

SCoRe: Saliency-Coverage Reduction for Vision Token Pruning in Vision-Language Models

Tong Xu^{1,2} Hailong Shi^{1,2*} Xingyu Gao^{1,2*}

¹Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China

²University of Chinese Academy of Sciences, Beijing, China

{xutong25, shihailong, gaoxingyu}@ime.ac.cn

Abstract

The heavy computational burden of Large Vision-Language Models (LVLMs) stems primarily from the lengthy visual token sequences generated by their vision encoders. To mitigate this, recent work has shifted towards pruning tokens within the vision encoder. However, we observe that these methods predominantly rely on a suboptimal decoupled heuristic method. This strategy is conceptually flawed: it is prone to sampling collapse, fails to fundamentally eliminate token redundancy, and tends to systematically discard secondary yet important semantic clusters. Addressing this limitation, this paper proposes to formalize visual token pruning as a unified Representativeness Optimization Problem. We introduce **SCoRe** (Saliency-Coverage **Reduction**), a unified optimization method theoretically grounded in the Weighted k -Center Problem. SCoRe constructs the final token set greedily: at each iteration, it selects the token that maximizes the unified representativeness score, thereby optimizing global representativeness. Extensive experiments demonstrate that SCoRe achieves State-of-the-Art (SOTA) performance across multiple benchmarks. Notably, with negligible computational overhead, our method reduces tokens by 94.4% while retaining 95% of the full performance.

1. Introduction

Large Vision-Language Models [3, 6, 21, 25, 27, 42] (LVLMs) have demonstrated exceptional general capabilities in complex visual understanding and reasoning tasks, fundamentally reshaping the field of artificial intelligence. However, a critical design dilemma is inherent in their mainstream architecture: to facilitate linguistic processing of visual input, the Vision Encoder must first grid the image and transform it into a sequence of visual tokens. Crucially, the number of visual tokens significantly surpasses

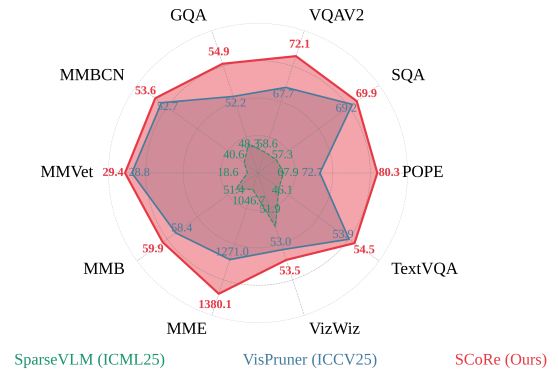


Figure 1. **Performance comparison across ten LVLM benchmarks.** All methods are evaluated under the identical setting of retaining 32 tokens (i.e., a 94.4% pruning rate). SCoRe (red) achieves comprehensive superiority over previous SOTA methods on key benchmarks such as VQA^{Text}, GQA, and MMBench. The performance area of SCoRe (red region) is the largest, verifying the superiority of its unified optimization framework. While pruning 94.4% of tokens, it maintains 95% of the baseline performance, outperforming mainstream methods: SparseVLM [49] (77.9%), VisPruner [47] (91.5%).

that of textual tokens [11, 31, 43]. This sequence length (L) grows rapidly when dealing with challenging tasks such as high-resolution images or long videos. The core bottleneck is not the sequence length itself, but the quadratic computational complexity ($O(L^2)$) intrinsic to the Self-Attention mechanism of the transformer [1, 2, 38, 51]. Consequently, this quadratic complexity makes scaling LVLMs to high-resolution [14, 50] or long-video [8, 20, 24, 33] regimes computationally prohibitive, thus fundamentally hindering their widespread deployment on edge devices and in real-world scenarios [35].

To address this bottleneck, existing visual token pruning methods primarily employ two distinct strategies to evaluate token importance. Early attempts relied on cross-modal attention within the LLM. However, this direction suffers

* Corresponding authors

from fundamental flaws: attention scores in the early layers of the LLM are highly dispersed and severely contaminated by inherent positional biases [47]. These noisy signals fail to accurately reflect the tokens’ true semantic importance. Consequently, recent work [37, 44, 48] has shifted towards leveraging more stable features derived from the Vision Encoder itself.

Despite adopting more reliable feature sources, these methods universally employ a decoupled heuristic strategy. They first select the Top- k visual tokens based on the Vision Encoder’s attention scores; then, in a secondary stage, they attempt to compensate for diversity via post-hoc methods such as clustering or merging. This two-stage heuristic is fundamentally sub-optimal as it inherently neglects optimizing maximum coverage over the semantic token space from the outset. First, the initial Top- k selection systematically ignores minor yet critical semantic clusters, a flaw that subsequent post-hoc compensation cannot effectively rectify. Second, the selected high-importance tokens exhibit high semantic redundancy. As a result, this strategy is highly prone to sampling collapse (*e.g.*, tokens heavily concentrating on the main subject while neglecting the background), yielding a retained subset that fundamentally lacks global coverage of the entire image.

To address these limitations, we formalize the visual token pruning task as a unified representativeness optimization problem. This problem is defined as finding a token subset of size k within the given semantic space that optimally represents the semantic information of the original full set. This novel formalization allows us to establish a theoretical link between token pruning and the classic *Weighted k -Center Problem*.

Based on this theoretical connection, we propose **SCoRe** (Salience-Coverage Reduction), a training-free and computationally efficient algorithm. SCoRe iteratively constructs the representative subset by optimizing a single, unified composite score. This score dynamically balances two factors: the token’s inherent salience and its spatial coverage. This unified optimization process intrinsically suppresses candidate tokens redundant to the selected set, ensuring that the newly added token at each step is the most globally representative point. Simultaneously, it incentivizes the algorithm to select from previously overlooked semantic clusters, thereby theoretically mitigating the sampling collapse phenomenon. Our main contributions are threefold:

- We conceptually formalize visual token pruning as a unified representativeness optimization problem, moving beyond previous heuristic-based methods, and establish its profound theoretical connection to the *Weighted k -Center Problem*.
- We propose a training-free and highly efficient algorithm, SCoRe. SCoRe is the first method to simultaneously optimize salience and coverage through a single, unified com-

posite score, making it conceptually superior to all existing decoupled heuristics.

- Extensive experiments on multiple benchmarks demonstrate that our method consistently surpasses existing state-of-the-art (SOTA) approaches. It achieves a 94.4% token reduction while maintaining 95% of the baseline performance on average with negligible computational overhead.

2. Related Work

2.1. Vision-Language Models (VLMs)

Large Vision-Language Models (LVLMs) extend the reasoning power of Large Language Models (LLMs) to the multimodal domain. Mainstream architectures [4, 9, 30] typically comprise three core components: (1) a pre-trained Vision Encoder [36, 46] to convert the input image into a sequence of visual tokens; (2) a Modal Alignment Module (*e.g.*, MLP [27, 30] or Q-Former [22]); and (3) an LLM Decoder serving as the core reasoning and generation engine. Despite their immense success, these models face a critical computational efficiency bottleneck [10, 39]. The adoption of higher-resolution inputs to enhance visual understanding leads to a sharp increase in the visual token sequence length (L). For instance, moving from LLaVA-1.5 to LLaVA-NeXT increases the sequence length from 576 to 2880 tokens [28, 29]. This excessively long sequence constitutes a critical obstacle to VLM inference efficiency due to the quadratic computational complexity ($O(L^2)$) inherent in the Transformer’s Self-Attention mechanism [39], severely hindering deployment on resource-constrained devices.

2.2. Visual Token Pruning

To mitigate this computational bottleneck, various visual token pruning methods have been proposed. These approaches are broadly categorized by the signal source used to evaluate token importance.

Pruning based on LLM Internal Cross-Attention. Early attempts primarily relied on cross-modal attention scores within the LLM to assess visual token importance [19, 45]. For instance, some methods drop visual tokens after specific LLM layers upon observing that they receive significantly lower attention scores than textual tokens. However, using early-layer cross-modal attention as a pruning criterion is unreliable, as these signals are highly dispersed and inherently biased by positional encodings, preventing an accurate reflection of true semantic salience.

Pruning based on Vision Encoder Features. To circumvent the unreliability of internal LLM signals, recent works have shifted to utilizing earlier, more stable features within the Vision Encoder, making them generally *text-agnostic* [34, 37]. These methods often reduce sequence length by

merging similar tokens. ToMe [5] progressively merges the most similar tokens across ViT blocks, while VisionZip [44] uses a hybrid approach: selecting dominant tokens based on [CLS] attention and merging the remainder.

Although these methods achieve state-of-the-art performance, they universally employ a decoupled heuristic strategy. As discussed, this two-stage strategy is fundamentally sub-optimal, as it inherently neglects optimizing maximum coverage over the global semantic space from the outset, thus rendering it prone to sampling collapse. To address this issue, SCoRe formulates visual token pruning as a unified representativeness optimization problem and proposes an algorithm that balances salience and coverage simultaneously by optimizing a single unified composite score.

3. Motivation

In the Introduction, we argued that existing state-of-the-art (SOTA) pruning methods, especially those relying on the decoupled heuristic strategy, are theoretically sub-optimal. In this section, we provide core visual and analytical evidence to support this argument. Through two key motivations and two corresponding observations, we will establish the design principles that a query-agnostic pruning algorithm must satisfy.

3.1. The Importance of Coverage

To provide analytical evidence at the data level, we must first address a fundamental question: what is the intrinsic feature space structure of visual tokens prior to their ingestion by the LLM? For this purpose, we extract the token features from the second-to-last layer of the Vision Encoder.

Observation 1: Visual tokens are non-uniformly distributed and form discrete semantic clusters. We use UMAP to project these high-dimensional features into a two-dimensional space. As shown in Figure 2a, the visualization reveals a critical insight: tokens are not uniformly distributed but form discrete, high-density semantic clusters. After clustering and coloring with DBSCAN, these clusters clearly divide the feature space into distinct semantic islands. Different semantic entities occupy independent positions in the feature space.

This clustering characteristic reveals a fundamental requirement for pruning algorithms: an ideal algorithm must ensure coverage diversity across these different semantic clusters. If a pruning algorithm only retains tokens from a single cluster, no matter how important that cluster is, the model will lose the remaining semantic information of the image (e.g., completely ignoring secondary subjects), thereby leading to inference failure. This observation demonstrates that any pruning algorithm must incorporate a mechanism to maximize spatial coverage.

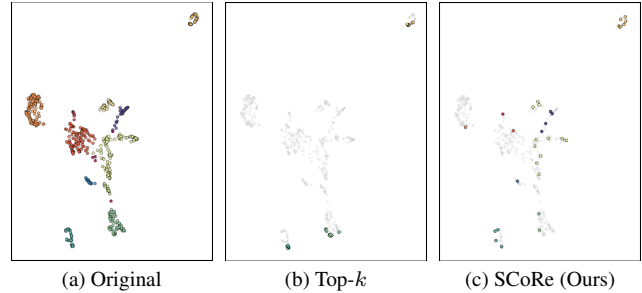


Figure 2. UMAP-based visualization: (a) Original tokens after UMAP dimensionality reduction and DBSCAN clustering with coloring; (b) Tokens retained by the Top- k attention score strategy, where sampling collapse occurs; (c) Tokens retained by the SCoRe algorithm, which achieves higher coverage.

3.2. Tokens Are Not Equally Important

Semantic clustering reveals the necessity of coverage, but it does not address the issue of how to efficiently allocate the pruning budget, as the semantic values of different clusters vary significantly. We measure token salience using [CLS] token attention scores and visualize the selection results of the Top- k strategy.

Observation 2: High-salience regions are highly semantically redundant; oversampling the same cluster ultimately leads to sampling collapse. Figure 2b exposes a salience trap: even within high-salience semantic clusters, a high degree of semantic redundancy exists. When employing the Top- k strategy to prioritize tokens with high attention scores, the retained tokens densely concentrate within a few specific clusters, completely neglecting the majority of other semantic regions. This observation reveals a key flaw: importance does not equate to representativeness. A token with a high attention score may be important, but it may be redundant for the cluster it belongs to. This also explains why greedy strategies relying solely on salience (such as Top- k) are destined to fail: they lead to sampling collapse, i.e., oversampling on a single high-salience cluster, wasting valuable budget that could have been used to cover other secondary clusters.

3.3. Design Principles for a Query-Agnostic Pruning Algorithm

Based on the preceding analysis, we jointly establish two core design principles for a query-agnostic token pruning algorithm:

- **Global Coverage:** The algorithm must sample from all discrete semantic clusters.
- **Salience Priority:** The algorithm must prioritize regions with high information content.

The existing decoupled heuristic strategy inevitably leads to the sampling collapse observed in Section 3.2 (Figure 2b), resulting in a high coverage loss. In contrast,

an ideal pruning algorithm must balance the seemingly conflicting goals of salience and coverage. Our method, SCoRe, is the realization of this unified framework: it automatically allocates tokens to the high-salience primary and secondary subjects simultaneously, while reducing the coverage loss, as evidenced in Figure 2c.

4. Methodology

In this section, we propose SCoRe, a unified framework grounded in classic k -Center theory. SCoRe formalizes token pruning as a representativeness optimization problem, thereby offering a principled solution that dynamically balances both coverage and salience.

4.1. Formalizing the Pruning Problem

In Sec. 3, our analysis established that an ideal pruning algorithm must simultaneously satisfy the twin objectives of global coverage and salience priority. Moving beyond these findings, we now formalize visual token pruning as a unified optimization problem.

Definition 1: Representativeness optimization problem Given:

- Let $V = \{v_1, v_2, \dots, v_L\}$ be the complete set of L visual tokens extracted from the Vision Encoder (where $v_i \in \mathbb{R}^D$ is a D -dimensional feature vector), and let k ($k \ll L$) be the target subset size.
- Let $w_i \geq 0$ be the salience score (weight) of token v_i (e.g., its [CLS] attention score) quantifying its intrinsic importance, and let $d(v_i, v_j)$ be the distance between two tokens (e.g., Euclidean or cosine distance).
- A representativeness loss function $\mathcal{R}(S, V)$, which quantifies the lack of representativeness of subset S with respect to the complete set V .

Objective:

Find a subset $S^* \subset V$, such that $|S^*| = k$, and S^* minimizes the representativeness loss:

$$S^* = \arg \min_{S \subset V, |S|=k} \mathcal{R}(S, V) \quad (1)$$

From Representativeness to the Weighted k -Center Problem

Based on our analysis in Section 3.3, an effective loss function must intrinsically balance a token v_i 's salience denoted as w_i and its semantic distance to the subset S denoted as $d(v_i, S)$. We establish the core premise that a representative subset S should ensure any token v_i in the original set V satisfies one of the following conditions: (1) It is semantically close to S , meaning $d(v_i, S)$ is small; (2) It is intrinsically unimportant, meaning w_i is low.

Thus, the representativeness loss incurred by the worst-represented point can be defined as the maximum weighted

Algorithm 1 SCoRe (Salience-Coverage Reduction)

Require: Visual tokens $V = \{v_1, \dots, v_L\}$, salience scores $W = \{w_1, \dots, w_L\}$, target size k , hyperparameter α

Ensure: Representative indices S_{idx}

```

1:  $s^* \leftarrow \arg \max(W)$ 
2:  $S_{\text{idx}} \leftarrow \{s^*\}$ 
3:  $d_{\min} \leftarrow [d_{\cos}(v_i, v_{s^*}) \mid \forall v_i \in V]$ 
4: while  $|S_{\text{idx}}| < k$  do
5:    $\text{comp\_s} \leftarrow d_{\min} \odot (W^\alpha)$ 
6:    $\text{comp\_s}[S_{\text{idx}}] \leftarrow -\infty$ 
7:    $i^* \leftarrow \arg \max(\text{comp\_s})$ 
8:    $S_{\text{idx}} \leftarrow S_{\text{idx}} \cup \{i^*\}$ 
9:    $d_{\text{new}} \leftarrow [d_{\cos}(v_i, v_{i^*}) \mid \forall v_i \in V]$ 
10:   $d_{\min} \leftarrow \min(d_{\min}, d_{\text{new}})$ 
11: end while
return  $S_{\text{idx}}$ 

```

distance:

$$\mathcal{R}(S, V) = \max_{v_i \in V} (w_i \cdot d(v_i, S)) \quad (2)$$

Substituting this loss function into the objective from Section 4.1, our optimization goal becomes:

$$S^* = \arg \min_{S \subset V, |S|=k} \left(\max_{v_i \in V} (w_i \cdot d(v_i, S)) \right) \quad (3)$$

Corollary: This problem is formally equivalent to the classic *Weighted k -Center Problem* [16].

This theoretical connection provides us with a firm foundation, allowing us to move beyond previous heuristic methods. Although this problem is NP-hard, it admits efficient greedy approximation algorithms. Inspired by this, we propose an algorithm that efficiently applies this greedy framework to visual token pruning, used to iteratively construct the representative subset.

4.2. SCoRe (Salience-Coverage Reduction)

Based on the theoretical framework established in Section 4.1, we formally introduce SCoRe, an efficient algorithm tailored for the visual token pruning task.

SCoRe's iterative strategy (Algorithm 1) strictly follows the greedy criterion derived in Section 4.1. The algorithm proceeds as follows: First, select the token s_1 with the highest salience score from the original set V as the initial representative point. Next, iterate $k - 1$ times: In each step t (from 2 to k), evaluate all unselected tokens $v_i \in V \setminus S_{t-1}$ and calculate their unified representativeness score—the product of two key components: the token's minimum cosine distance to the current selected set S_{t-1} and its own salience score raised to the power of α . Finally, the token with the highest unified score is chosen as the new representative point s_t and added to S_t .

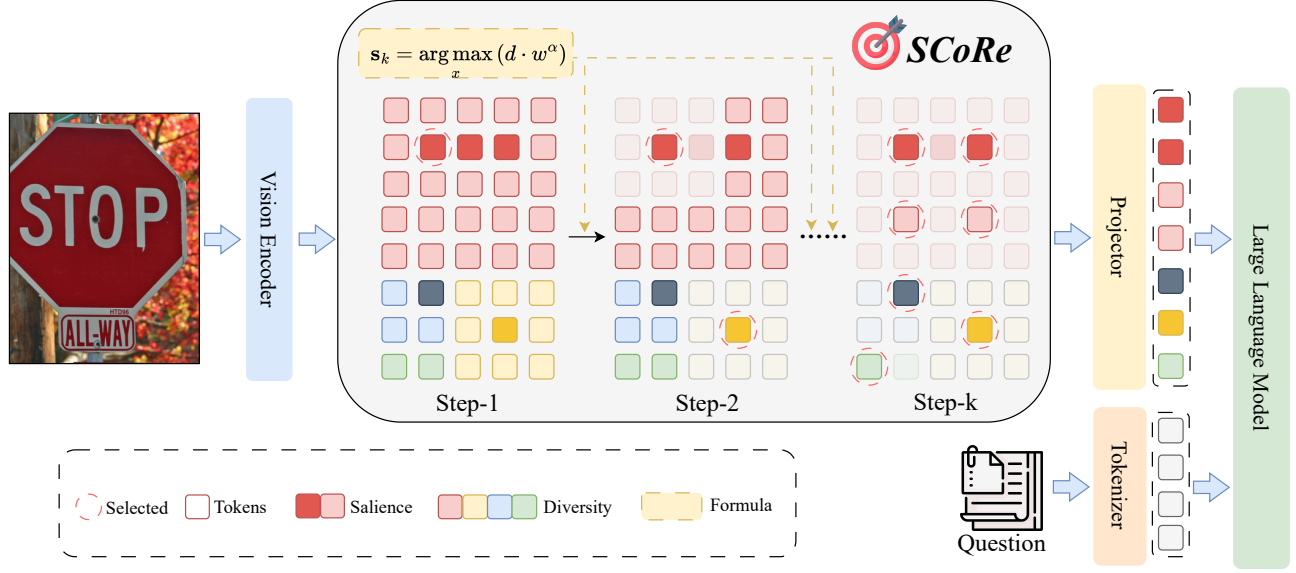


Figure 3. **SCoRe’s visual token pruning process.** The diagram illustrates how SCoRe iteratively selects representative tokens by balancing saliency (from [CLS] attention) and coverage (via cosine distance), integrating into a VLM pipeline where pruned tokens are passed to the Projector and LLM for downstream inference.

This iterative process is summarized by the SCoRe core formula:

$$s_t = \arg \max_{v_i \in V \setminus S_{t-1}} (d_{\cos}(v_i, S_{t-1}) \cdot (\text{score}(v_i))^\alpha) \quad (4)$$

Where the key symbols are defined as follows:

- $d_{\cos}(v_i, S)$ (Cosine Distance): It is the distance measure for semantic difference, defined as the minimum distance from token v_i to the selected set S : $d_{\cos}(v_i, S) = \min_{s \in S} d_{\cos}(v_i, s)$, where $d_{\cos}(v_i, s) = 1 - \text{cosine similarity}(v_i, s)$.
- $\text{score}(v_i)$ (Saliency Score): This score quantifies the intrinsic semantic importance of token v_i . It is defined as the attention score of token v_i to the [CLS] token.
- α (Balance Hyperparameter): A hyperparameter that balances the coverage/diversity objective.

The combination ensures that a token must possess both high saliency ($\text{score}(v_i)$) and high coverage ($d_{\cos}(v_i, S_{t-1})$) to be selected. This fundamentally prevents sampling collapse: selecting a point from a high-saliency cluster immediately decreases the $d_{\cos}(v_i, S_{t-1})$ score of nearby points, making them less attractive in subsequent iterations.

4.3. SCoRe Integration in VLMs

SCoRe operates as a training-free, plug-and-play module embedded between the Vision Encoder and the Modality Projector. The complete pipeline is shown in Figure 3. During inference, the input image I is first processed by the Vision Encoder to simultaneously output two key tensors: L original visual token features $V = \{v_1, \dots, v_L\}$

and their associated saliency weights $W = \{w_1, \dots, w_L\}$. These two tensors (V and W) are fed into the SCoRe algorithm, which efficiently selects k representative indices $S_{\text{idx}} = \{s_1, \dots, s_k\}$. The compact token subset V_S , selected via indices S_{idx} , is subsequently projected by the Modality Projector into the LLM’s textual semantic space, yielding F_S . Next, these projected visual tokens F_S are concatenated with the text tokens T derived from the user query, forming the unified multimodal input sequence $Z = [F_S; T]$. Finally, this merged sequence Z is fed into the LLM Decoder to generate the final response A .

5. Experiments

In this section, we validate our method across various VLM architectures on comprehensive multi-modal tasks, including high-resolution image and video understanding. We compare our approach with multiple existing methods and conduct an ablation study and an efficiency analysis.

5.1. Experimental setup

Datasets We conduct comprehensive experiments across 13 multi-modal benchmarks, encompassing both image-based and video-based tasks. For image-centric evaluations, we select 10 widely adopted benchmarks: common visual question answering (VQA) tasks including VQAv2 [13], GQA [17], VizWiz [15], ScienceQA-IMG, and TextVQA; as well as general multi-modal benchmarks such as POPE [23], MME [12], MMBench, MMBench-CN, and MM-Vet. For video understanding, we evaluate

Method	VQA ^{V2}	GQA	VizWiz	SQA ^{IMG}	VQA ^{Text}	POPE	MME	MMB	MMB ^{CN}	MMVet	Acc.	Rel.
<i>Upper Bound, All 576 Tokens (100%)</i>												
LLaVA-1.5-7B	78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	31.1	63.1	100.0%
<i>Retain 128 Tokens (↓ 77.8%)</i>												
<i>Cross-Attention</i>												
FastV [7] (ECCV24)	61.8	49.6	51.3	60.2	50.6	59.6	1208.9	56.1	51.4	28.1	52.9	85.4%
SparseVLM [49] (ICML25)	73.8	56.0	51.4	67.1	54.9	80.5	1376.2	60.0	51.1	30.0	59.4	94.4%
<i>Vision Feature</i>												
ToMe [5] (ICLR23)	63.0	52.4	50.5	59.6	49.1	62.8	1088.4	53.3	48.8	27.2	52.1	83.9%
VisionZip [44] (CVPR25)	75.6	57.6	52.0	68.9	56.8	83.2	1432.4	62.0	56.7	32.6	61.7	98.4%
VisPruner [47] (ICCV25)	75.8	58.2	52.7	69.1	57.0	84.6	1461.4	62.7	57.3	33.7	62.4	99.6%
SCoRe (Ours)	76.4	59.3	53.7	70.1	57.2	87.3	1451.2	62.6	57.8	33.7	63.1	100.6%
<i>Retain 64 Tokens (↓ 88.9%)</i>												
<i>Cross-Attention</i>												
FastV [7] (ECCV24)	55.0	46.1	50.8	51.1	47.8	48.0	1019.6	48.0	42.7	25.8	46.6	75.9%
SparseVLM [49] (ICML25)	68.2	52.7	50.1	62.2	51.8	75.1	1221.1	56.2	46.1	23.3	54.7	86.4%
<i>Vision Feature</i>												
ToMe [5] (ICLR23)	57.1	48.6	50.2	50.0	45.3	52.5	922.3	43.7	38.9	24.1	45.7	73.9%
VisionZip [44] (CVPR25)	72.4	55.1	52.9	69.0	55.5	77.0	1365.6	60.1	55.4	31.7	59.7	95.6%
VisPruner [47] (ICCV25)	72.7	55.4	53.3	69.1	55.8	80.4	1369.9	61.3	55.1	32.3	60.4	96.6%
SCoRe (Ours)	74.7	56.9	53.5	69.9	56.5	83.1	1414.1	60.2	55.6	32.0	61.3	97.8%
<i>Retain 32 Tokens (↓ 94.4%)</i>												
<i>Cross-Attention</i>												
FastV [7] (ECCV24)	43.4	41.5	51.7	42.6	42.5	32.5	884.6	37.8	33.2	20.7	39.0	64.1%
SparseVLM [49] (ICML25)	58.6	48.3	51.9	57.3	46.1	67.9	1046.7	51.4	40.6	18.6	49.3	77.9%
<i>Vision Feature</i>												
ToMe [5] (ICLR23)	46.8	43.6	51.3	41.4	38.3	39.0	828.4	31.6	28.1	17.3	37.9	61.4%
VisionZip [44] (CVPR25)	67.1	51.8	52.9	68.8	53.1	68.7	1247.4	57.7	50.3	25.5	55.8	89.0%
VisPruner [47] (ICCV25)	67.7	52.2	53.0	69.2	53.9	72.7	1271.0	58.4	52.7	28.8	57.2	91.5%
SCoRe (Ours)	72.1	54.9	53.5	69.9	54.5	80.3	1380.1	59.9	53.6	29.4	59.7	95.0%

Table 1. **Performance comparison of different pruning methods on LLaVA-1.5-7B.** Here, **Acc.** denotes the average accuracy across 10 benchmarks, and **Rel.** represents the average percentage of performance maintained at the corresponding reduction ratio.

on 3 standard video question answering (Video-QA) benchmarks: TGIF-QA [18], MSVD-QA [40], and MSRVT-QA [40]. All experiments adhere to the default settings and evaluation metrics of each benchmark, with detailed task descriptions provided in the supplementary material.

Model architectures We evaluate SCoRe on diverse VLM architectures to demonstrate its generalizability: the LLaVA series (LLaVA-1.5 [28], LLaVA-NeXT [29] for high-resolution inputs), Video-LLaVA [26] for video understanding, and Qwen-VL [4]—a widely used general-purpose VLM. For consistency, we adopt the inference configurations specified in the original works of all evaluated models.

Comparison methods We compare SCoRe with five state-of-the-art (SOTA) token pruning methods, covering diverse pruning paradigms targeting different stages of the VLM pipeline: ToMe [5], FastV [7], SparseVLM [49], VisionZip [44], and VisPruner [47]. These methods represent the latest advances in visual token reduction, ensuring a comprehensive comparison of SCoRe’s performance.

5.2. Main results

We first apply SCoRe to the classic LLaVA-1.5 model and conduct a comprehensive comparison with existing methods. Tab. 1 shows the performance of different pruning methods on the LLaVA-1.5-7B model when retaining only 128, 64, and 32 tokens. With **77.8%** of the vi-

sual tokens pruned (retaining 128 tokens), SCoRe remarkably maintained **100.6%** of the original performance, effectively achieving the upper bound. When the number of retained tokens is further reduced to 64 (pruning rate 88.9%), SCoRe only decreases the original performance by **2.2%** (Rel. 97.8%), significantly outperforming the best cross-attention-based method, SparseVLM (Rel. 86.4%), by as much as **11.4%**. It also surpasses the latest vision-based pruning method, VisionZip (Rel. 95.6%). This result suggests that simultaneously optimizing salience and coverage for token selection is more crucial than merging similar tokens. Even when retaining only about **5%** of the tokens (32 tokens, pruning rate 94.4%), SCoRe is able to maintain **95.0%** of the performance, proving the effectiveness of its design that balances salient tokens and diverse semantic clusters as complements.

5.3. Extending SCoRe to Other VLM Frameworks

SCoRe with higher resolution. Some works [29, 32, 41] attempt to improve VLM performance on visual question answering tasks by increasing the input image resolution. However, this resolution increase introduces much more visual tokens, further intensifying the computational load on the VLM. In this section, we apply SCoRe to LLaVA-NeXT-7B, which can handle up to 2880 visual tokens. The results are presented in Tab. 2. Compared to LLaVA-1.5,

Method	VQA ^{V2}	GQA	VQA ^{Text}	POPE	MME	Acc.	Rel.
<i>Upper Bound, All 2880 Tokens (100%)</i>							
LLaVA-NeXT-7B	81.2	62.9	59.6	86.3	1513.8	73.1	100.0%
<i>Retain 640 Tokens (↓ 77.8%)</i>							
<i>Cross-Attention</i>							
FastV [7] (ECCV24)	78.9	60.4	58.4	83.1	1477.3	70.9	97.0%
SparseVLM [49] (ICML25)	78.2	59.1	56.2	80.9	1456.3	69.4	94.9%
<i>Vision Feature</i>							
VisionZip [44] (CVPR25)	79.2	60.1	58.5	82.2	1468.4	70.7	96.7%
VisPruner [47] (ICCV25)	79.8	61.6	59.3	85.9	1480.7	72.1	98.6%
SCoRe (Ours)	80.0	61.9	59.9	86.2	1475.3	72.4	99.1%
<i>Retain 320 Tokens (↓ 88.9%)</i>							
<i>Cross-Attention</i>							
FastV [7] (ECCV24)	71.9	55.9	55.7	71.7	1282.9	63.9	87.7%
SparseVLM [49] (ICML25)	71.4	56.5	52.4	73.5	1342.7	62.2	87.9%
<i>Vision Feature</i>							
VisionZip [44] (CVPR25)	74.2	58.1	55.3	75.0	1348.8	66.0	90.5%
VisPruner [47] (ICCV25)	75.7	58.4	57.6	80.4	1370.1	68.1	93.3%
SCoRe (Ours)	78.2	60.1	59.3	85.2	1403.1	70.6	96.6%
<i>Retain 160 Tokens (↓ 94.4%)</i>							
<i>Cross-Attention</i>							
FastV [7] (ECCV24)	61.8	49.8	51.9	51.7	1079.5	53.8	74.7%
SparseVLM [49] (ICML25)	62.2	50.2	45.1	54.6	1167.1	54.1	74.5%
<i>Vision Feature</i>							
VisionZip [44] (CVPR25)	67.3	54.3	54.7	59.4	1239.7	59.5	82.3%
VisPruner [47] (ICCV25)	70.6	54.7	56.0	72.9	1226.0	63.1	86.7%
SCoRe (Ours)	72.6	58.6	57.5	82.1	1395.8	68.1	93.4%

Table 2. **Performance comparison of different pruning methods on LLaVA-NeXT-7B.** **Acc.** denotes the average accuracy across benchmarks, and **Rel.** represents the average percentage of performance maintained at the corresponding reduction ratio.

Method	VQA ^{V2}	GQA	VizWiz	SQA ^{IMG}	VQA ^{Text}	Acc.	Rel.
<i>Upper Bound, All 256 Tokens (100%)</i>							
Qwen-VL-7B	78.8	59.3	35.2	67.1	63.8	60.8	100.0%
<i>Retain 128 Tokens (↓ 50%)</i>							
FastV [7] (ECCV24)	76.5	56.9	32.7	65.3	58.2	57.9	94.9%
VisPruner [47] (ICCV25)	77.4	57.8	33.4	65.9	59.6	58.8	96.4%
SCoRe (Ours)	78.0	58.5	34.0	66.5	60.5	59.5	97.9%

Table 3. **Performance of SCoRe on Qwen-VL-7B.** **Acc.** denotes the average accuracy across benchmarks, **Rel.** represents the average percentage of performance maintained.

LLaVA-NeXT involves a greater number of visual tokens, implying a higher degree of redundancy. At a **94.4%** reduction ratio, retaining only **160** visual tokens, SCoRe preserves **93.4%** of the original performance, significantly outperforming FastV (74.7%), SparseVLM (74.5%) as well as VisionZip (82.3%), demonstrating the value of SCoRe for high-resolution visual inputs.

SCoRe with Qwen-VL-7B. To validate the architectural generalizability of our method, we apply **SCoRe** to the **Qwen-VL-7B** model, which employs a Q-Former as a multimodal projector. The experiment benchmarks performance against the Upper Bound (256 tokens, 60.8% overall Acc.) and compares results when retaining only 128 tokens (↓ 50%). SCoRe consistently demonstrates superiority over FastV and VisPruner across all measured benchmarks after the 50% token reduction. Quantitatively, SCoRe achieves 59.5% overall accuracy, representing a 1.6 percentage point improvement over FastV (57.9%). The Relative Retention rate is maximized by SCoRe at 97.9%, significantly higher than FastV (94.9%) and VisPruner (96.4%).

The comprehensive advantage of SCoRe across the Qwen-VL-7B architecture confirms its core principle’s effectiveness: the strategy of balancing salience and coverage

Method	TGIF-QA		MSVD-QA		MSRVTT-QA		Average	
	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
<i>Upper Bound, All 2048 Tokens (100%)</i>								
Video-LLaVA [26]	19.8	2.53	70.5	3.93	57.5	3.50	49.3	3.32
<i>Retain 455 Tokens (↓ 77.8%)</i>								
FastV [7] (ECCV24)	19.2	2.50	69.1	3.91	54.4	3.42	47.6	3.28
VisPruner [47] (ICCV25)	18.4	2.49	70.2	3.95	56.7	3.50	48.4	3.31
SCoRe (Ours)	19.0	2.51	70.6	3.96	57.2	3.51	48.9	3.33
<i>Retain 227 Tokens (↓ 88.9%)</i>								
FastV [7] (ECCV24)	14.3	2.42	68.9	3.90	53.0	3.40	45.4	3.24
VisPruner [47] (ICCV25)	15.9	2.41	69.3	3.92	55.6	3.45	46.9	3.26
SCoRe (Ours)	16.5	2.43	69.8	3.93	56.2	3.47	47.5	3.28
<i>Retain 114 Tokens (↓ 94.4%)</i>								
FastV [7] (ECCV24)	10.6	2.29	64.1	3.78	52.4	3.39	42.4	3.15
VisPruner [47] (ICCV25)	14.1	2.35	65.4	3.79	54.1	3.41	44.5	3.18
SCoRe (Ours)	14.8	2.37	66.0	3.80	54.8	3.43	45.2	3.20

Table 4. Video understanding performance with **Video-LLaVA** across three commonly used video question answering benchmarks.

is a universal and effective method for visual token pruning, transcending diverse VLM projection mechanisms.

5.4. Application in Video Understanding

Video understanding is another scenario with high visual redundancy. We also apply SCoRe to Video-LLaVA [26], which accepts videos as input. Following [33] and [26], we conduct experiments on three video question answering benchmarks, using ChatGPT-Assistant for evaluation. The evaluation results are shown in Tab. 4. Video-LLaVA processes 8 frames of 224-resolution video, totaling **2048** visual tokens. SCoRe maintains strong performance across all reduction ratios, significantly outperforming FastV. This improvement is due to the higher redundancy in temporally continuous video frames compared to single images, and the [CLS] attention accurately identifies key tokens within the video sequence, enabling SCoRe to maintain strong performance even at high reduction ratios.

5.5. Ablation study

To verify the necessity of balancing diversity and salience, we conduct ablation experiments on LLaVA-1.5-7B with 32 retained tokens (Fig. 4).

We evaluate two extremes: Diverse ($\alpha = 0$), a pure diversity method, achieves 53.2% Acc. on TextVQA and 76.2% F1 on POPE, proving diversity alone is inadequate. Important ($\alpha \rightarrow \infty$), a pure salience method, improves to 54.0% Acc. on TextVQA and 77.8% F1 on POPE, yet misses valuable non-salient context.

Our full SCoRe method, by optimally balancing both properties, outperforms both extremes significantly: it achieves 54.5% Acc. on TextVQA and 80.3% F1-Score on POPE. This robust performance demonstrates that the complementary integration of diversity and salience is crucial for maintaining performance under aggressive token pruning.

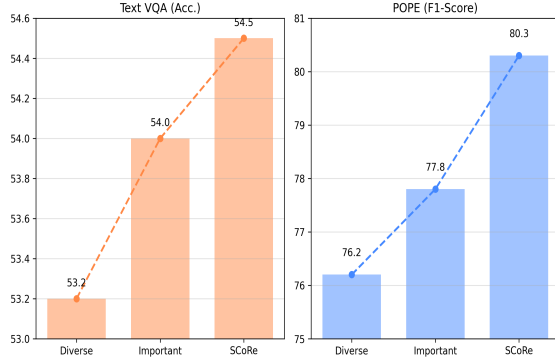


Figure 4. **Ablation study on the balance between diversity and salience.** All experiments retain 32 visual tokens. Diverse corresponds to SCoRe with $\alpha = 0$ (degrading to a pure diversity-based algorithm), Important corresponds to SCoRe with a sufficiently large α (degrading to a pure salience-based algorithm), and SCoRe represents our full method that balances both properties.

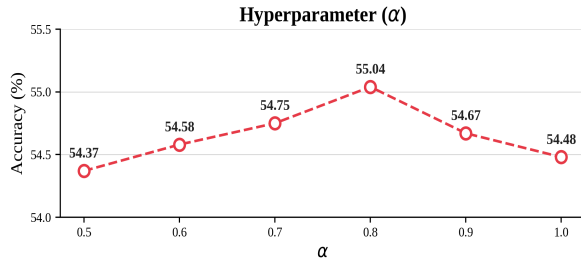


Figure 5. **Sensitivity analysis of hyperparameter α .** The experiment is conducted on the TextVQA with 32 retained tokens, where α controls the trade-off between diversity coverage and salience.

5.6. Hyperparameter Sensitivity (α)

The hyperparameter α in SCoRe controls the trade-off between coverage and salience during token selection. To investigate its sensitivity, we conduct experiments on the TextVQA benchmark with 32 retained tokens, varying α from 0.5 to 1.0 (Fig. 5). The results show that SCoRe achieves stable performance across a wide range of α values, with the highest accuracy (55.04%) observed at $\alpha = 0.8$. When α is too small (0.5 – 0.6), overemphasizing diversity leads to suboptimal performance (54.52% at $\alpha = 0.6$) due to the inclusion of less critical tokens. When α is too large (0.9 – 1.0), overemphasizing salience results in a slight drop (54.48% at $\alpha = 1.0$) as semantic coverage decreases. Notably, the performance remains above 54.5% for $\alpha \in [0.7, 0.9]$, indicating that SCoRe is robust to reasonable variations in α and does not require fine-grained tuning, which enhances its practicality.

Method	# Tokens	FLOPs (T)	Storage (MB)	GPU Memory (GB)	CUDA Time (ms)
LLaVA-NeXT-7B	2880	43.3	1440	16.7	853.2
FastV [7] (ECCV24)	640	13.5	380	16.9	403.4
SCoRe (Ours)	640	4.1	360	14.6	230.5
FastV [7] (ECCV24)	160	6.3	95	16.9	305.3
SCoRe (Ours)	160	3.8	80	14.7	158.3

Table 5. **Performance and Resource Comparison of Different Methods.** # Token denotes the number of retained vision tokens, FLOPs (T) represents tera operations, Storage (MB) is the memory storage overhead, GPU Memory (GB) is the graphics card memory usage, and CUDA Time (ms) is the inference time on CUDA.

5.7. Efficiency analysis

To demonstrate the efficiency of SCoRe, we perform a comparative analysis of FLOPs, cache storage, GPU memory, and CUDA latency with FastV on LLaVA-NeXT-7B. All experiments are performed on a single NVIDIA A100-40GB GPU, evaluated using the POPE benchmark. As shown in Tab. 5, by retaining only 160 tokens, SCoRe achieves significant reduction in inference latency for LLaVA-NeXT. Compared to FastV with the same reduction ratio, SCoRe requires less memory and achieves faster inference. Moreover, the design of pruning before the language model enables compatibility with faster attention implementations like FlashAttention [10], which are infeasible for methods like FastV that require access to the text-visual attention information within the language model.

6. Conclusion

In this paper, we identify a fundamental flaw in existing visual token pruning methods: their reliance on the decoupled heuristic strategy, which leads to sub-optimal sampling collapse. To address this, we propose SCoRe, a training-free token pruning framework. SCoRe abandons heuristics, formalizing pruning as a unified representativeness optimization problem and connecting it to the classic weighted k -center problem. By iteratively optimizing a composite score balancing salience and coverage, SCoRe fundamentally avoids prior limitations. Extensive experiments show SCoRe achieves state-of-the-art performance across benchmarks with enhanced efficiency. We believe SCoRe’s theoretically grounded unified framework offers a principled and scalable direction for designing VLMs to efficiently handle high-resolution images and videos.

Acknowledgments

This work was supported in part by Science and Technology Innovation (STI) 2030—Major Projects under Grant 2022ZD0208700, the National Natural Science Foundation of China under Grant 62376264, and the CAAI—Huawei AI Computing Power Acceleration Program.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 1
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*, 2023. 1
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 6
- [5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3, 6
- [6] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*, 2023. 1
- [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. *arXiv preprint arXiv:2403.06764*, 2024. 6, 7, 8
- [8] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 1
- [9] 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. 2
- [10] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 2, 8
- [11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1
- [12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 5
- [13] 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. 5
- [14] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024. 1
- [15] 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. 5
- [16] Dorit S Hochbaum and David B Shmoys. A unified approach to approximation algorithms for bottleneck problems. *Journal of the ACM (JACM)*, 33(3):533–550, 1986. 4
- [17] 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. 5
- [18] 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. 6
- [19] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 784–794, 2022. 2
- [20] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vignesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv:2312.14125*, 2023. 1
- [21] 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*, 2023. 1
- [22] 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. 2
- [23] 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. 5
- [24] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 1
- [25] 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:2403.18814*, 2024. 1

- [26] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6, 7
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023. 1, 2
- [28] 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. 2, 6
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 6
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2
- [31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1
- [32] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 6
- [33] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1, 7
- [34] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022. 2
- [35] Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. Mobile edge intelligence for large language models: A contemporary survey. *arXiv preprint arXiv:2407.18921*, 2024. 1
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [37] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 2
- [38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1
- [39] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2
- [40] 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. 6
- [41] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 6
- [42] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *CoRR*, abs/2408.10188, 2024. 1
- [43] Panqi Yang, Haodong Jing, Nanning Zheng, and Yongqiang Ma. Unihoi: Unified human-object interaction understanding via unified token space. *arXiv preprint arXiv:2511.15046*, 2025. 1
- [44] 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. 2, 3, 6, 7
- [45] Deming Ye, Yankai Lin, Yufei Huang, and Maosong Sun. Tr-bert: Dynamic token reduction for accelerating bert inference. *arXiv preprint arXiv:2105.11618*, 2021. 2
- [46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2
- [47] Qizhe Zhang, Aosheng Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20857–20867, 2025. 1, 2, 6, 7
- [48] Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025. 2
- [49] 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. 1, 6, 7
- [50] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 1
- [51] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 1