Semantically Comprehensive Token Pruning in LVLMs via Maximizing Concept Coverage

Xueting Li[1], **Qi liu**[1], **Chenghao Xu**[2], **Xu Yang**[1], **Guangtao Lyu**[1], **Jiahua Li**[1], **Cheng Deng**[1]*
[1] School of Electronic Engineering, Xidian University, Xi'an, China,
[2] College of Information Science and Engineering, Hohai University, Changzhou, China,
{lxt,qiliu,guangtaolyu}@stu.xidian.edu.cn, chx@hhu.edu.cn
{xuyang.xd,ljhxdu,chdeng.xd}@gmail.com

Abstract

High-resolution visual tokens impose substantial computational burdens owing to extreme redundancy in Large Visual Language Models (LVLMs). Existing visual token pruning methods typically leverage simple metrics derived from human experience, such as attention or similarity, to rank and select tokens within a highly entangled feature space. However, these metrics lack interpretability and often introduce human bias, failing to capture the genuine semantic significance of tokens, especially amidst the inherent semantic complexity and ambiguity of visual tokens. To mitigate this limitation, we propose a novel Semantically Comprehensive Token Selection (SCTS) method for unbiased, interpretable visual token pruning via a concept-driven paradigm. To unravel the model's intrinsic semantic representation mechanism, we first introduce a Sparse Autoencoder to disentangle visual features into an interpretable space, with each dimension encoding a distinct semantic concept. We then formulate the token pruning task as a Maximum Concept Coverage problem, quantifying the Marginal Semantic Gain (MSG) of each token's contribution to uncovered concepts and iteratively selecting tokens with the highest MSG. This concept-centric approach prioritizes tokens with unique semantic contributions, guaranteeing semantic comprehensiveness while preserving robust performance even at high compression ratios. Extensive experiments across multiple LVLM architectures and benchmarks verify that SCTS consistently outperforms state-of-the-art approaches, achieving a superior trade-off between computational efficiency and semantic completeness.

1 Introduction

Large Vision Language Models (LVLMs) (Zhu et al., 2024; Liu et al., 2024a,c,b; Li et al., 2024b, 2025; Bai et al., 2025) have significantly enhanced the capability to capture fine-grained details through high-resolution inputs. However, this advancement comes at a steep price: the proliferation of visual tokens triggers quadratic computational complexity within Transformer architectures. This massive spatial redundancy, introduced by high-resolution processing, imposes a prohibitive computational overhead, which has become the primary bottleneck restricting the practical application of LVLMs (Kim et al., 2024; Qu et al., 2025). Consequently, visual token pruning has emerged as a critical strategy to reduce inherent redundancy and alleviate this computational pressure.

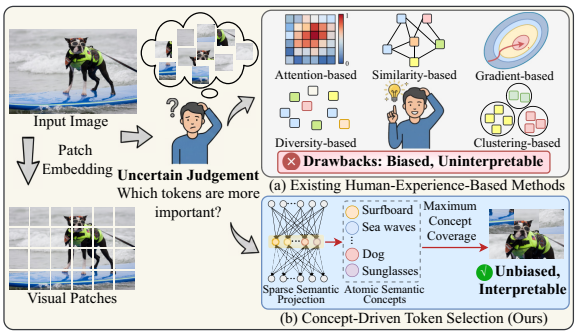


Figure 1: (a) Existing methods rely on human experience metrics, leading to subjective bias and poor interpretability. (b) Ours uncovers the intrinsic semantic concepts of LVLMs and reformulates pruning as a Maximum Concept Coverage problem. By maximizing the coverage of unique semantic concepts, our SCTS ensures objective and unbiased token selection while preserving both interpretability and semantic integrity.

Despite progress in visual token pruning, existing approaches predominantly rely on metrics derived from human experience, such as attention weights (Chen et al., 2024; Zhang et al., 2025c; Hu et al., 2025; Zhang et al., 2025a) or similarity (Li et al., 2024a; Sun et al., 2025a), to assess token importance, operating within a highly entangled feature space. Fundamentally, this adherence to

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27829–27846
July 2-7, 2026

external definitions imposes artificially set experiential information on the selection process. Since cognitive perspectives on "how to identify critical visual tokens" naturally vary among different researchers, such subjective definitions inevitably introduce human-induced inductive biases and limit interpretability (Figure 1 (a)). Furthermore, given the inherent polysemy and high complexity of visual features, relying on simple human experience metrics often fails to capture their multifaceted semantic content comprehensively. As a result, the selected tokens frequently misalign with the intrinsic semantic information required for model reasoning, compromising information integrity.

In this paper, we argue that token pruning should be grounded in the model's intrinsic semantics, rather than relying on external, manually defined metrics. To this end, we propose Semantically Comprehensive Token Selection (SCTS), a novel unbiased and interpretable visual token pruning framework. Diverging from approaches that rely on human experience, SCTS achieves unbiased token selection via an objective mechanism rooted in the model's intrinsic semantic concepts. First, to reveal the model's intrinsic semantic representation, we leverage a Sparse Autoencoder (SAE) as an internal knowledge decoupler, mapping visual token features into an interpretable space composed of atomic semantic concepts. Subsequently, based on these disentangled concepts, we reformulate the token pruning task as a Maximum Concept Coverage (MCC) problem (Figure 1 (b)). *Our goal is not to rank tokens, but to identify the minimal subset of tokens capable of covering all activated semantic concepts within the image.*

To effectively address this MCC problem, we define Marginal Semantic Gain (MSG) to quantify the incremental contribution of each token toward semantic concepts uncovered by the current subset. We then employ an efficient greedy strategy to select tokens that maximize this MSG iteratively. This mechanism intrinsically penalizes redundancy, by which tokens replicating previously covered concepts deliver zero marginal gain and are discarded, whereas those encoding unique, uncovered semantic concepts are retained. By grounding the selection process in the model's intrinsic conceptual mechanism, we adhere to a "concept-first" principle—prioritizing tokens that capture unique semantic concepts. This approach not only ensures an unbiased and interpretable filtration of visual information but also significantly enhances semantic completeness, enabling superior performance even under high compression ratios.

We validate SCTS through extensive experiments across diverse vision-language tasks and multiple advanced LVLMs, such as LLaVA-1.5, LLaVA-Next, and Qwen-2.5. Experimental results demonstrate that our method achieves superior performance, effectively retaining unique semantic details even under high compression ratios. Our main contributions are summarized as follows:

- We propose a novel visual token pruning framework, SCTS, which leverages an SAE to uncover the mechanisms of semantic expression in LVLMs and redefines pruning as an interpretable, concept-driven process, effectively addressing subjective human biases.
- We are the first to formally define visual token pruning as an MCC problem and design the MSG to select tokens based on their unique semantic contributions, ensuring semantic integrity and providing fine-grained interpretability in the selection process.
- We apply SCTS to various advanced LVLMs and conduct comprehensive experiments on a range of vision-language tasks, demonstrating that our method maintains high performance even at extreme reduction ratios.

2 Related Work

Large Vision-Language Models. Driven by the paradigm shift in Large Language Models, models such as LLaVA (Liu et al., 2024a,b), Qwen-VL (Bai et al., 2023), and Mini-Gemini (Li et al., 2025) have propelled the dominance of LVLMs by aligning visual and linguistic spaces. To overcome perception bottlenecks, modern architectures (e.g., LLaVA-Next (Liu et al., 2024c) and Qwen2.5-VL (Bai et al., 2025)) widely adopt dynamic resolution strategies, encoding high-resolution images into extensive sequences of visual tokens. However, this pursuit of fine-grained detail comes at a significant cost: the self-attention mechanism in Transformers incurs a computational complexity that grows quadratically with the number of tokens. This renders inference in high-resolution LVLMs extremely inefficient, underscoring an urgent need for effective token compression schemes.

Visual Token Pruning and Compression. To address efficiency bottlenecks, visual token pruning has become a key research direction. Existing

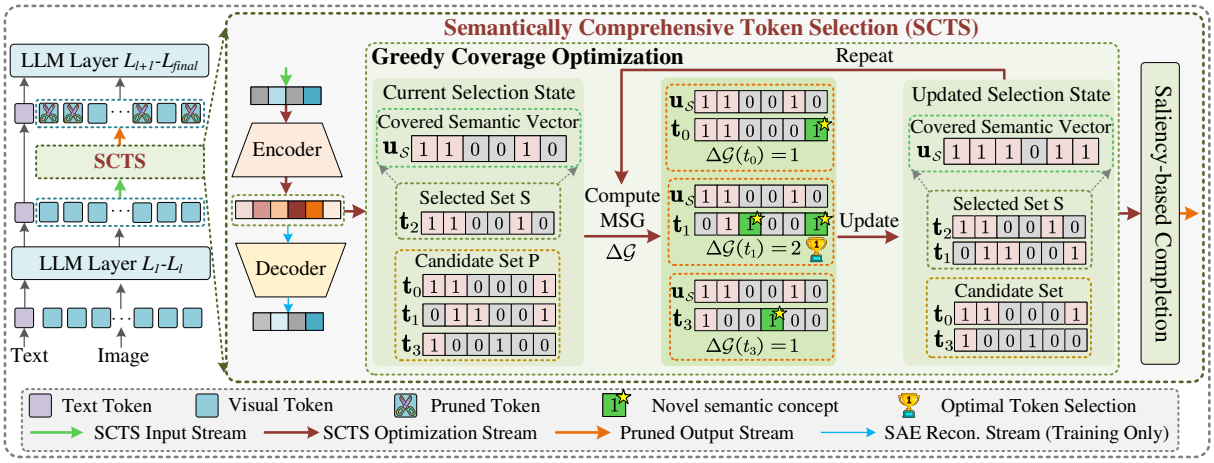


Figure 2: Overview of SCTS. Left: Integration of SCTS within the LVLM pipeline for efficient inference. Right: Details of the SCTS workflow. First, the Greedy Coverage Optimization mechanism iteratively selects tokens with the maximum MSG ($\Delta\mathcal{G}$, Equation 7) to construct a minimal subset ensuring semantic completeness. Subsequently, any remaining token budget is allocated to Saliency-based Completion to preserve fine-grained spatial details.

methods mainly rely on human experience metrics, utilizing metrics such as attention weights (Chen et al., 2024; Hu et al., 2025; Zhang et al., 2025a; Lyu et al., 2026b; Hu et al., 2025; Xu et al., 2026), gradients (Kim et al., 2025), similarity scores (Li et al., 2024a; Sun et al., 2025a), or clustering (Ma et al., 2026) to assess token importance. However, operating in a highly entangled feature space, their reliance on scalar proxies often fails to capture complex semantics, frequently leading to the loss of critical tokens. While learning-based alternatives (Huang et al., 2024; Zhang et al., 2025b; Lyu et al.; Zeng et al., 2025; Sun et al., 2025b) offer different approaches, they are often hindered by high re-training costs and opacity. In contrast, our SCTS framework leverages SAE to uncover intrinsic semantic representations. By reformulating pruning as an MCC problem, SCTS prioritizes intrinsic semantic completeness.

3 Methodology

In this section, we first introduce Sparse Semantic Projection in Sec. 3.1, which disentangles features into atomic concepts to lay an interpretable semantic foundation. Subsequently, in Sec. 3.2, we detail the core token selection mechanism, where we reformulate the pruning task as a MCC problem to prioritize semantic completeness. The overall architecture of our SCTS is illustrated in Figure 2.

3.1 Sparse Semantic Projection

To reveal the model's intrinsic semantic expression, we introduce a sparse semantic projection mechanism to map visual token features into an interpretable space of atomic concepts, laying a disentangled semantic foundation for subsequent token selection. Specifically, we leverage the SAE architecture from Bricken et al. (Bricken et al., 2023; Lyu et al., 2026a) to decompose dense, polysemantic representations into independent, monosemantic features.

Formally, let $\mathbf{z}_i^{(l)} \in \mathbb{R}^D$ denote the hidden state of the i-th token at layer l, where D is the hidden dimension. The SAE projects $\mathbf{z}_i^{(l)}$ into an overcomplete latent space of dimension M and subsequently reconstructs it:

$$\mathbf{h}_i = \text{ReLU}(W_{enc}(\mathbf{z}_i^{(l)} - \mathbf{b}_{pre}) + \mathbf{b}_{enc}), \quad (1)$$

$$\hat{\mathbf{z}}_i^{(l)} = W_{dec}\mathbf{h}_i + \mathbf{b}_{dec}, \quad (2)$$

where $\mathbf{h}_i \in \mathbb{R}^M$ is the sparse feature vector, $\hat{\mathbf{z}}_i^{(l)}$ is the reconstructed state, W_{enc}, W_{dec} are learnable weight matrices, and $\mathbf{b}_{pre}, \mathbf{b}_{enc}, \mathbf{b}_{dec}$ are bias vectors. To ensure $\mathbf{h}_i$ captures disentangled atomic semantics, the network is optimized in an unsupervised manner via the composite objective:

$$\mathcal{L} = \|\mathbf{z}^{(l)} - \hat{\mathbf{z}}^{(l)}\|_2^2 + \lambda\|\mathbf{h}\|_1. \quad (3)$$

The first term ensures reconstruction fidelity, while the L_1 penalty, weighted by λ, explicitly enforces sparsity by driving most neurons to zero. Consequently, each active dimension in $\mathbf{h}_i$ signifies a distinct atomic semantic concept.

3.2 Concept-Driven Token Selection

Problem Reformulation. Building upon the disentangled semantic foundation, we fundamentally

reformulate the pruning task. Departing from existing methods that depend on simple metrics derived from human experience for token selection, we model the process as an MCC Problem. Given the full set of visual token indices $\mathcal{V} = \{1, \ldots, N\}$, where N denotes the total number of tokens, our primary objective is to identify an optimal subset $\mathcal{S} \subset \mathcal{V}$ that achieves comprehensive coverage of the global semantic universe $\mathcal{U}$ (defined below). This perspective prioritizes semantic completeness, ensuring that no unique visual concept is omitted.

Discrete Concept Generation. To enable discrete optimization, we binarize the continuous sparse activations. For each token $i \in \mathcal{V}$, we generate its corresponding binary concept vector $\mathbf{t}_i \in \{0, 1\}^M$ via thresholding:

$$t_{i,j} = \begin{cases} 1, & \text{if } h_{i,j} \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where τ is the significance threshold and $j \in \{1, \ldots, M\}$ indexes semantic dimensions. Thus, $\mathbf{t}_i$ serves as the discrete representation of token i in the concept space. We formally define the global semantic universe of the image as:

$$\mathcal{U} = \bigcup_{i \in \mathcal{V}} \{j \mid t_{i,j} = 1\}, \tag{5}$$

where the operator $\bigcup$ denotes the set-theoretic union, aggregating all unique semantic indices present across the token population.

Greedy Coverage Optimization. Solving the MCC problem exactly is an NP-hard combinatorial optimization task. To ensure efficiency, we adopt a greedy search strategy based on our proposed MSG. Let $\mathbf{u}_\mathcal{S} \in \{0, 1\}^M$ denote the cumulative semantic coverage vector of the currently selected subset of token indices $\mathcal{S} \subseteq \mathcal{V}$. It is updated by aggregating the binary concept vectors of all selected tokens:

$$\mathbf{u}_\mathcal{S} = \bigvee_{k \in \mathcal{S}} \mathbf{t}_k, \tag{6}$$

where the symbol $\bigvee$ represents the element-wise logical OR operation. Let $\mathcal{P} = \mathcal{V} \setminus \mathcal{S}$ denote the set of unselected candidate tokens. For any candidate $p \in \mathcal{P}$, we define the MSG $\Delta\mathcal{G}(p)$ as the number of new semantic concepts it activates. This is calculated by masking the candidate's vector with the inverse of the current coverage:

$$\Delta\mathcal{G}(p) = \|\mathbf{t}_p \odot (\mathbf{1} - \mathbf{u}_\mathcal{S})\|_1, \tag{7}$$

where $\odot$ denotes the Hadamard product, $\mathbf{1}$ is a vector of all ones, and $\|\cdot\|_1$ represents the L_1 norm. The term $(\mathbf{1} - \mathbf{u}_\mathcal{S})$ acts as a filter, preserving only the dimensions that are not yet covered.

The optimization process iteratively identifies the optimal token p^* that maximizes this gain:

$$p^* = \arg\max_{p \in \mathcal{V} \setminus \mathcal{S}} \Delta\mathcal{G}(p). \tag{8}$$

To resolve ambiguity, if multiple tokens provide the same $\Delta\mathcal{G}(p)$, we select the one with the highest semantic density $\|\mathbf{t}_p\|_1$, prioritizing tokens with richer information content.

Following the selection of p^*, we update the subset via $\mathcal{S} \leftarrow \mathcal{S} \cup \{p^*\}$ and the coverage vector via $\mathbf{u}_\mathcal{S} \leftarrow \mathbf{u}_\mathcal{S} \vee \mathbf{t}_{p^*}$. This iterative procedure continues until the target budget $|\mathcal{S}| = B$ is reached or the maximum MSG drops to zero, ensuring that each step minimizes semantic redundancy while maximizing the visual diversity.

Implementation Note. While theoretically formulated on $\mathbb{R}^M$, our implementation exploits SAE sparsity by restricting operations to the effective universe $\mathcal{U}$. This significantly reduces computational and memory overhead. Further details are provided in Appendix B.

Saliency-based Completion. In scenarios where the maximum MSG drops to zero before the token budget is exhausted, the selection strategy transitions to a saliency-based completion. The remaining slots in $\mathcal{S}$ are filled by prioritizing tokens from the candidate set $\mathcal{P}$ with the highest information density, quantified by the L_1 norm $\|\mathbf{t}_i\|_1$. This unified strategy ensures that the pruned subset first guarantees semantic completeness and subsequently maximizes information richness.

3.3 Integration with LVLM Inference

Upon identifying the semantically complete subset $\mathcal{S}$, we refine the input sequence for subsequent processing. Specifically, the original hidden states are pruned such that $\mathbf{Z}^{(l)}_{pruned} = \{\mathbf{z}^{(l)}_i \mid i \in \mathcal{S}\}$, effectively condensing the sequence length from N to $|\mathcal{S}|$. These representative tokens are then propagated through the remaining layers. This reduction significantly mitigates the quadratic complexity of subsequent self-attention operations.

4 Experiments

4.1 Experimental Setup

Implementation Details. To verify the efficacy of our framework, we conduct extensive experiments

Methods	GQA	MMB^{EN}	MMB^{CN}	MME	POPE	SQA^{IMG}	VQA^{V2}	VQA^{Text}	VizWiz	Average
Upper Bound, 576 Tokens	61.9	64.7	58.1	1862	85.9	69.5	78.4	58.2	50	100%
LLaVA-1.5-7B	Retain 192 Tokens (↓ 66.7%)									
ToMe (ICLR23)	54.3	60.5	-	1563	72.4	65.2	68.0	52.1	-	88.5%
FastV (ECCV24)	52.7	61.2	57.0	1612	64.8	67.3	67.1	52.5	50.8	90.5%
LLaVA-PruMerge (ICCV25)	54.3	59.6	52.9	1632	71.3	67.9	70.6	54.3	50.1	91.4%
PDrop (CVPR25)	57.1	63.2	56.8	1766	82.3	68.8	75.1	56.1	51.1	96.9%
HiRED (AAAI25)	58.7	62.8	54.7	1737	82.8	68.4	74.9	47.4	50.1	94.6%
VisionZip (CVPR25)	**59.3**	64.5	57.3	1767	**86.4**	68.9	76.8	57.3	51.6	98.7%
SparseVLM (ICML25)	57.6	62.5	53.7	1721	83.6	69.1	75.6	56.1	50.5	96.1%
DART (EMNLP25)	58.9	63.6	57.0	1856	82.8	**69.8**	76.7	57.4	51.1	98.5%
HoloV (NeurIPS25)	59.0	**65.4**	58.0	1820	85.6	**69.8**	76.7	57.4	50.9	99.1%
SCTS (Ours)	59.1	65.3	**58.2**	**1864**	86.1	69.7	**76.9**	**57.7**	**52.1**	**99.8%**
LLaVA-1.5-7B	Retain 128 Tokens (↓ 77.8%)									
ToMe (ICLR23)	52.4	53.3	-	1343	62.8	59.6	63.0	49.1	-	80.4%
FastV (ECCV24)	49.6	56.1	56.4	1490	59.6	60.2	61.8	50.6	51.3	85.4%
LLaVA-PruMerge (ICCV25)	53.3	58.1	51.7	1554	67.2	67.1	68.8	54.3	50.3	89.4%
PDrop (CVPR25)	56.0	61.1	56.6	1644	82.3	68.3	72.9	55.1	51.0	94.9%
HiRED (AAAI25)	57.2	61.5	53.6	1710	79.8	68.1	73.4	46.1	51.3	93.1%
VisionZip (CVPR25)	57.6	63.4	56.7	1768	84.7	68.8	75.6	56.8	52.0	97.7%
SparseVLM (ICML25)	56.0	60.0	51.1	1696	80.5	67.1	73.8	54.9	51.4	93.8%
DART (EMNLP25)	57.9	63.2	57.0	**1845**	80.1	69.1	75.9	56.4	51.7	97.5%
HoloV (NeurIPS25)	57.7	**63.9**	56.5	1802	84.0	**69.8**	75.5	56.8	51.5	98.0%
SCTS (Ours)	**58.4**	63.7	**57.1**	1837	**85.6**	69.3	**76.1**	**57.0**	**51.9**	**98.6%**
LLaVA-1.5-7B	Retain 64 Tokens (↓ 88.9%)									
ToMe (ICLR23)	48.6	43.7	-	1138	52.5	50.0	57.1	45.3	-	70.1%
FastV (ECCV24)	46.1	48.0	52.7	1256	48.0	51.1	55.0	47.8	50.8	76.7%
LLaVA-PruMerge (ICCV25)	51.9	55.3	49.1	1549	65.3	68.1	67.4	54.0	50.1	87.8%
PDrop (CVPR25)	41.9	33.3	50.5	1092	55.9	68.6	69.2	45.9	50.7	77.4%
HiRED (AAAI25)	54.6	60.2	51.4	1599	73.6	68.2	69.7	44.2	50.2	89.4%
VisionZip (CVPR25)	55.1	60.1	55.4	1690	77.0	69.0	72.4	55.5	**52.9**	94.5%
SparseVLM (ICML25)	52.7	56.2	46.1	1505	75.1	62.2	68.2	51.8	50.1	87.3%
DART (EMNLP25)	55.9	60.6	53.2	**1765**	73.9	**69.8**	72.4	54.4	51.6	94.0%
HoloV (NeurIPS25)	55.3	**63.3**	55.1	1715	80.3	69.5	72.8	55.4	52.8	95.7%
SCTS (Ours)	**57.2**	63.1	**56.2**	1730	**81.9**	67.9	**73.5**	**55.8**	51.3	**96.1%**

Table 1: Main results on LLaVA-1.5-7B across three token budgets. Gray rows indicate pruning configurations. **Bold** and underline denote the best and second-best performance among reduction methods, respectively.

across three representative LVLM architectures: LLaVA-1.5 (Liu et al., 2024a), LLaVA-Next (Liu et al., 2024c), and Qwen2.5-VL (Bai et al., 2025). For the implementation of SCTS, we integrate the SAE into the 3rd layer of the respective LLM backbones to capture foundational semantic features. All training and evaluations are conducted on a single NVIDIA RTX A6000 Pro (96GB) GPU.

Benchmarks. To evaluate the efficacy of our framework, we conduct experiments on several widely recognized benchmarks, including POPE (Li et al., 2023), VizWiz (Bigham et al., 2010), MMBench and MMB-CN (Liu et al., 2024e), GQA (Hudson and Manning, 2019), MME (Fu et al., 2025), ScienceQA-IMG (Lu et al., 2022), VQA V2 (Goyal et al., 2017), and TextVQA (Singh et al., 2019).

Comparison Methods. We compare our approach with a broad spectrum of token reduction methods, ranging from representative baselines to recent state-of-the-art (SOTA) approaches, including ToMe (Bolya et al., 2023), FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2025c), HiRED (Arif et al., 2025), LLaVA-PruMerge (Shang et al., 2025), PDrop (Xing et al., 2025), MustDrop (Liu et al., 2024d), FasterVLM (Zhang et al., 2024), GlobalCom^2 (Liu et al., 2025), VisionZip (Yang et al., 2025), DART (Wen et al., 2025) and HoloV (Zou et al., 2025). Unlike these methods that depend on human experience metrics, our proposed SCTS leverages the model's intrinsic semantic representations to formulate pruning as an MCC problem. By prioritizing semantic completeness, this principled approach achieves superior performance, particularly at high compression ratios.

Further details about implementation settings and benchmarks are provided in the Appendix A.

4.2 Main Results

Performance on LLaVA-1.5-7B. As shown in Table 1, our method consistently outperforms baselines relying on simple metrics derived from human experience across all budgets. At 66.7% pruning, it achieves 99.8% accuracy, surpassing recent SOTA methods, including SparseVLM (+2.7%),

Methods	GQA	MMB^{EN}	MMB^{CN}	MME	POPE	SQA^{IMG}	VQA^{V2}	VQA^{Text}	VizWiz	Average
Upper Bound, 2880 Tokens	64.2	67.4	60.6	1851	86.5	70.1	81.8	64.9	57.6	100%
LLaVA-NeXT-7B	Retain 320 Tokens (↓ 88.9%)									
FastV (ECCV24)	55.9	61.6	51.9	1661	71.7	62.8	71.9	55.7	53.1	88.00%
LLaVA-PruMerge (ICCV25)	53.6	61.3	55.3	1534	60.8	66.4	69.7	50.6	54.0	85.60%
PDrop (CVPR25)	56.4	63.4	56.2	1663	77.6	67.5	73.5	54.4	54.1	90.90%
MustDrop (2024.11)	57.3	62.8	55.1	1641	82.1	68.0	73.7	59.9	54.0	92.20%
FasterVLM (ICCV25)	56.9	61.6	53.5	1701	83.6	66.5	74.0	56.5	52.6	91.20%
HiRED (AAAI25)	59.3	64.2	55.9	1690	83.3	66.7	75.7	58.8	54.2	93.30%
SparseVLM (ICML25)	56.1	60.6	54.5	1533	82.4	66.1	71.5	58.4	52.0	89.70%
GlobalCom² (AAAI26)	57.1	61.8	53.4	1698	83.8	67.4	76.7	57.2	54.6	92.20%
DART (EMNLP25)	61.7	65.3	**58.2**	1710	84.1	68.4	79.1	58.7	**56.1**	95.60%
HoloV (NeurIPS25)	61.7	65.3	57.5	1738	83.9	68.9	79.5	58.7	55.3	95.60%
SCTS (Ours)	**62.1**	**65.4**	58.1	**1756**	**84.3**	**69.2**	**79.6**	**58.9**	**56.1**	**96.20%**

Table 2: Performance comparison on LLaVA-Next-7B across varying token budgets. **Bold** and underline denote the best and second-best performance, respectively.

Method	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB^{EN}	Avg.
Upper Bound	84.7	84.8	86.1	2304	82.8	100%
Qwen2.5-VL-7B	Retain 192 Tokens (↓ 66.7%)					
FastV (ECCV24)	78.5	77.9	82.2	2072	75.7	92.3%
HoloV (NeurIPS25)	79.8	**78.9**	85.0	2093	78.3	94.3%
SCTS (Ours)	**85.5**	77.3	**86.4**	**2317**	**78.6**	**97.6%**
Qwen2.5-VL-7B	Retain 128 Tokens (↓ 77.8%)					
FastV (ECCV24)	78.0	69.0	80.7	2036	74.9	89.2%
HoloV (NeurIPS25)	79.8	70.3	82.3	2043	76.5	90.8%
SCTS (Ours)	**84.0**	**75.5**	**85.9**	**2218**	**78.1**	**95.7%**
Qwen2.5-VL-7B	Retain 64 Tokens (↓ 88.9%)					
FastV (ECCV24)	77.4	60.3	78.6	1940	69.2	84.3%
HoloV (NeurIPS25)	79.5	61.8	80.7	2006	72.4	87.0%
SCTS (Ours)	**82.9**	**71.0**	**81.5**	**2026**	**73.9**	**90.7%**

Table 3: Performance on Qwen2.5-VL-7B across varying token budgets. **Bold** denotes the best performance.

PDrop (+2.9%), and HoloV (+0.7%). Even under extreme compression (88.9% pruning), it retains 96.1% of the original performance. Remarkably, our approach matches or exceeds the performance upper bound on benchmarks like MMB, POPE, and VizWiz. These results confirm our concept-prioritized semantically complete selection reduces redundancy and enhances performance by eliminating interference.

SCTS with Higher Resolution. LLaVA-NeXT employs a dynamic processing strategy with variable token lengths. Table 2 highlights SCTS's superior scalability on this architecture: even when compressed from 2880 to just 320 tokens, it achieves an average score of 96.2%, significantly surpassing the 95.6% SOTA. This confirms the efficacy of SCTS for high-resolution visual inputs.

SCTS with Qwen Architecture. To verify generalization beyond LLaVA, we evaluated SCTS on Qwen2.5-VL-7B. As shown in Table 3, SCTS consistently outperforms human-experience-based baselines. Notably, with 128 retained tokens, it

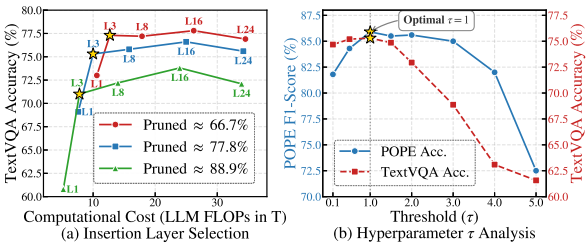


Figure 3: Ablation analyses on the insertion layer and threshold τ (at 77.8% pruning rate).

achieves a retention rate of 95.7%, significantly surpassing FastV (+6.5%) and HoloV (+4.9%). Even under the extreme pruning ratio of 88.9% (64 tokens), SCTS maintains high retention (90.7%), outperforming the strongest baseline, HoloV (+3.7%). These findings demonstrate the robustness of our pruning strategy across diverse architectures.

4.3 Ablation Study

Impact of the SCTS Insertion Layer. We evaluate the accuracy-computation trade-off across varying depths in Figure 3 (a). Insertion at L_1 degrades performance due to premature information loss, while deeper layers ($l \geq 8$) incur excessive overhead with diminishing returns. We select L_3 as the Pareto-optimal choice, which balances significant accuracy recovery with a minimal FLOPs budget compared to deeper layers.

Impact of Hyperparameter τ. Figure 3 (b) shows that lower thresholds ($\tau < 1$) underperform due to insufficient noise filtering, while aggressive thresholds ($\tau > 2$) degrade metrics by discarding critical semantic concepts. The setting $\tau = 1$ strikes the optimal balance, achieving peak performance on both benchmarks. Thus, we adopt $\tau = 1$ for all subsequent experiments.

Impact of Token Selection Strategy. As shown

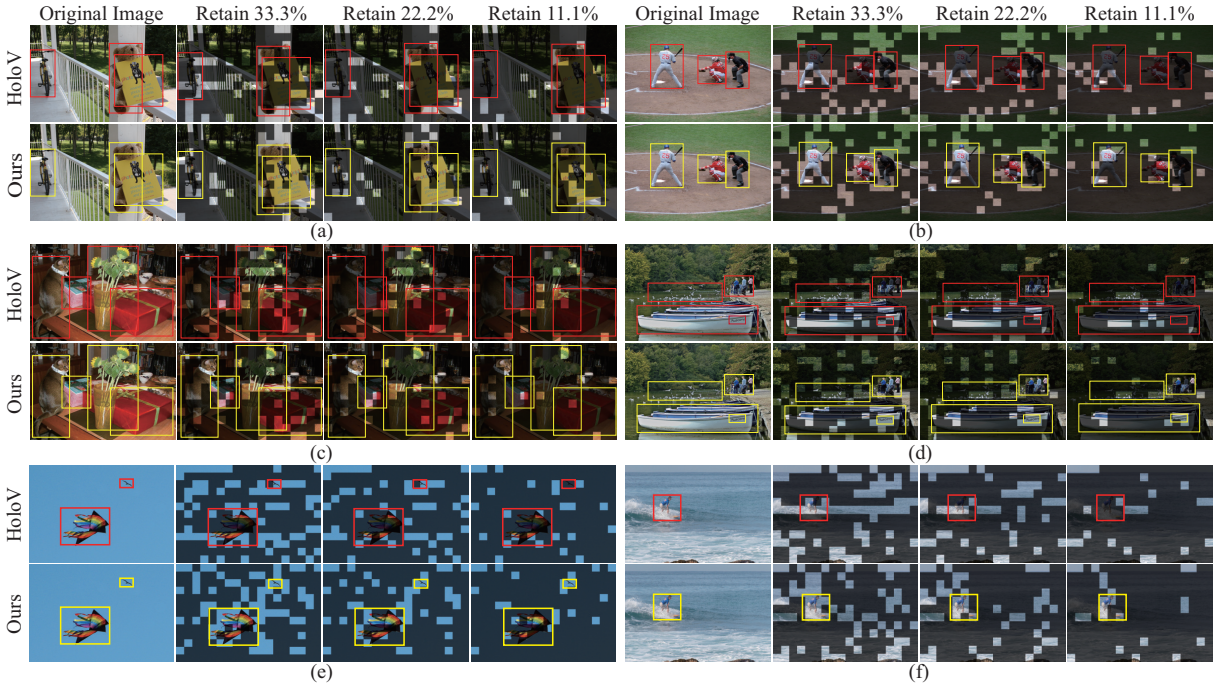


Figure 4: Visual comparison with HoloV across retention rates of 33.3%, 22.2%, and 11.1%. Bounding boxes highlight fine-grained text and objects.

Strategy	POPE			VQA$^{\text{Text}}$		
	↓ 66.7%	↓ 77.8%	↓ 88.9%	↓ 66.7%	↓ 77.8%	↓ 88.9%
Random	83.8	81.6	77.3	71.6	68.0	59.5
Uniform	84.2	83.4	78.9	73.0	70.2	59.1
L_0 Top-k	86.2	85.7	75.8	72.9	68.0	58.1
Ours	**86.4**	**85.9**	**81.5**	**77.3**	**75.5**	**71.0**

Table 4: Ablation study on token selection strategies.

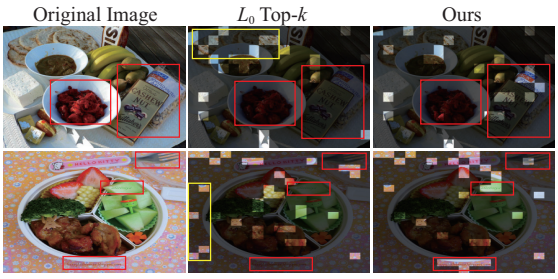


Figure 5: Visualization of token selection masks comparing our method with the L_0 Top-k baseline, where red boxes highlight fine-grained details and yellow boxes indicate redundant regions retained by L_0 Top-k.

in Table 4 and Figure 5, we evaluate SCTS against three baselines: Random sampling (stochastic selection), Uniform sampling (fixed-interval grid retention), and L_0 Top-k (prioritizing tokens activating the most concepts). Surprisingly, L_0 Top-k lags behind even random sampling. Figure 5 reveals the cause: equating activation count with importance wastes the token budget on redundant dominant features, crowding out subtle yet vital details (e.g., text, fork). Conversely, by prioritizing Semantic Coverage, SCTS minimizes redundancy and prevents semantic collapse, ensuring the retention of essential cues.

Detailed ablation on SAE latent dimension M is provided in the Appendix C.

4.4 Qualitative Analysis

Qualitative Visualization. Figure 4 compares the retained tokens of our method and HoloV on Qwen2.5-VL-7B across varying pruning ratios. The results highlight our method's robustness in high compression scenarios. While HoloV, based on human experience, discards semantically important but low-scoring features, leading to the loss of fine details (e.g., bicycle and poster in (a), jersey numbers in (b), and text in (d)) and small objects (e.g., cat in (c), kite in (e), surfer in (f)), our SCTS successfully retains these semantically distinct instances. Additionally, our method preserves structural integrity in text-rich scenes (e.g., (a)), preventing semantic fragmentation and ensuring reliable visual cues for LVLMs.

Visualization of Interpretable Features. Figure 7 validates the SAE's disentanglement capability, revealing that individual dimensions capture distinct atomic concepts. Features #578, #2791, and #804 exhibit precise selectivity for 'Cat', 'Ze-

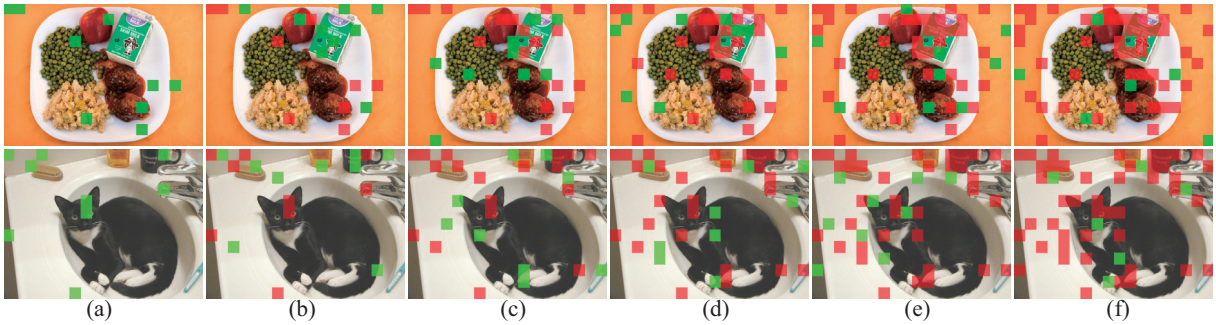
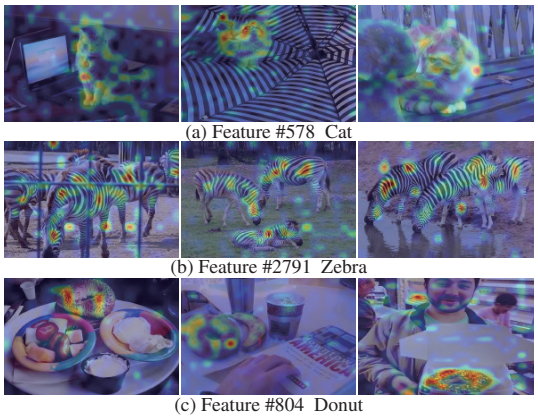


Figure 6: Visualization of the progressive token selection process at 77.8% pruning rate. We visualize the selection sequence across 6 steps ((a)→(f)). Green boxes indicate tokens selected at the *current* step, while Red boxes represent the accumulated selections from *previous* steps. As the process advances, our method incrementally captures unique semantic concepts, aiming for comprehensive semantic preservation while minimizing redundancy.



(a) Feature #578 Cat

(b) Feature #2791 Zebra

(c) Feature #804 Donut

Figure 7: Visualization of learned interpretable features.

bra', and 'Donut' respectively, effectively isolating targets from background clutter. This confirms the decomposition of entangled representations into an orthogonal interpretable space, laying a solid foundation for our concept-based pruning.

Visualization of Token Selection Trajectory Dynamics. Figure 6 illustrates the coarse-to-fine evolution of the selection trajectory at a 77.8% pruning rate. By distinguishing current choices (Green) from the accumulated history (Red), a continuous complementary pattern emerges. Taking the first row as an example, the process initially prioritizes distinct semantic anchors (e.g., apple, plate) to establish the scene skeleton, then progressively covers previously unselected semantic regions to complete the global context (e.g., milk carton), and finally densifies features of critical concepts to enhance local details (e.g., text patterns). This dynamic mechanism ensures that our strategy effectively mitigates redundancy, securing semantic diversity before refining feature fidelity.

Additional visualization results are provided in Appendix G.

Methods	Tokens	Latency	FLOPs	Acc.
Qwen2.5-VL-7B	1296	175 ms	36.6 T	86.1
HoloV (NeurIPS25)	144	158 ms	**2.2 T**	80.7 (↓ 6.3%)
SCTS (Ours)	**144**	**109 ms**	6.0 T	**81.5 (↓ 5.3%)**

Table 5: Efficiency analysis on POPE benchmark.

4.5 Efficiency Analysis

We evaluate efficiency on the POPE benchmark using Qwen2.5-VL-7B, reporting the average end-to-end latency and LLM-decoder FLOPs per sample on a single NVIDIA RTX PRO 6000 Blackwell GPU (96GB). As detailed in Table 5, SCTS achieves a 1.6× speedup over the baseline, with the SAE introducing negligible computational overhead (0.14 T). Compared to HoloV at the same pruning ratio, SCTS yields higher accuracy (81.5% vs. 80.7%) and significantly lower latency (109 ms vs. 158 ms). SCTS replaces HoloV's complex indexing with GPU-optimized matrix operations, effectively converting theoretical efficiency into real-world speedups.

5 Conclusion

In this paper, we propose SCTS, a novel visual token pruning framework for LVLMs. We introduce an SAE to disentangle visual features into atomic semantic concepts, then reformulate pruning as an MCC problem to quantify token contributions via MSG. This enables SCTS to select tokens while ensuring semantic completeness. Our method prioritizes semantically distinct tokens, maintaining high performance even under aggressive compression. Extensive experiments across various LVLM architectures and benchmarks demonstrate that SCTS

outperforms SOTA methods.

6 Limitations

While SCTS enhances efficiency and interpretability, we acknowledge certain limitations. First, although the SAE's marginal overhead is offset by LLM savings, the pipeline requires further optimization to maximize throughput. Second, pre-trained SAE weights cannot be directly transferred across LVLMs due to dimensional and distributional mismatches; while adaptation is lightweight, universal weight transfer remains challenging. Third, as the SAE learns from internal representations, it may inherently reflect training data distributions, necessitating careful evaluation for safety-critical deployment. Future work will focus on cross-model adaptation, distributional robustness, and broader practical applications.

Acknowledgments

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600) and the National Natural Science Foundation of China (U25B2048, 62132016).

References

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. 2025. HiRED: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1773–1781.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and 1 others. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 333–342.

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

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 19–35.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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 (CVPR)*, pages 6904–6913.

Lianyu Hu, Fanhua Shang, Wei Feng, and Liang Wan. 2025. LightVLM: Acceleraing large multimodal models with pyramid token merging and KV cache compression. *arXiv preprint arXiv:2509.00419*.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. 2024. IVTP: Instruction-guided visual token pruning for large vision-language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 214–230.

Drew A. Hudson and Christopher D. Manning. 2019. 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 (CVPR)*, pages 6700–6709.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Youngeun Kim, Youjia Zhang, Huiling Liu, Aecheon Jung, Sunwoo Lee, and Sungeun Hong. 2025. Training-free token pruning via zeroth-order gradient estimation in vision-language models. *arXiv preprint arXiv:2509.24837*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Xueqi Li, Gao Cong, Guoqing Xiao, Yang Xu, Wenjun Jiang, and Kenli Li. 2024a. On evaluation metrics for diversity-enhanced recommendations. In *Proceedings of the 33th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1286–1295.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024b. LLaMA-VID: An image is worth 2 tokens in large language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 323–340.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2025. Mini-Gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pages 740–755.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024e. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 216–233.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 35:2507–2521.

Guangtao Lyu, Xinyi Cheng, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Fen Fang, and Cheng Deng. 2026a. Towards interpretable hallucination analysis and mitigation in lvlms via contrastive neuron steering. *arXiv preprint arXiv:2602.00621*.

Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. 2026b. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in lvlms. *arXiv preprint arXiv:2602.15556*.

Guangtao Lyu, Chenghao Xu, Jiexi Yan, Muli Yang, and Cheng Deng. Towards unified human motion-language understanding via sparse interpretable characterization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Junpeng Ma, Qizhe Zhang, Ming Lu, Zhibin Wang, Qiang Zhou, Jun Song, and Shanghang Zhang. 2026. MMG-Vid: Maximizing marginal gains at segment-level and token-level for efficient video LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. 2025. Mobile edge intelligence for large language models: A contemporary survey. *IEEE Communications Surveys & Tutorials*, pages 3820–3860.

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

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. 2025a. mDP3: A training-free approach for list-wise frame selection in video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24090–24101.

Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Theresa Batista-Navarro. 2025b. LVPruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4299–4308.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. 2025. Stop looking for important tokens in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 9961–9980.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and 1 others. 2025. PyramidDrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chenghao Xu, Jiexi Yan, Muli Yang, Fen Fang, Huilin Chen, and Cheng Deng. 2026. Editing is a bargaining game: Balanced knowledge editing in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pages 34097–34105.

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

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. 2025. Atp-llava: Adaptive token pruning for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24972–24982.

Quan-Sheng Zeng, Yunheng Li, Qilong Wang, Peng-Tao Jiang, Zuxuan Wu, Ming-Ming Cheng, and Qibin Hou. 2025. A glimpse to compress: Dynamic visual token pruning for large vision-language models. *arXiv preprint arXiv:2508.01548*.

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

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024. [CLS] attention is all you need for training-free visual token pruning: Make VLM inference faster. *arXiv preprint arXiv:1503.06733*.

Renshan Zhang, Rui Shao, Gongwei Chen, Miao Zhang, Kaiwen Zhou, Weili Guan, and Liqiang Nie. 2025b. FALCON: Resolving visual redundancy and fragmentation in high-resolution multimodal large language models via visual registers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23530–23540.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2025c. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42th International Conference on Machine Learning (ICML)*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. 2025. Don't just chase" highlighted tokens" in MLLMs: Revisiting visual holistic context retention. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

A Detailed Experiment Settings

A.1 Reproducibility

To achieve disentanglement of visual token features, we construct a sparse interpretable concept space with a latent dimension of $M = 16D$, where each dimension represents a specific semantic concept. We employ a Sparse Autoencoder (SAE) architecture consisting of a single-layer encoder and decoder, trained on the VQA v2 (Goyal et al., 2017) validation set. During training, all LVLM parameters remain frozen. The process is completed in approximately 24 hours on a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GB), with a configuration of 1 epoch, a batch size of 64, and a learning rate of 5×10^{-4}. To ensure highly sparse concept generation, an L_1 penalty with weight $\lambda = 0.5$ is applied. It should be noted that the SAE is specifically optimized for the frozen hidden states of the backbone; thus, a retraining process is required if the LVLM architecture changes. At inference, a significance threshold of $\tau = 1.0$ is applied to filter noisy activations, ensuring that pruning is guided solely by the most salient concepts. All experimental tasks, including training and evaluation across multiple benchmarks, were conducted on a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GB).

A.2 Evaluation Benchmarks

VQA v2 (Goyal et al., 2017): VQAv2 evaluates open-ended visual recognition using 265,016 images from MSCOCO (Lin et al., 2014). It features an adversarially balanced design, ensuring each question corresponds to at least two images with different answers to reduce statistical bias. We evaluate on the test-dev set (107,394 pairs) using 10 ground-truth answers per question and standard automated metrics.

GQA (Hudson and Manning, 2019): The GQA benchmark evaluates structured understanding and reasoning in visual scenes. It comprises three components: scene graphs, questions, and images. Derived from the Visual Genome dataset (Krishna et al., 2017), the scene graphs provide structured descriptions of objects, attributes, and relations, while the image component includes spatial and object-level features. Questions are generated via a semantic engine based on scene graphs to ensure clear reasoning paths. We evaluate our method on the 12,578 samples of the "test-dev" subset, using accuracy as the primary metric.

ScienceQA (SQA^{IMG}) (Lu et al., 2022): This benchmark evaluates zero-shot generalization on scientific topics across three subjects: natural, language, and social sciences. The questions are hierarchically organized across 26 topics. For evaluation, we utilize the SQA-IMG subset, which consists of 2,017 image-question pairs.

TextVQA (Singh et al., 2019): TextVQA assesses the capability to recognize and reason over textual information embedded in diverse real-world scenarios, such as signs and packaging. It challenges models to integrate Optical Character Recognition (OCR) with natural language understanding to perform contextual reasoning. The benchmark provides reference OCR tokens to facilitate text-based answering. We evaluate performance on the standard validation set of 5,000 samples, reporting accuracy.

VizWiz (Bigham et al., 2010): The VizWiz benchmark evaluates a model's visual understanding using real-world images captured by blind users. Due to the uncontrolled capture environment, images often present challenges like blur or poor lighting, and some questions may be irrelevant to the image content. Each question is paired with 10 crowdsourced answers for automated evaluation. We report performance on the test-dev set, which includes 8,000 image-question pairs.

POPE (Li et al., 2023): The POPE benchmark evaluates object hallucination by probing models with targeted yes/no questions regarding object existence. It employs metrics such as accuracy, precision, and F1 score across three sampling strategies to quantify the model's susceptibility to hallucination. In our experiments, we report the F1 score on the test split comprising 9,000 samples.

MME (Fu et al., 2025): MME is a comprehensive benchmark evaluating perceptual and cognitive capabilities across 14 subtasks. It covers a wide spectrum of perception tasks, including OCR as well as coarse-grained recognition (e.g., object count, position, and color) and fine-grained identification (e.g., celebrities, landmarks, and artworks). It consists of binary judgment questions, and we report the perception score based on the total 2,374 image-question pairs.

MMBench (Liu et al., 2024e): This benchmark assesses multi-modal capabilities through a three-level hierarchical taxonomy. It spans from fundamental perception and reasoning skills, through 6 specific capabilities, to 20 concrete subtasks. Utilizing a multiple-choice format to ensure rigorous evaluation, we test on both the English version

(4,377 pairs) and the Chinese version (MMBench-CN, 4,329 pairs).

A.3 Generalizability of the Concept Space

To optimize the SAE within the SCTS framework, we utilize the VQAv2 (Goyal et al., 2017) validation set as the primary training corpus. Although this subset consists of approximately 41k images, it encompasses over 214k question-answer pairs characterized by high semantic density and adversarial balancing. This diversity provides a sufficiently dense distribution of visual features, enabling the SAE to learn robust and interpretable latent decompositions.

Our empirical results demonstrate that SCTS exhibits remarkable cross-domain generalizability. Despite being trained on object-centric VQA samples, the framework achieves state-of-the-art performance on benchmarks with significant domain shifts, such as OCR-intensive tasks (TextVQA (Singh et al., 2019), VizWiz (Bigham et al., 2010)) and specialized scientific reasoning (ScienceQA (Lu et al., 2022)). For instance, at a 77.8% compression ratio on LLaVA-1.5-7B, SCTS reaches 57.0% accuracy on TextVQA, significantly outperforming FastV's 50.6%.

This consistent performance gain across disparate domains underscores the robustness of our concept-driven paradigm. It suggests that by decomposing complex visual features into sparse, fundamental semantic units, SCTS captures universal visual primitives that transcend specific training distributions. This enables effective token selection even in specialized out-of-distribution scenarios, confirming SCTS as a robust, general-purpose pruning strategy for LVLMs.

B Complexity and Scalability Analysis

A potential concern regarding the proposed Semantically Complete Token Selection (SCTS) framework is the scalability of the greedy selection process, particularly given the high-dimensional latent space ($M = 16D$) utilized by the SAE. In this section, we provide a detailed analysis of the theoretical complexity and the effective complexity achieved through our sparsity-aware implementation.

B.1 Theoretical Formulation

Let N denote the number of visual tokens, M the dimension of the SAE dictionary, and B the target token budget. A naive implementation of the greedy Maximum Concept Coverage (MCC) algorithm requires recalculating the Marginal Semantic Gain (MSG) for all remaining candidates against all uncovered concepts in each iteration. The worst-case time complexity of such a naive implementation is:

$$\mathcal{O}_{\text{naive}} = \mathcal{O}(B \cdot N \cdot M). \tag{1}$$

Given that M is significantly expanded (e.g., $M = 65536$ for LLaVa-1.5-7B and $M = 57344$ for Qwen2.5-VL-7B), direct computation could theoretically become a bottleneck compared to attention-based ranking methods ($\mathcal{O}(N \log N)$).

B.2 Effective Complexity via Sparsity

However, the theoretical upper bound ignores a critical property of our framework: the high sparsity of SAE features. Controlled by the sparsity coefficient λ and the significance threshold τ, the number of unique concepts activated in any single image is extremely small. To exploit this, we implement an Active Dimension Reduction mechanism prior to the greedy selection loop. This process involves two steps:

1. **Global Filtering:** We first compute a global boolean mask to identify the subset of semantic dimensions that are active in at least one token. Let M_{active} denote the size of this subset. This pre-processing step has a complexity of $\mathcal{O}(N \cdot M)$, but is executed as a highly parallelized element-wise operation on the GPU.

2. **Sparse Iteration:** The subsequent greedy selection operates solely on the reduced feature matrix of size $N \times M_{active}$. Consequently, the iterative complexity reduces to:

$$\mathcal{O}_{\text{effective}} = \mathcal{O}(B \cdot N \cdot M_{active}). \tag{2}$$

B.3 Empirical Efficiency

Empirical statistics on the Qwen2.5-VL-7B model across our evaluation benchmarks indicate that while the total encoded latent dimension is $M = \alpha \times D$ (where $\alpha = 16$, totaling $57,344$), the number of active concepts M_{active} typically resides within a sparse subspace. Taking the POPE benchmark as a representative case due to its sensitivity to object-level feature precision and hallucination, we observe that M_{active} averages approximately $3,889$ (representing less than 7% of M). As illustrated in Figure 8, while M_{active} adaptively scales based on the visual complexity of the scene—ranging from

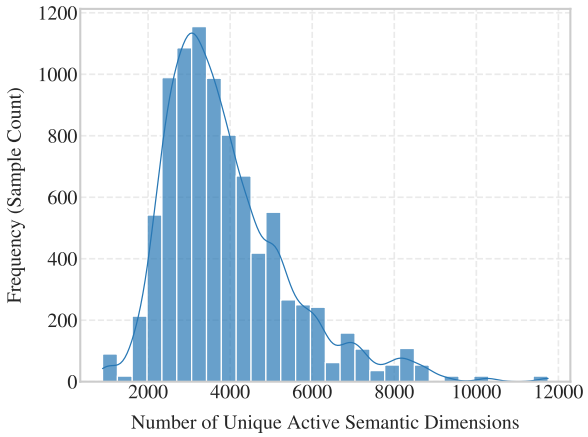


Figure 8: Distribution of the union of active semantic dimensions (M_{active}) per sample on the POPE benchmark.

M	POPE			VQA^{Text}		
	↓ 66.7%	↓ 77.8%	↓ 88.9%	↓ 66.7%	↓ 77.8%	↓ 88.9%
8D	85.2	84.7	81.0	75.3	73.5	69.7
12D	85.9	84.9	81.2	75.8	73.6	70.1
16D	**86.4**	**85.9**	**81.5**	**77.3**	**75.3**	**71.0**
20D	86.5	85.0	81.6	77.5	75.4	71.5
24D	86.3	85.1	81.7	77.3	75.4	71.8

Table 6: Performance comparison across different dictionary dimensions (M) and token compression ratios. M represents the dimension of the high-dimensional sparse features projected by the SAE encoder.

a minimum of 890 to a maximum of $11,749$—it consistently remains an order of magnitude smaller than M. Consequently, the total computational cost T is formulated as:

$$T = \underbrace{\mathcal{O}(N \cdot M)}_{\text{Parallel Pre-filtering}} + \underbrace{\mathcal{O}(B \cdot N \cdot M_{active})}_{\text{Iterative Selection}}. \quad (3)$$

By filtering out over 93% of inactive dimensions on average in the first stage, the burden of iterative selection is substantially reduced. This adaptive mechanism ensures that the selection process does not become a system bottleneck, even in adversarial scenarios requiring high semantic granularity to distinguish real objects from hallucinatory ones. Furthermore, our implementation optimizes this process through bitwise operations and vectorized matrix multiplication, maintaining minimal actual latency as evidenced in Table 5.

C Further Ablation Studies

Analysis of Expansion Factor α. To further investigate the impact of the SAE encoded latent dimension (M) on token selection, we conduct experiments on the Qwen2.5-7B-VL model with various expansion factors $\alpha \in \{8, 12, 16, 20, 24\}$, where $M = \alpha \times D$ and D represents the model's hidden dimension. As summarized in Table 6, increasing the expansion factor α generally enhances performance across both POPE and VQA^{Text} benchmarks, yet the marginal gains diminish as the dictionary scales. Specifically, on the POPE benchmark, increasing the expansion factor α from 8 to 16 yields a significant 1.2% accuracy improvement, whereas further expansion up to 24 only results in negligible gains or slight fluctuations. This suggests that an expansion factor of 16 already provides sufficient capacity to capture the essential visual primitives for this architecture. A similar saturation effect is observed in the text-dense VQA^{Text} task, where accuracy plateaus after α reaches 16. Given that a larger M increases the computational complexity of the parallel pre-filtering stage ($\mathcal{O}(N \cdot M)$), we identify $\alpha = 16$ as the optimal expansion factor that ensures high semantic granularity without compromising inference efficiency.

Analysis of Regularization Weight λ. Table 7 details the impact of the regularization weight λ. The results demonstrate that λ is the core factor for regulating the sparsity of feature activations (L_0 norm), which directly impacts the quality of semantic decoupling. Specifically, when λ is small, the number of activated neurons remains high (resulting in a higher L_0 norm), leading to insufficient sparsity and persistent semantic entanglement between features. As λ increases, the activations become sparser, allowing the decoupled atomic concepts to exhibit stronger monosemanticity. Our experiments indicate that $\lambda = 0.5$ achieves an optimal balance between semantic disentanglement (indicated by Avg. L_0) and downstream task performance.

Table 7: Ablation of λ on Qwen2.5-VL-7B (Pruning Ratio 66.7%).

λ	Avg. L_0	POPE	VQAText
0.05	7238	86.0	69.7
0.1	5961	86.2	69.8
0.3	4573	86.2	77.1
0.5	**3889**	**86.4**	**77.3**
0.7	3247	86.3	77.2
0.9	2594	86.1	77.2

Fine-grained Sensitivity Analysis of Hyperparameter τ. Table 8 presents the fine-grained ablation results for the semantic filtering thresh-

old τ on Qwen2.5-VL-7B (at a 66.7% pruning ratio), with a step size of 0.1. The results indicate that within the broad interval of $\tau \in [0.7, 1.3]$, the POPE metric fluctuates by only 0.2% and VQA-Text by only 0.8%, demonstrating the exceptional robustness of SCTS to the parameter τ. Consistent with the analysis in Figure 3(b) of the main paper, τ serves as a core parameter for controlling the intensity of semantic filtering, exhibiting a distinct performance window. Specifically, when τ is too small, the low filtering threshold introduces redundant background noise features. Conversely, when τ is excessively large (exceeding 1.3), overly aggressive filtering leads to the loss of critical sparse atomic semantics. Experimental results confirm that $\tau = 1.0$ represents the optimal balance between "preserving key semantics" and "suppressing noise".

Table 8: Fine-grained Ablation of τ.

Threshold τ	POPE	VQAText
0.7	85.7	76.5
0.8	86.0	76.9
0.9	86.2	77.2
1.0	**86.4**	**77.3**
1.1	86.4	77.1
1.2	86.3	76.7
1.3	86.4	76.5

D Analysis of Training Overhead and Inference Efficiency

In this section, we provide a comprehensive analysis of both the training costs and inference efficiency of SCTS.

D.1 Training and Storage Overhead

To accelerate the training process, we implemented several key optimizations. Throughout the training phase, the LVLM backbone remains frozen, with updates applied only to the lightweight SAE module. Furthermore, to optimize the pipeline, the forward pass is truncated immediately after the third layer once the SAE loss is computed, which significantly reduces the computational burden.

We conducted an experimental analysis under a consistent hardware environment (a single NVIDIA RTX PRO 6000 Blackwell GPU with 96GB memory) to evaluate the accessibility and cost of SAE training across different LVLMs. The results, including training time and memory footprint, are summarized in Table 9.

Table 9: Resource Consumption across various LVLMs (VQAv2 Dataset, Expansion $\alpha = 16$).

Model	Training Time	Batch Size	GPU Memory (Training)
LLaVA-v1.5-7B	3.8h	64	38.6 GB
LLaVA-NeXT-7B	13.5h	16	58.4 GB
Qwen2.5-VL-7B	23.5h	48	70.8 GB

The results indicate that the training cost of SCTS is highly acceptable relative to the benefits gained. This training process is essential, as its core objective lies in the model's self-exploration and semantic decoupling of its internal hidden features. With minimal training overhead, SCTS effectively transforms semantically entangled features into monosemantic atomic concepts. This transformation not only mitigates performance loss at high compression ratios and enhances inference efficiency but also significantly improves model interpretability, providing a theoretical foundation for interpretable pruning in large-scale models.

D.2 Detailed Latency Breakdown and Scalability Analysis

To provide a transparent view of the end-to-end efficiency, we measured the detailed runtime breakdown of LLaVA-NeXT-7B processing a large-scale input of 2880 tokens on a single NVIDIA RTX 6000 Ada/Blackwell GPU (96GB).

Although the MSG-based token selection introduces a latency of 10.2 ms, this overhead accounts for only 6.5% of the total inference time. By incurring this minor computational cost, we successfully reduced the workload of the subsequent LLM layers, saving 141 ms in end-to-end latency. Even under the extreme scenario of an 88.9% pruning ratio, SCTS maintains 97.5% of its original performance. This demonstrates that the token selection cost is highly acceptable and provides a superior performance-to-cost ratio for practical deployment. By significantly reducing the number of tokens processed by the LLM backbone, SCTS not only lowers individual request latency but also enhances potential system throughput, making it highly suitable for large-scale vision-language tasks. The specific results are summarized in Table 10.

D.3 Latency Analysis of Batch Inference and Streaming Scenarios.

We evaluate SCTS in high-throughput scenarios to demonstrate its real-world utility. Extensive batch inference experiments were conducted on

Table 10: Detailed Latency Breakdown and Efficiency Comparison.

Model	Tokens	SAE Encoding Latency	Token Selection Latency	Total Latency	Speedup	Accuracy
LLaVA-NeXT-7B	2880	0 ms	0 ms	298 ms	1.0x	86.5
SCTS (Ours)	320	1.4 ms	10.2 ms	157 ms	1.9x	84.3

Table 11: Performance Comparison of SCTS with LVPruning.

Method	Retain tokens	GQA	ViViZ	SQA-IMG	VQAText	POPE	MMB-CN	MMB-EN	Average
LLaVA-1.5-7B	576	61.9	50	69.5	58.2	85.9	58.1	64.7	100%
LVPruning	260	59.3	50.8	68.6	57.5	85.8	56.3	63.4	98.4%
SCTS (Ours)	192	59.1	52.1	69.7	57.7	86.1	58.2	65.3	98.8%

a single NVIDIA RTX PRO 6000 Blackwell GPU (96GB) to simulate real-time deployment requirements. As shown in Table 12, SCTS maintains a robust speedup (up to 1.53x) even at a batch size of 8. This stable performance gain is particularly critical for latency-sensitive applications. Given that video sequences can be viewed as continuous batches of frames, the significant reduction in per-frame latency strongly supports the potential of SCTS for real-time video-LMM applications.

Table 12: Latency Comparison on Qwen2.5-VL-7B (POPE Task, 88.9% Pruning Ratio).

Batch Size	Qwen2.5-VL-7B (ms/sample)	SCTS (ms/sample)	Speedup
1	175.7	**109.0**	**1.61x**
4	162.2	**104.5**	**1.55x**
8	152.9	**100.1**	**1.53x**

E Comparison with Existing Training-based Methods

In this section, we present a comprehensive performance comparison between SCTS and representative training-based pruning methods, including LVPruning (Sun et al., 2025b) and ATP-LLaVA (Ye et al., 2025).

Tables 11 and 13 show the performance comparison across multiple datasets. The results demonstrate that, with fewer retained tokens, SCTS achieves superior performance across multiple datasets, with average metrics higher than those of LVPruning and ATP-LLaVA, while maintaining stability across several key indicators. These results strongly demonstrate that SCTS can achieve better and more stable performance with a lower token retention rate, significantly outperforming existing training-based pruning methods. This further confirms the effectiveness of the concept-driven approach introduced in this work.

Table 13: Performance Comparison of SCTS with ATP-LLaVA.

Method	Retain tokens	GQA	SQA-IMG	POPE	Average
LLaVA-1.5-7B	576	61.9	69.5	85.9	100%
ATP-LLaVA	144	59.5	69.1	84.2	97.5%
SCTS (Ours)	128	59.1	69.7	85.6	98.7%
ATP-LLaVA	88	56.8	67.2	82.6	95.0%
SCTS (Ours)	64	58.4	67.9	81.9	95.8%

F SCTS Implementation Pseudo-code

Algorithm 1 Semantically Comprehensive Token Selection (SCTS)

Input: Hidden states $\{\mathbf{z}_i\}_{i=1}^{N}$, SAE parameters, budget B, threshold τ.
Output: Selected index set $\mathcal{S}$.
Initialize: $\mathcal{S} \leftarrow \emptyset$, $\mathcal{P} \leftarrow \{1, \ldots, N\}$, $\mathbf{u}_{\mathcal{S}} \leftarrow \mathbf{0}$.
Generate binary concept vectors $\mathbf{t}_i$ for all $i \in \mathcal{P}$.
Stage 1: Greedy Coverage Optimization
while $|\mathcal{S}| < B$ **do**
  $\Delta\mathcal{G}(p) \leftarrow \|\mathbf{t}_p \odot (\mathbf{1} - \mathbf{u}_{\mathcal{S}})\|_1$
  $p^* \leftarrow \arg\max_{p \in \mathcal{P}} (\Delta\mathcal{G}(p),\ \|\mathbf{t}_p\|_1)$
  if $\Delta\mathcal{G}(p^*) = 0$ **then**
    break
  end if
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{p^*\}, \quad \mathcal{P} \leftarrow \mathcal{P} \setminus \{p^*\}$
  $\mathbf{u}_{\mathcal{S}} \leftarrow \mathbf{u}_{\mathcal{S}} \vee \mathbf{t}_{p^*}$ {Update coverage}
end while
Stage 2: Saliency-based Completion
if $|\mathcal{S}| < B$ **then**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \mathrm{Top}_{B-|\mathcal{S}|}(\{p \in \mathcal{P}\}$ by $\|\mathbf{t}_p\|_1)$
end if
return $\mathcal{S}$

G Additional Visualization Results

As shown in Figure 9, we present additional visual comparisons between HoloV and our method to further demonstrate its superior robustness in high-

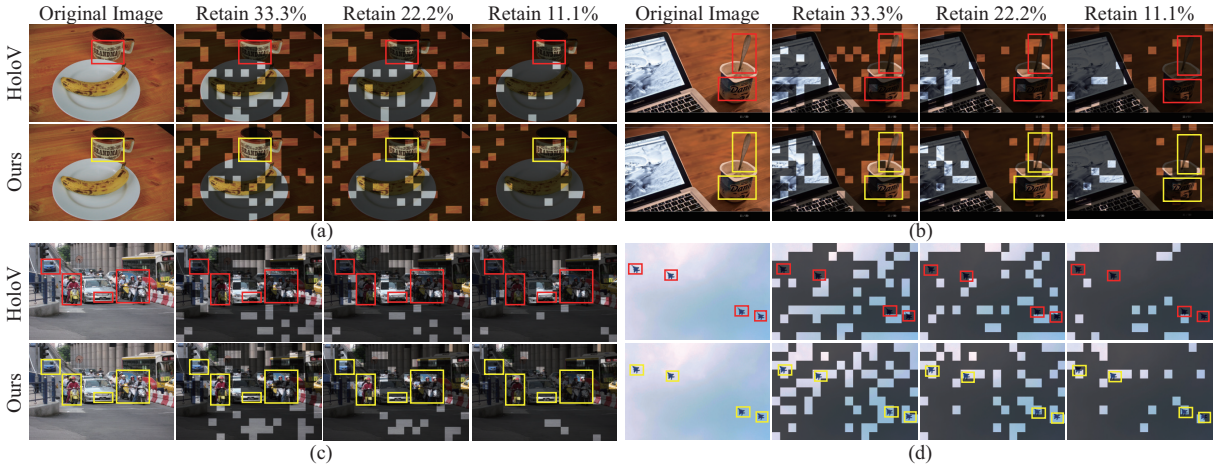


Figure 9: Additional visual comparison between HoloV and our method. It presents original images alongside their pruned versions at retention rates of 33.3%, 22.2%, and 11.1%. The bounding boxes highlight regions with fine-grained text and objects, where our method preserves semantic integrity even under aggressive pruning.

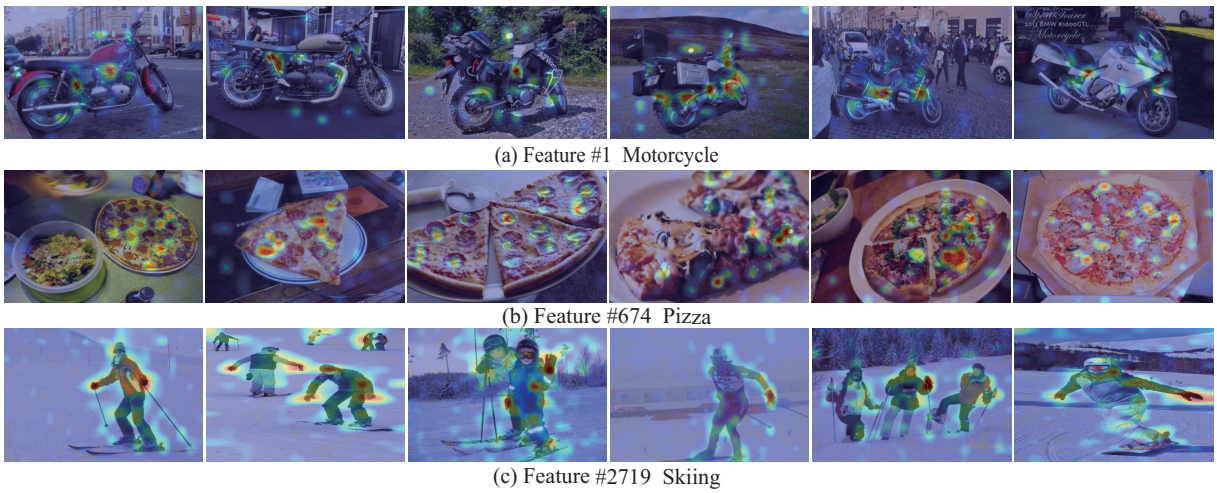


Figure 10: Additional heatmap visualization of learned interpretable features.

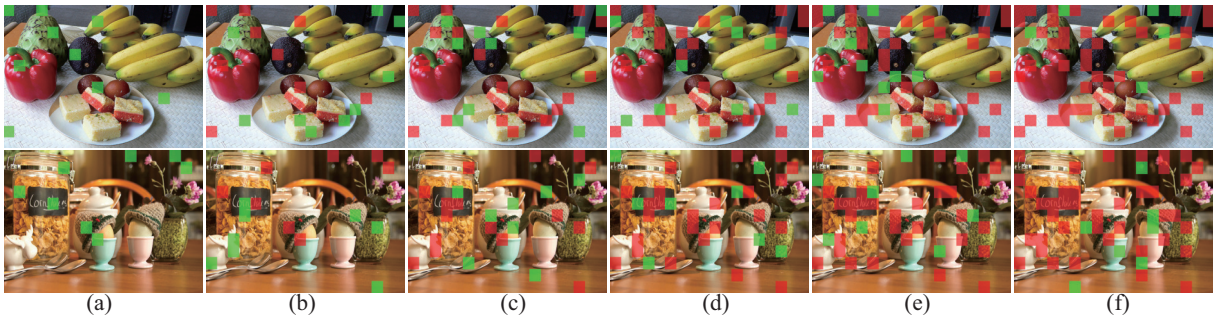


Figure 11: Additional visualization of the progressive token selection process at 77.8% pruning rate. We visualize the selection sequence across 6 steps ((a)→(f)). Green boxes indicate tokens selected at the *current* step, while Red boxes represent the accumulated selections from *previous* steps. As the process advances, our method incrementally captures unique semantic concepts, aiming for comprehensive semantic preservation while minimizing redundancy.

compression scenarios.

As shown in Figure 10, we also provide additional heatmap visualizations of learned interpretable features to further demonstrate that individual dimensions can capture distinct atomic concepts.

As shown in Figure 11, we visualize an additional progressive token selection process at a 77.8% pruning rate to demonstrate that our method effectively mitigates redundancy and to illustrate how semantic diversity is prioritized before refining feature fidelity.