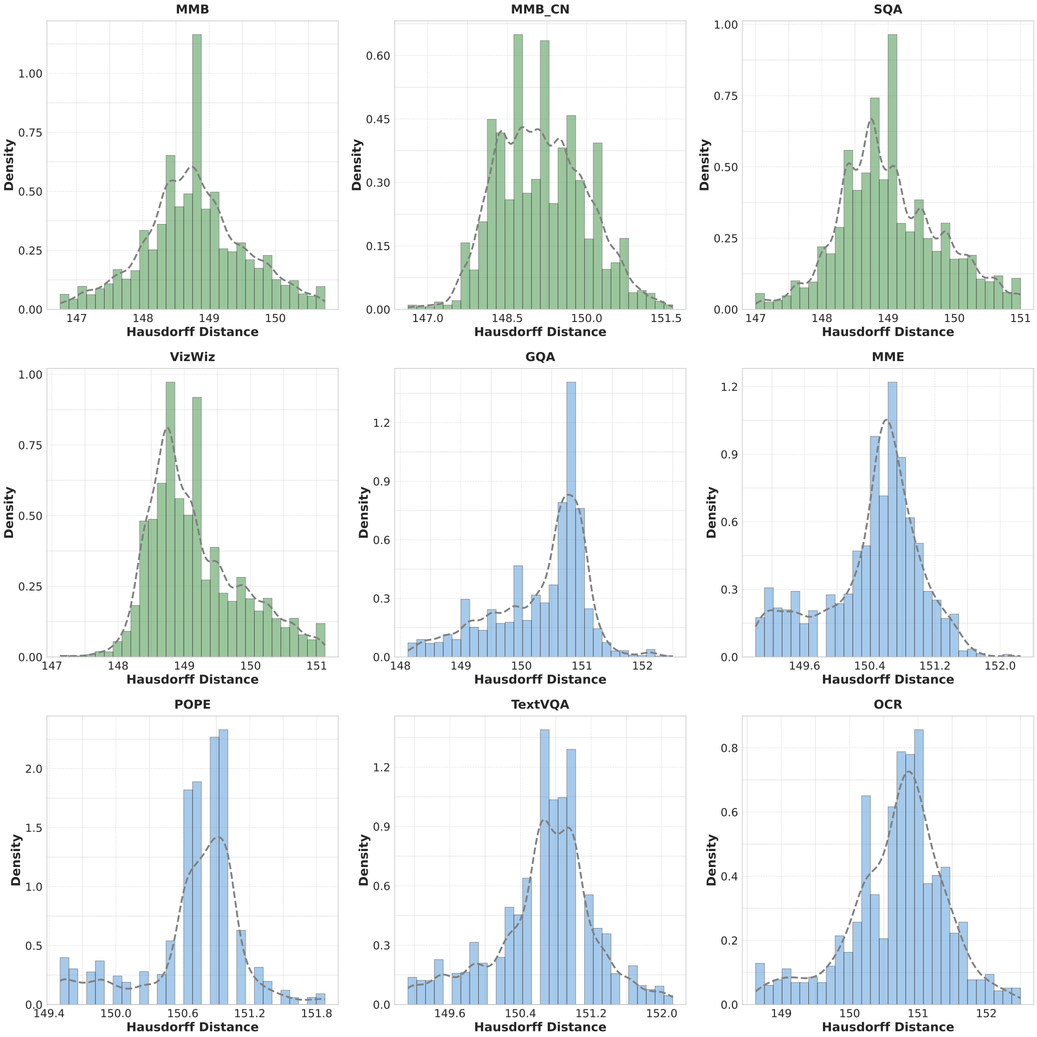


Figure 9: Observations of prompt-visual coupling η across 9 popular benchmarks.

E Omitted Technical Details

E.1 Proof of Lemma 1

Restatement of Lemma 1 (An Error Bound for Visual Token Pruning)**.** *Under Assumption 1, given any token set with its pruned counterpart $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P},\ \ \mathcal{X}_{\mathrm{s}} = \mathcal{S} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$, the pruning error bound is given by:*

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_{\mathrm{s}})\| \ \le\ C_\ell\ \max\Big\{\min\big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{V},\mathcal{P})\big\},\, \min\big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{S},\mathcal{P})\big\}\Big\}.$$

Remark. *Here $d_H(\mathcal{S},\mathcal{P})$ and $d_H(\mathcal{S},\mathcal{V})$ describe the prompt alignment and visual preservation, while $d_H(\mathcal{V},\mathcal{P})$ is an inherent term that describes the prompt-visual coupling of input data.*

Proof. The intermediate input for any layer and its pruned counterpart are given by

$$\mathcal{X} = \mathcal{V} \sqcup \mathcal{P} \text{ and } \mathcal{X}_{\mathrm{s}} = \mathcal{S} \sqcup \mathcal{P}.$$

By Equation (1), the Hausdorff distance is symmetric, *i.e.*,

$$d_H(\mathcal{S},\mathcal{V}) = d_H(\mathcal{V},\mathcal{S}), \tag{E1-1}$$

and induced by Euclidean distance.

Step 1. Bound the one-sided distances.

We analyze the distances by considering the membership of the points in the subsets.

Direction 1 ($\mathcal{X} \to \mathcal{X}_{\mathrm{s}}$) For any $x \in \mathcal{X}$:

Case (i): If $x \in \mathcal{P}$, then since $\mathcal{P} \subset \mathcal{X}_{\mathrm{s}}$,

$$\inf_{y\in\mathcal{X}_{\mathrm{s}}} \|x-y\| = 0.$$

Case (ii): If $x \in \mathcal{V}$, then the candidate points in $\mathcal{X}_{\mathrm{s}} = \mathcal{S} \sqcup \mathcal{P}$ can be chosen either from $\mathcal{S}$ or $\mathcal{P}$. Thus,

$$\inf_{y\in\mathcal{X}_{\mathrm{s}}} \|x-y\| \ \le\ \min\Big\{\inf_{s\in\mathcal{S}} \|x-s\|,\ \inf_{p\in\mathcal{P}} \|x-p\|\Big\}.$$

Taking the supremum over $x \in \mathcal{V}$ yields

$$\sup_{x\in\mathcal{V}} \inf_{y\in\mathcal{X}_{\mathrm{s}}} \|x-y\| \ \le\ \min\Big\{\sup_{x\in\mathcal{V}} \inf_{s\in\mathcal{S}} \|x-s\|,\ \sup_{x\in\mathcal{V}} \inf_{p\in\mathcal{P}} \|x-p\|\Big\}.$$

$$\sup_{x\in\mathcal{V}} \inf_{p\in\mathcal{P}} \|x-p\| \ \le\ \max\Big\{\sup_{x\in\mathcal{V}} \inf_{p\in\mathcal{P}} \|x-p\|,\ \sup_{p\in\mathcal{P}} \inf_{x\in\mathcal{V}} \|p-x\|\Big\} = d_H(\mathcal{V},\mathcal{P}),$$

By Equation (1), we derive the distance in direction 1:

$$\sup_{x\in\mathcal{X}} \inf_{y\in\mathcal{X}_{\mathrm{s}}} \|x-y\| \ \le\ \min\Big\{d_H(\mathcal{V},\mathcal{S}),\ d_H(\mathcal{V},\mathcal{P})\Big\}. \tag{E1-2}$$

Direction 2 ($\mathcal{X}_{\mathrm{s}} \to \mathcal{X}$) For any $y \in \mathcal{X}_{\mathrm{s}}$:

Case (i): If $y \in \mathcal{P}$, then as $\mathcal{P} \subset \mathcal{X}$,

$$\inf_{x\in\mathcal{X}} \|y-x\| = 0.$$

Case (ii): If $y \in \mathcal{S}$, the candidate points in $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P}$ can be chosen from either $\mathcal{V}$ or $\mathcal{P}$; hence

$$\inf_{x\in\mathcal{X}} \|y-x\| \ \le\ \min\Big\{\inf_{v\in\mathcal{V}} \|y-v\|,\ \inf_{p\in\mathcal{P}} \|y-p\|\Big\}.$$

Taking the supremum over $y \in \mathcal{S}$ yields

$$\sup_{y\in\mathcal{S}} \inf_{x\in\mathcal{X}} \|y-x\| \ \le\ \min\Big\{\sup_{y\in\mathcal{S}} \inf_{v\in\mathcal{V}} \|y-v\|,\ \sup_{y\in\mathcal{S}} \inf_{p\in\mathcal{P}} \|y-p\|\Big\}.$$

$$\sup_{y\in\mathcal{S}}\inf_{p\in\mathcal{P}}\|y-p\| \ \le\ \max\Big\{\sup_{y\in\mathcal{S}}\inf_{p\in\mathcal{P}}\|y-p\|,\ \sup_{p\in\mathcal{P}}\inf_{y\in\mathcal{S}}\|p-y\|\Big\} = d_H(\mathcal{S},\mathcal{P}),$$

By Equation (1), we derive the distance in direction 2:

$$\sup_{y\in\mathcal{X}_\mathrm{s}}\inf_{x\in\mathcal{X}}\|y-x\| \ \le\ \min\Big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{S},\mathcal{P})\Big\}. \tag{E1-3}$$

Step 2. Combine the bounds.

By Equation (1), combining the bounds in (E1-2) and (E1-3), we obtain

$$d_H\big(\mathcal{X},\mathcal{X}_\mathrm{s}\big) \le \max\Big\{\ \min\big\{d_H(\mathcal{V},\mathcal{S}),\ d_H(\mathcal{V},\mathcal{P})\big\},\ \min\big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{S},\mathcal{P})\big\}\Big\}.$$

Based on (E1-1), we have

$$d_H\big(\mathcal{X},\mathcal{X}_\mathrm{s}\big) \le \max\Big\{\ \min\big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{V},\mathcal{P})\big\},\ \min\big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{S},\mathcal{P})\big\}\Big\}.$$

Loading the Assumption 1, we have the output discrepancy is bounded by

$$\begin{aligned}\|\mathcal{F}(\mathcal{X})-\mathcal{F}(\mathcal{X}_\mathrm{s})\| \ &\le\ C_\ell\ d_H\big(\mathcal{X},\mathcal{X}_\mathrm{s}\big).\\ &= C_\ell\ \max\Big\{\ \min\big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{V},\mathcal{P})\big\},\ \min\big\{d_H(\mathcal{S},\mathcal{V}),\ d_H(\mathcal{S},\mathcal{P})\big\}\Big\}.\end{aligned}$$

This completes the proof. □

E.2 Proof of Lemma 2

Restatement of Lemma 2 (A Relaxed Error Bound under Practical Budgets) **.** *Under Assumptions 1 and 2, let $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P}, \;\; \mathcal{X}_{\mathrm{s}} = \mathcal{S} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$ with $|\mathcal{S}| = K \ll N$. Partition the retained token set $\mathcal{S}$ into two disjoint subsets: $\mathcal{S} = \mathcal{S}_{\mathrm{p}} \sqcup \mathcal{S}_{\mathrm{v}}$, devoted to prompt alignment $d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P})$ and visual preservation $d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V})$, respectively. Then, the pruning error bound reduces to*

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_{\mathrm{s}})\| \;\le\; C_\ell \;\max\bigl\{d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}),\; d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V})\bigr\} + C_\ell\,\eta.$$

Proof. By Lemma 1, we obtain

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_{\mathrm{s}})\| \;\le\; C_\ell \;\max\Bigl\{\min\bigl\{d_H(\mathcal{S}, \mathcal{V}),\; d_H(\mathcal{V}, \mathcal{P})\bigr\}, \min\bigl\{d_H(\mathcal{S}, \mathcal{V}),\; d_H(\mathcal{S}, \mathcal{P})\bigr\}\Bigr\}.$$

Since $\min\{a, b\} \le \max\{a, b\}$, we have

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_{\mathrm{s}})\| \;\le\; C_\ell \;\max\Bigl\{\; d_H\bigl(\mathcal{S}, \mathcal{P}\bigr),\; d_H\bigl(\mathcal{S}, \mathcal{V}\bigr),\; d_H\bigl(\mathcal{V}, \mathcal{P}\bigr)\Bigr\}. \tag{E2-1}$$

For any $p \in \mathcal{P}$, we have

$$\inf_{s \in \mathcal{S}} \|p - s\| = \min\Bigl\{\; \inf_{s \in \mathcal{S}_{\mathrm{p}}} \|p - s\|,\; \inf_{s \in \mathcal{S}_{\mathrm{v}}} \|p - s\|\Bigr\} \le \inf_{s \in \mathcal{S}_{\mathrm{p}}} \|p - s\|.$$

Taking the supremum over $p \in \mathcal{P}$ yields

$$\sup_{p \in \mathcal{P}} \inf_{s \in \mathcal{S}} \|p - s\| \le \sup_{p \in \mathcal{P}} \inf_{s \in \mathcal{S}_{\mathrm{p}}} \|p - s\|.$$

Similarly, since $\mathcal{S}_{\mathrm{v}} \subset \mathcal{S}$,

$$\sup_{s \in \mathcal{S}_{\mathrm{v}}} \inf_{p \in \mathcal{P}} \|s - p\| \le \sup_{s \in \mathcal{S}} \inf_{p \in \mathcal{P}} \|s - p\|.$$

Thus, by Equation (1),

$$d_H(\mathcal{S}, \mathcal{P}) \le \max\Bigl\{d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}),\; d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{P})\Bigr\}.$$

Using Assumption 2 ($d_H(\mathcal{V}, \mathcal{P}) \le \eta$) and the triangle inequality for Hausdorff distance, we have

$$\begin{aligned} d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{P}) &\le d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) + d_H(\mathcal{V}, \mathcal{P}) \le d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) + \eta, \\ d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{V}) &\le d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}) + d_H(\mathcal{P}, \mathcal{V}) \le d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}) + \eta. \end{aligned}$$

Hence,

$$d_H(\mathcal{S}, \mathcal{P}) \le \max\Bigl\{d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}),\; d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) + \eta\Bigr\}. \tag{E2-2}$$

Similarly, one can show that

$$d_H(\mathcal{S}, \mathcal{V}) \le \max\Bigl\{d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}),\; d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}) + \eta\Bigr\}. \tag{E2-3}$$

Loading the maximum of (E2-2), (E2-3) and $d_H(\mathcal{V}, \mathcal{P})$ into (E2-1), we obtain

$$\begin{aligned} \|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_{\mathrm{s}})\| &\le C_\ell \;\max\Bigl\{d_H(\mathcal{S}, \mathcal{P}),\, d_H(\mathcal{S}, \mathcal{V}),\; d_H(\mathcal{V}, \mathcal{P})\Bigr\} \\ &\le C_\ell \;\max\Bigl\{d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}),\; d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) + \eta,\;\; d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}),\; d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}) + \eta,\;\; \eta\Bigr\} \end{aligned}$$

Since $d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}) \ge 0$, $d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) \ge 0$, $\eta \ge 0$, we have

$$\begin{aligned} \max\bigl\{d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}),\; d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}) + \eta,\; \eta\bigr\} &= d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}) + \eta, \\ \max\bigl\{d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}),\; d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) + \eta,\; \eta\bigr\} &= d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) + \eta. \end{aligned}$$

Hence

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathcal{X}_{\mathrm{s}})\| \;\le C_\ell \;\max\Bigl\{d_H(\mathcal{S}_{\mathrm{p}}, \mathcal{P}),\; d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V})\Bigr\} + C_\ell\,\eta.$$

This completes the proof. □

E.3 Proof of Lemma 3

Restatement of Lemma 3 (d_{eff}-regular lower bound on covering numbers)**.** *Given $\mathcal{P}, \mathcal{V} \subset \mathbb{R}^d$ with an effective dimension d_{eff}. Suppose their δ-dilations $\mathcal{V}_\delta := \bigcup_{v \in \mathcal{V}} B(v, \delta)$, $\mathcal{P}_\delta := \bigcup_{p \in \mathcal{P}} B(p, \delta)$ ($\delta \ll \eta$) satisfy d_{eff}-dimensional covering regularity; thus, there exist constants $b > a > 0$, $b' > a' > 0$ and $\epsilon_0 > \delta$ such that*

$$a\,\epsilon_{\text{p}}^{-d_{\text{eff}}} \le \mathcal{N}(\mathcal{P}, \epsilon_{\text{p}}) \le b\,\epsilon_{\text{p}}^{-d_{\text{eff}}}, \qquad a'\,\epsilon_{\text{v}}^{-d_{\text{eff}}} \le \mathcal{N}(\mathcal{V}, \epsilon_{\text{v}}) \le b'\,\epsilon_{\text{v}}^{-d_{\text{eff}}}, \qquad \forall\, \epsilon_{\text{p}}, \epsilon_{\text{v}} \in (\delta, \epsilon_0],$$

Remark *Previous work suggests that both visual and language embeddings concentrate on a low-dimensional manifold, so the effective covering dimension satisfies the typical relation $d_{\text{eff}} \ll d$.*

Proof. We prove the two-sided bound for $\mathcal{P}$; the argument for $\mathcal{V}$ is identical.

Notation.

- $\mathcal{N}(X, r)$: minimal number of closed balls of radius r covering X.
- $X_\delta = \bigcup_{x \in X} B(x, \delta)$, with $B(x, \delta) = \{y : \|y - x\| \le \delta\}$.

Step 1. Transfer trick for small ϵ.

Fix $\epsilon \in (\delta, \epsilon_0]$ and define $\epsilon' = \min\{\epsilon + \delta,\ \epsilon_0\}$.

If $\epsilon \le \epsilon_0 - \delta$ (so $\epsilon' = \epsilon + \delta$), then any ϵ-cover $\{z_i\}_{i=1}^m$ of $\mathcal{P}$ satisfies for each $y \in \mathcal{P}_\delta$:

$$\exists\, x \in \mathcal{P} : \|y - x\| \le \delta, \quad \exists\, i : \|x - z_i\| \le \epsilon \implies \|y - z_i\| \le \epsilon + \delta = \epsilon'.$$

Hence

$$\mathcal{P}_\delta \ \subseteq\ \bigcup_{i=1}^m B(z_i, \epsilon') \quad \Longrightarrow \quad \mathcal{N}(\mathcal{P}_\delta, \epsilon') \ \le\ \mathcal{N}(\mathcal{P}, \epsilon). \tag{E3-1}$$

Note: For $\epsilon > \epsilon_0 - \delta$, the above transfer argument is not applied.

Step 2. Lower bound on $\mathcal{N}(\mathcal{P}, \epsilon)$.

Split into two cases:

Case I: $\epsilon \le \epsilon_0 - \delta$**.** Since $\mathcal{P}_\delta$ satisfies d_{eff}-dimensional covering regularity; loading the lower-bound for $\mathcal{P}_\delta$ at radius $\epsilon' = \epsilon + \delta$, there exists a constant $a_\delta \ge 0$ such that

$$\mathcal{N}(\mathcal{P}_\delta, \epsilon') = \mathcal{N}(\mathcal{P}_\delta, \epsilon + \delta) \ \ge\ a_\delta\,(\epsilon + \delta)^{-d_{\text{eff}}}.$$

Based on (E3-1), we obtain

$$a_\delta(\epsilon + \delta)^{-d_{\text{eff}}} \ \le\ \mathcal{N}(\mathcal{P}_\delta, \epsilon') \ \le\ \mathcal{N}(\mathcal{P}, \epsilon)$$

Since $\delta \le \epsilon$, it follows that $\epsilon + \delta \le 2\epsilon$; thus, we have

$$\mathcal{N}(\mathcal{P}, \epsilon) \ \ge\ a_\delta\, 2^{-d_{\text{eff}}}\, \epsilon^{-d_{\text{eff}}}. \tag{E3-2}$$

Case II: $\epsilon > \epsilon_0 - \delta$**.** Define $\widetilde{a} \ := \ (\epsilon_0 - \delta)^{d_{\text{eff}}}$, such that

$$(\epsilon_0 - \delta)^{-d_{\text{eff}}} = \widetilde{a}^{-1}.$$

Since $\epsilon > \epsilon_0 - \delta$, we have

$$\epsilon^{-d_{\text{eff}}} \ \le\ (\epsilon_0 - \delta)^{-d_{\text{eff}}}.$$

Hence

$$\epsilon^{-d_{\text{eff}}} \ \le\ \widetilde{a}^{-1} \quad \Longleftrightarrow \quad \widetilde{a}\,\epsilon^{-d_{\text{eff}}} \ \le\ 1.$$

Since any nonempty set $\mathcal{P}$ has covering number at least one, the following holds

$$\widetilde{a}\,\epsilon^{-d_{\text{eff}}} \ \le\ 1 \ \le\ \mathcal{N}(\mathcal{P}, \epsilon). \tag{E3-3}$$

Therefore, set $a := \min\{a_\delta 2^{-d_{\text{eff}}}, \widetilde{a}\} > 0$, combining (E3-2) and (E3-3) yields

$$\mathcal{N}(\mathcal{P}, \epsilon) \ \geq \ a\, \epsilon^{-d_{\text{eff}}}, \quad \forall\, \epsilon \in (\delta, \epsilon_0]. \tag{E3-4}$$

Similarly, $\mathcal{V}$ holds $\mathcal{N}(\mathcal{V}, \epsilon) \geq a'\, \epsilon^{-d_{\text{eff}}}, \quad \forall\, \epsilon \in (\delta, \epsilon_0]$.

Step 3. Upper bound on $\mathcal{N}(\mathcal{P}, \epsilon)$.

Since $\mathcal{P}_\delta$ satisfies d_{eff}-dimensional covering regularity, there exists a constant $b_\delta \geq a_\delta \geq 0$ such that

$$\mathcal{N}(\mathcal{P}_\delta, \epsilon) \ \leq \ b_\delta\, \epsilon^{-d_{\text{eff}}}.$$

Since $\mathcal{P} \subseteq \mathcal{P}_\delta$, we have $\mathcal{N}(\mathcal{P}, \epsilon) \leq \mathcal{N}(\mathcal{P}_\delta, \epsilon)$; thus, the following holds

$$\mathcal{N}(\mathcal{P}, \epsilon) \leq \mathcal{N}(\mathcal{P}_\delta, \epsilon) \leq \ b_\delta\, \epsilon^{-d_{\text{eff}}}.$$

Based on the *monotonicity of covering numbers*, for every radius $\epsilon \geq \delta$, we have

$$\mathcal{N}(\mathcal{P}, \epsilon) \leq \mathcal{N}(\mathcal{P}, \delta).$$

Therefore, set $b \ := \ \max\{b_\delta, \ \mathcal{N}(\mathcal{P}, \delta)\}$, for all $\epsilon \in (\delta, \epsilon_0]$ we have

$$\mathcal{N}(\mathcal{P}, \epsilon) \ \leq \ b\, \epsilon^{-d_{\text{eff}}}. \tag{E3-5}$$

Likewise for $\mathcal{V}$, the following holds $\mathcal{N}(\mathcal{V}, \epsilon) \leq b'\, \epsilon^{-d_{\text{eff}}}, \quad \forall\, \epsilon \in (\delta, \epsilon_0]$.

Step 4. Combine the bounds.

Based on (E3-4) and (E3-5), for all $\epsilon \in (\delta, \epsilon_0]$ the following holds

$$a\, \epsilon^{-d_{\text{eff}}} \leq \mathcal{N}(\mathcal{P}, \epsilon) \leq b\, \epsilon^{-d_{\text{eff}}}, \quad a'\, \epsilon^{-d_{\text{eff}}} \leq \mathcal{N}(\mathcal{V}, \epsilon) \leq b'\, \epsilon^{-d_{\text{eff}}}.$$

This completes the proof. □

E.4 Proof of Theorem 1

Restatement of Theorem 1 (Trade-off between Prompt Alignment and Visual Preservation)**.** *Under Assumption 2 and the covering-regularity hypothesis of Lemma 3 with constants $a, a', d_{\text{eff}} > 0$, there exist a radius-scaling factor $z > 1$ such that $\eta/z > \delta$ and $K < \mathcal{N}(\mathcal{P}, \eta/z) + \mathcal{N}(\mathcal{V}, \eta/z)$, for every pruning results $\mathcal{S} = (\mathcal{S}_{\text{p}} \sqcup \mathcal{S}_{\text{v}}) \subseteq \mathcal{V}$ with budget K satisfying*

$$\max\{D_1 K^{-2/d_{\text{eff}}},\ D_2\, \eta^2\} \ \le\ d_H(\mathcal{S}_{\text{p}}, \mathcal{P})\, d_H(\mathcal{S}_{\text{v}}, \mathcal{V}),$$

where $D_1 := (a\, a')^{1/d_{\text{eff}}}\, 4^{1/d_{\text{eff}}} > 0$, $D_2 := 1/z^2 > 0$.

Remark (Optimal Attainment Level)**.** *The term $D_1\, K^{-2/d_{\text{eff}}}$ is completely determined by the pruning budget, while $D_2\, \eta^2$ quantifies the effect of prompt-visual coupling. Hence, the optimal attainment level per objective is given by $\epsilon^* = \max\{\eta/z,\ \sqrt{D_1}\, K^{-1/d_{\text{eff}}}\}$. Any attempt to reduce one objective below ϵ^* forces the other above ϵ^*, thereby increasing the overall pruning error.*

Remark (Effect of Budget and Coupling Strength)**.** *As K decreases, z correspondingly shrinks (D_2 growing as a power function), ultimately making $D_2\, \eta^2$ dominate the bound; while as K increases, both of the terms reduce, thereby diminishing the trade-off and tightening the overall error bound.*

Proof. We begin the proof by noting

$$\epsilon_{\text{p}} = d_H(\mathcal{S}_{\text{p}}, \mathcal{P}), \quad \epsilon_{\text{v}} = d_H(\mathcal{S}_{\text{v}}, \mathcal{V}), \quad K_{\text{p}} = |\mathcal{S}_{\text{p}}|, \quad K_{\text{v}} = |\mathcal{S}_{\text{v}}|, \quad K_{\text{p}} + K_{\text{v}} = K.$$

Step 1. Quantify the impact of budget K.

By Lemma 3, for all $\epsilon_{\text{p}}, \epsilon_{\text{v}} \in (\delta, \epsilon_0]$, we have

$$a\, \epsilon_{\text{p}}^{-d_{\text{eff}}} \le \mathcal{N}(\mathcal{P}, \epsilon_{\text{p}}) \le K_{\text{p}}, \quad a'\, \epsilon_{\text{v}}^{-d_{\text{eff}}} \le \mathcal{N}(\mathcal{V}, \epsilon_{\text{v}}) \le K_{\text{v}}. \tag{E4-1}$$

By AM-GM inequality, we have $K_{\text{p}} K_{\text{v}} \le \left(\frac{K}{2}\right)^2$; thus, loading (E4-1) we have

$$(a\, a')\, (\epsilon_{\text{p}}\, \epsilon_{\text{v}})^{-d_{\text{eff}}} \ \le\ \left(\frac{K}{2}\right)^2 \implies \epsilon_{\text{p}}\, \epsilon_{\text{v}} \ \ge\ (a\, a')^{1/d_{\text{eff}}}\, 4^{1/d_{\text{eff}}}\, K^{-2/d_{\text{eff}}}.$$

Define $D_1 := (a\, a')^{1/d_{\text{eff}}}\, 4^{1/d_{\text{eff}}} > 0$, the K-bound is established by

$$\epsilon_{\text{p}}\, \epsilon_{\text{v}} \ \ge\ D_1\, K^{-2/d_{\text{eff}}}. \tag{E4-2}$$

Step 2. Quantify the impact of prompt-visual coupling η.

Based on the budget condition, the radius-scaling factor z holds

$$K < \mathcal{N}\big(\mathcal{P}, \tfrac{\eta}{z}\big) + \mathcal{N}\big(\mathcal{V}, \tfrac{\eta}{z}\big). \tag{E4-3}$$

For contradiction, we suppose two covering radii is simultaneously small, such that $\epsilon_{\text{p}} < \eta/z$ and $\epsilon_{\text{v}} < \eta/z$. Then, the monotonicity of covering numbers gives

$$\mathcal{N}(\mathcal{P}, \epsilon_{\text{p}}) \ge \mathcal{N}\big(\mathcal{P}, \tfrac{\eta}{z}\big), \quad \mathcal{N}(\mathcal{V}, \epsilon_{\text{v}}) \ge \mathcal{N}\big(\mathcal{V}, \tfrac{\eta}{z}\big).$$

Hence

$$K \ \ge\ \mathcal{N}(\mathcal{P}, \epsilon_{\text{p}}) + \mathcal{N}(\mathcal{V}, \epsilon_{\text{v}}) \ \ge\ \mathcal{N}\big(\mathcal{P}, \tfrac{\eta}{z}\big) + \mathcal{N}\big(\mathcal{V}, \tfrac{\eta}{z}\big),$$

contradicting (E4-3). Therefore *at least one* of $\epsilon_{\text{p}}, \epsilon_{\text{v}}$ is $\ge \eta/z$. Consequently

$$\epsilon_{\text{p}}\, \epsilon_{\text{v}} \ \ge\ \left(\frac{\eta}{z}\right)^2,$$

Define $D_2 := \frac{1}{z^2} > 0$, the η-bound is given by

$$\epsilon_{\text{p}}\, \epsilon_{\text{v}} \ \ge\ D_2\, \eta^2. \tag{E4-4}$$

Step 3. Combine the impacts.

By (E4-2) and (E4-4), we have

$$\epsilon_{\text{p}} \epsilon_{\text{v}} \ge D_1 K^{-2/d_{\text{eff}}} \quad \text{and} \quad \epsilon_{\text{p}} \epsilon_{\text{v}} \ge D_2 \eta^2 \implies \epsilon_{\text{p}}\, \epsilon_{\text{v}} \ \ge\ \max\{D_1 K^{-2/d_{\text{eff}}},\ D_2 \eta^2\}.$$

This completes the proof. □

E.5 Proof of Theorem 2

Restatement of Theorem 2 (Performance Guarantee). *Under Assumption 1 and the covering-regularity of Lemma 3 with constants $a, a', d_{\text{eff}} > 0$ and $b > a,\ b' > a'$, for any budget split $(K_{\text{p}},\ K - K_{\text{p}})$, covering fold k, and token set $\mathcal{X} = \mathcal{V} \sqcup \mathcal{P} \subseteq \mathbb{R}^d$ with $|\mathcal{V}| = N$, $|\mathcal{P}| = L$, and $d_H(\mathcal{V}, \mathcal{P}) \leq \eta$, the following hold:*

(a) ***Performance bound:*** *The Performance degradation caused by MoB is upper bounded by*

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathrm{MoB}(\mathcal{X}))\| \ \leq\ C_\ell\ \max\Big\{\alpha(\eta, k, L)\,(K_{\text{p}})^{-1/d_{\text{eff}}},\ \beta\,(K - K_{\text{p}})^{-1/d_{\text{eff}}}\Big\}\ +\ C_\ell\,\eta,$$

where $\alpha(\eta, k, L)\ =\ \eta\,\big(b\,k\,L/a\big)^{1/d_{\text{eff}}}, \quad \beta\ =\ 2(b')^{1/d_{\text{eff}}}.$

(b) ***Multilinear complexity:*** *The complexity of MoB is given by $T_{\text{MoB}} = \mathcal{O}(N\,(L + K)\,d)$.*

Remark (Coupling Trade-off). *Under weak coupling (large $\alpha(\eta, k, L)$), minimizing the bound requires a larger K_{p}. Conversely, under strong coupling (small $\alpha(\eta, k, L)$), the alignment term decays rapidly, favoring visual preservation (increasing $K - K_{\text{p}}$). Specially, under perfect coupling ($\eta = 0$), the bound simplifies to $\|\Delta y\|\ \leq\ C_\ell\,\beta\,(K - K_{\text{p}})^{-1/d_{\text{eff}}}$,*i.e.*, MoB reduces to pure visual preservation.*

Remark (Budget Scaling). *As the total budget K increases, the preservation term $\beta\,(K - K_{\text{p}})^{-1/d_{\text{eff}}}$ decays, requiring a corresponding increase in K_{p} (and thus a reduction in the alignment term) to rebalance the trade-off and further lower the overall error bound.*

Remark (Scalability). *MoB exhibits a multilinear scalability with respect to visual tokens N, prompt tokens L, and retained tokens K (especially $K, L \ll N$), making it readily adaptable to more challenging scenarios, such as advanced MLLMs with higher-resolution inputs or multi-frame video.*

Notation.

- The intermediate input $\mathcal{X}$ is formulated as

$$\mathcal{X} = \mathcal{V} \sqcup \mathcal{P}\ \subseteq \mathbb{R}^d \quad \text{where} \quad |\mathcal{V}| = N,\ \ |\mathcal{P}| = L,\ \ \text{and}\ \ N \gg L.$$

  Particularly, $\mathcal{V}$, $\mathcal{P}$ are compact sets with d_{eff} effective dimensions.

- We define the pruned intermediate input as

$$\mathrm{MoB}(\mathcal{X}) \coloneqq \mathcal{X}_{\text{s}}, \quad \text{where} \quad \mathcal{X}_{\text{s}} = \mathcal{S} \sqcup \mathcal{P} \quad \text{where} \quad |\mathcal{S}| = K.$$

- The budget configuration is given by $\langle K_{\text{p}}, K_{\text{v}} \rangle$, where $K_{\text{p}} + K_{\text{v}} = K$.

Proof. We separately proof the Performance Guarantee & Complexity in Part A & Part B

Part A: Performance Guarantee

Part A-1: Performance Guarantee of prompt alignment

Step A-1.1: Bound of the radius derived by k-fold NN-covering

Given any union set before K_{p}-truncation

$$\mathcal{S}'_{\text{p}} \coloneqq \bigcup_{p \in \mathcal{P}} \arg\mathop{\text{top-k}}_{s_{\text{p}} \in \mathcal{V}}(\cos(s_{\text{p}}, p), k) \quad \text{where} \quad |\mathcal{S}'_{\text{p}}| = K'_{\text{p}} \quad \text{and} \quad K_{\text{p}} \leq K'_{\text{p}} \leq kL,$$

we define

$$\epsilon'_{\text{p}}\ =\ d_H\big(\mathcal{S}'_{\text{p}},\ \mathcal{P}\big).$$

By previous work [13], NN-covering achieves a 1-approximation for the k-center problem with sufficient budget; *i.e.*, specifically for any $p \in \mathcal{P}$ we have

$$\inf_{s'_{\text{p}} \in \mathcal{S}'_{\text{p}}} \|p - s'_{\text{p}}\| = \inf_{v \in \mathcal{V}} \|p - v\|.$$

Thus,

$$\sup_{p \in \mathcal{P}} \inf_{s'_{\text{p}} \in \mathcal{S}'_{\text{p}}} \|p - s'_{\text{p}}\| = \sup_{p \in \mathcal{P}} \inf_{v \in \mathcal{V}} \|p - v\|.$$

Based on Assumption 2, since $s \in \mathcal{S}'_{\mathrm{p}} \subseteq \mathcal{V}$, the upper bound of the radius ϵ'_{p} is given by

$$\begin{aligned} \epsilon'_{\mathrm{p}} = d_H\big(\mathcal{S}'_{\mathrm{p}},\, \mathcal{P}\big) &\coloneqq \max\{ \sup_{s'_{\mathrm{p}} \in \mathcal{S}'_{\mathrm{p}}} \inf_{p \in \mathcal{P}} \|p - s'_{\mathrm{p}}\|,\ \sup_{p \in \mathcal{P}} \inf_{s'_{\mathrm{p}} \in \mathcal{S}'_{\mathrm{p}}} \|p - s'_{\mathrm{p}}\|\} \\ &\leq \max\{\sup_{v \in \mathcal{V}} \inf_{p \in \mathcal{P}} \|p - v\|,\ \sup_{p \in \mathcal{P}} \inf_{v \in \mathcal{V}} \|p - s'_{\mathrm{p}}\|\} \qquad \text{(E5-1)} \\ &\coloneqq d_H\big(\mathcal{V},\, \mathcal{P}\big) \leq \eta. \end{aligned}$$

Step A-1.2: Impact of K_{p}-truncation on the radius

Based on Lemma 3, we have

$$a r^{-d_{\mathrm{eff}}} \leq \mathcal{N}(\mathcal{P}, r) \leq b\, r^{-d_{\mathrm{eff}}}.$$

In particular:

$$b\,(\epsilon_{\mathrm{p}})^{-d_{\mathrm{eff}}} \ \geq\ K_{\mathrm{p}} \quad \Longrightarrow \quad \epsilon_{\mathrm{p}} \ \leq\ \left(\tfrac{b}{K_{\mathrm{p}}}\right)^{1/d_{\mathrm{eff}}}.$$

and also

$$a\,(\epsilon'_{\mathrm{p}})^{-d_{\mathrm{eff}}} \ \leq\ K'_{\mathrm{p}} \quad \Longrightarrow \quad \epsilon'_{\mathrm{p}} \ \geq\ \left(\tfrac{a}{K'_{\mathrm{p}}}\right)^{1/d_{\mathrm{eff}}}.$$

Combining the upper and lower bound for ϵ_{p} and ϵ'_{p}, respectively in terms of $b, K_{\mathrm{p}}, K'_{\mathrm{p}}$, we obtain

$$\epsilon_{\mathrm{p}} \ \leq\ \left(\tfrac{b}{K_{\mathrm{p}}}\right)^{1/d_{\mathrm{eff}}} \ =\ \left(\tfrac{bK'_{\mathrm{p}}}{aK_{\mathrm{p}}}\right)^{1/d_{\mathrm{eff}}} \cdot \left(\tfrac{a}{K'_{\mathrm{p}}}\right)^{1/d_{\mathrm{eff}}} \ \leq\ \left(\tfrac{bK'_{\mathrm{p}}}{aK_{\mathrm{p}}}\right)^{1/d_{\mathrm{eff}}} \epsilon'_{\mathrm{p}}.$$

That is, truncating from K'_{p} to K_{p} centers increases the radius by at most the factor

$$\epsilon_{\mathrm{p}} \ \leq\ \big(bK'_{\mathrm{p}}/aK_{\mathrm{p}}\big)^{1/d_{\mathrm{eff}}}\, \epsilon'_{\mathrm{p}}.$$

Since $kL \geq K'_{\mathrm{p}}$, loading into above, we have

$$\epsilon_{\mathrm{p}} \ \leq\ \big(bkL/aK_{\mathrm{p}}\big)^{1/d_{\mathrm{eff}}}\, \epsilon'_{\mathrm{p}}.$$

By loading (E5-1) into the above, the performance guarantee of prompt alignment is given by

$$\epsilon_{\mathrm{p}} \coloneqq d_H\big(\mathcal{S}_{\mathrm{p}},\, \mathcal{P}\big) \ \leq\ \alpha(\eta, k, L)\,(K_{\mathrm{p}})^{-1/d_{\mathrm{eff}}} \quad \text{where} \quad \alpha(\eta, k, L) \coloneqq \eta\,\big(bkL/a\big)^{1/d_{\mathrm{eff}}}. \qquad \text{(E5-2)}$$

Part A-2: Performance Guarantee of Visual Preservation

By previous work [33], FPS achieves a 2-approximation for the k-center problem:

$$\epsilon_{\mathrm{v}} \leq 2\,\epsilon^{\star}(K_{\mathrm{v}}), \qquad \text{(E5-3)}$$

where $\epsilon^{\star}(K_{\mathrm{v}})$ is the optimal radius with K_{v} centers. Based on Lemma 3, we have

$$\mathcal{N}(\mathcal{V}, r) \leq b'\, r^{-d_{\mathrm{eff}}},$$

thereby, the upper bound of optimal radius is given by

$$\epsilon^{\star}(K_{\mathrm{v}}) \leq (b'/K_{\mathrm{v}})^{1/d_{\mathrm{eff}}}.$$

By loading the above into (E5-3), the performance guarantee of visual preservation is given by

$$\epsilon_{\mathrm{v}} \coloneqq d_H(\mathcal{S}_{\mathrm{v}}, \mathcal{V}) \leq \beta\,(K_{\mathrm{v}})^{-1/d_{\mathrm{eff}}}, \quad \text{where} \quad \beta \coloneqq 2\, {b'}^{1/d_{\mathrm{eff}}}. \qquad \text{(E5-4)}$$

Part A-3: Performance Guarantee of MoB

By substituting (E5-2) and (E5-4) into Lemma 2, the performance guarantee of the MoB is given by:

$$\|\mathcal{F}(\mathcal{X}) - \mathcal{F}(\mathrm{MoB}(\mathcal{X}))\| \ \leq\ C_\ell\ \max\Big\{\alpha(\eta, k, L)\,(K_{\mathrm{p}})^{-1/d_{\mathrm{eff}}},\ \beta\,(K_{\mathrm{v}})^{-1/d_{\mathrm{eff}}}\Big\} \ +\ C_\ell\,\eta,$$

where $\alpha(\eta, k, L) \ =\ \eta\,\big(b\,k\,L/a\big)^{1/d_{\mathrm{eff}}}, \quad \beta \ =\ 2\, {b'}^{1/d_{\mathrm{eff}}}.$

This completes the proof of Part A.

Part B: Complexity

Since $k \ll K_\mathrm{p} \le K \sim L \ll N$, we restrict our complexity analysis to the leading-order terms.

Part B-1: Normalization

MoB do a L_2 normalization for each token $x \in \mathcal{X} \subseteq \mathbb{R}^d$; thus, the complexity is given by

$$T_\mathrm{norm} = \mathcal{O}((N+L)\,d). \tag{E5-5}$$

Part B-2: Selection of Prompt Center

Firstly, MoB calculates the cosine similarity with each $p \in \mathcal{P}$ and $v \in \mathcal{V}$ via a matrix multiplication:

$$\mathbf{M}_\mathrm{sim} = \mathbf{P}\,\mathbf{V}^\top \in \mathbb{R}^{L\times N} \quad \text{where} \quad \mathbf{V} \in \mathbb{R}^{N\times d} \ \text{ and } \ \mathbf{P} \in \mathbb{R}^{L\times d},$$

which leads a complexity of $T_{\text{step 1-1}} = \mathcal{O}(N\,L\,d)$. Subsequent, MoB do a top-k retrieval in the first dimension of $\mathbf{M}_\mathrm{sim}$ the select k most closed centers for each prompt token $p \in \mathcal{P}$, which can be reduced to a partial sorting, thereby leading to a complexity of $T_{\text{step 1-2}} = \mathcal{O}(N\,L\,\log k)$. Finally, MoB merge the selected result of each $p \in \mathcal{P}$, and truncated the top-K_p ones with largest similarity, leading to a $T_{\text{step 1-3}} = \mathcal{O}(L\,k\,\log K_\mathrm{p})$. Consequently, the total complexity $T_{\text{p-select}}$ of prompt center selection is given by:

$$\begin{aligned} T_{\text{p-select}} &= T_{\text{step 1-1}} + T_{\text{step 1-2}} + T_{\text{step 1-3}}, \\ &= \mathcal{O}(N\,L\,d) + \mathcal{O}(N\,L\,\log k) + \mathcal{O}(L\,k\,\log K_\mathrm{p}), \\ &= \mathcal{O}(N\,L\,d). \end{aligned} \tag{E5-6}$$

Part B-3: Selection of Visual Center

Initially, MoB calculates the minimum distance (used in FPS) with each visual token $v \in \mathcal{V} \backslash \mathcal{S}_\mathrm{p} := \mathcal{V}'$ and the selected prompt centers via a matrix multiplication together with an argmin operator:

$$\mathbf{d}_\mathrm{FPS} = \arg\min \mathbf{V}'^\top\,\mathbf{S}_\mathrm{p} \in \mathbb{R}^{N-K_\mathrm{p}} \quad \text{where} \quad \mathbf{V}' \in \mathbb{R}^{(N-K_\mathrm{p})\times d} \ \text{ and } \ \mathbf{S}_\mathrm{p} \in \mathbb{R}^{K_\mathrm{p}\times d},$$

thus, the complexity is given by

$$\begin{aligned} T_{\text{step 2-1}} &= \underbrace{\mathcal{O}((N-K_\mathrm{p})\,K_\mathrm{p}\,d)}_{\text{matrix multiplication}} + \underbrace{\mathcal{O}((N-K_\mathrm{p})\,K_\mathrm{p})}_{\text{argmin}}, \\ &= \mathcal{O}((N-K_\mathrm{p})\,K_\mathrm{p}\,d). \end{aligned}$$

Subsequently, in $K - K_\mathrm{p}$ iterations, MoB add the tokens with largest minimum distance with an argmax operator in $\mathbf{d}_\mathrm{FPS}$, and update the $\mathbf{d}_\mathrm{FPS}$ with an inner production together with an $N - K_\mathrm{p}$-dimensional element-wise comparison; thus the complexity is given by

$$\begin{aligned} T_{\text{step 2-2}} &= \underbrace{\mathcal{O}((N-K_\mathrm{p})(K-K_\mathrm{p}))}_{\text{argmax}} + \underbrace{\mathcal{O}((K-K_\mathrm{p})\,N\,d)}_{\text{inner productioin}} + \underbrace{\mathcal{O}((K-K_\mathrm{p})\,d)}_{\text{ele}-\text{wise comparision}}, \\ &= \mathcal{O}((K-K_\mathrm{p})\,N\,d). \end{aligned}$$

Consequently, the total complexity $T_{\text{v-select}}$ of visual center selection is given by:

$$\begin{aligned} T_{\text{v-select}} &= T_{\text{step 2-1}} + T_{\text{step 2-2}}, \\ &= \mathcal{O}((N-K_\mathrm{p})\,K_\mathrm{p}\,d) + \mathcal{O}((K-K_\mathrm{p})\,N\,d), \\ &= \mathcal{O}(N\,K\,d). \end{aligned} \tag{E5-7}$$

Part B-4: Totally complexity

By (E5-5), (E5-6) and (E5-7), the totally complexity of MoB is given by

$$\begin{aligned} T_\mathrm{MoB} &= T_\mathrm{norm} + T_{\text{p-select}} + T_{\text{v-select}}, \\ &= \mathcal{O}((N+L)\,d) + \mathcal{O}(N\,L\,d) + \mathcal{O}(N\,K\,d), \\ &= \mathcal{O}(N\,L\,d) + \mathcal{O}(N\,K\,d), \\ &= \mathcal{O}(N\,(L+K)\,d). \end{aligned}$$

This completes the proof of Part B.

Combining the Part A & B, we complete the proof.

□