LRCP: LOW-RANK COMPRESSIBILITY GUIDED VISUAL TOKEN PRUNING FOR EFFICIENT LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) achieve strong multimodal understanding, but their inference cost grows rapidly with the number of visual tokens, especially for high-resolution images and long videos. Existing attention-based methods estimate token importance from attention scores, which may introduce positional bias, while representation-based methods reduce visual redundancy based on feature relations or reconstruction errors, overlooking the global structure of the visual token set. In this paper, we revisit visual token compression from the perspective of low-rank compressibility. Across models and datasets, we observe that visual token representations exhibit a pronounced low-rank structure, with a dominant subspace that remains stable even after a large fraction of tokens is randomly removed. Motivated by this finding, we propose LRCP, a training-free compression framework that first estimates the dominant low-rank subspace of visual tokens via PCA, and then scores each token by its projection residual onto this subspace, retaining tokens that are poorly explained by the low-rank background. Extensive experiments and direct patch-level validation show that LRCP preserves 94.7% of the original image-understanding performance with an 88.9% token reduction and 97.8% of the average video-understanding accuracy with an 87.5% token reduction.

1 INTRODUCTION

Large vision-language models (LVLMs) have substantially advanced multimodal understanding (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Dai et al., 2023; Zhu et al., 2024; OpenAI, 2023), with strong results in visual question answering (Antol et al., 2015; Goyal et al., 2017; Singh et al., 2019; Mathew et al., 2021), image captioning (Lin et al., 2014; Young et al., 2014; Agrawal et al., 2019), and video understanding (Li et al., 2024a; Lin et al., 2024; Zhang et al., 2025a). As these models are applied to higher-resolution images (Chen et al., 2024b; Li et al., 2024a; Guo et al., 2025) and long videos (Lin et al., 2024; Zhang et al., 2025a), the number of visual tokens increases rapidly. This growth raises both computation and memory costs during prefilling and also increases inference latency. Reducing visual token overhead has therefore become critical for efficient LVLM deployment (Liu et al., 2025b).

To address this, recent work has explored visual token compression, with existing methods broadly categorized as attention-based (Chen et al., 2024a; Arif et al., 2025; Zhang et al., 2024a; Liu et al., 2025a; Yang et al., 2024) or representation-based (Zhang et al., 2024b; Xing et al., 2024; Zhang et al., 2025b; Ma et al., 2026; Shang et al., 2024; Ye et al., 2025; Wang et al., 2025; Xing et al., 2025). Attention-based methods estimate token importance from attention scores, but they can be affected by positional bias and are less compatible with efficient attention implementations such as FlashAttention (Dao et al., 2022; Dao, 2024). Representation-based methods reduce visual redundancy by exploiting relations among token features, but they mainly capture local relations and often rely on heuristic distance measures. As a result, existing methods do not explicitly model the global structure of the visual token set, which limits their ability to preserve important semantic information under high compression ratios.

This paper studies visual token compression from the perspective of low-rank compressibility. We observe that visual tokens in LVLMs concentrate around a small number of dominant directions in the feature space, forming an approximate low-rank structure. This suggests that, beyond individual token importance or local similarity, visual token redundancy is rooted in the global structure of visual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

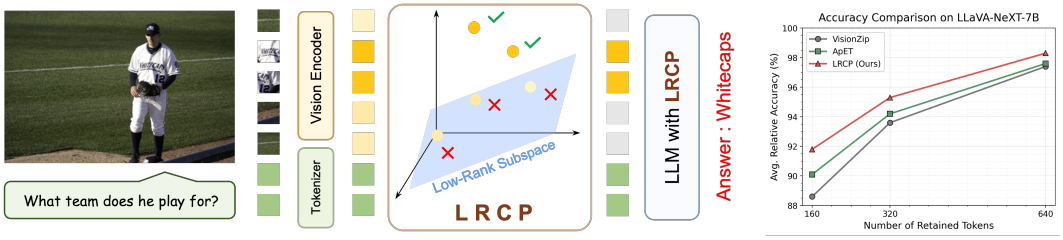

Figure 1: **(Left)** LRCP estimates the dominant low-rank subspace of visual tokens via PCA and selects tokens based on their projection residuals. **(Right)** Accuracy comparison with LLaVA-NeXT-7B across 7 benchmarks, showing that ours outperforms both VisionZip (attention-based) and ApET (representation-based).

representations. Specifically, this global structure can be described by a shared low-rank background. We further find that this dominant subspace remains highly stable even when a large fraction of tokens is randomly removed, indicating that the low-rank structure is a collective property rather than driven by a few specific tokens. Based on these observations, we formulate visual token compression as a lossy compression problem under a low-rank background, where tokens with larger deviations from the dominant subspace are more likely to carry discriminative visual information.

Motivated by these observations, we propose LRCP (Low-Rank Compressibility Guided Pruning), a training-free visual token compression framework. LRCP estimates the dominant low-rank subspace via PCA and scores each token by its projection residual (Figure 1, left). Under the same compression ratio, LRCP outperforms both attention-based and representation-based methods (Figure 1, right). The method can be applied at intermediate layers of either the visual encoder or the LLM. Since it does not require attention weights, it is fully compatible with efficient inference implementations (Dao, 2024; Kwon et al., 2023).

Our main contributions are summarized as follows:

1. We conduct a systematic empirical analysis verifying that the low-rank structure of visual tokens is consistent across models and datasets, with a highly stable dominant subspace.
2. Based on this finding, we formulate visual token pruning as a lossy compression problem under a low-rank background, providing a principled surrogate objective that connects token selection with projection residuals onto the dominant subspace.
3. We propose LRCP, which scores tokens by their projection residuals onto the dominant low-rank subspace. LRCP is training-free and compatible with efficient attention implementations, enabling plug-and-play integration into existing LVLM architectures.
4. Extensive experiments on image and video benchmarks show that LRCP preserves over 94% of the original performance with an 88.9% token reduction, consistently outperforming existing methods under various compression ratios.

2 Related Work

2.1 Large Vision-Language Models

Modern LVLMs couple a vision encoder with a language model via a lightweight projector, converting image patches into visual tokens for autoregressive decoding (Liu et al., 2023; Li et al., 2023a; Dai et al., 2023). Within this paradigm, recent efforts have rapidly scaled the visual token budget: LLaVA-v1.5 (Liu et al., 2024a) encodes a 336×336 image into 576 tokens, while LLaVA-NeXT (Liu et al., 2024b) raises this to several thousand via multi-crop encoding. The same trend extends to native-resolution modeling (Wang et al., 2024; Bai et al., 2025; Guo et al., 2025), multi-image understanding (Zhu et al., 2025; Li et al., 2024a), and long-video reasoning (Lin et al., 2024; Zhang et al., 2025a). Although longer visual sequences improve fine-grained perception, they also make visual tokens the dominant source of prefilling cost and inference latency, motivating effective token compression.

2.2 Token Compression for LVLMs

To reduce the inference cost introduced by visual tokens, recent studies have explored token compression for LVLMs. Existing methods can be grouped into attention-based (Chen et al., 2024a; Arif et al., 2025; Zhang et al., 2024a; Liu et al., 2025a; Yang et al., 2024) and representation-based (Zhang et al., 2024b; Xing et al., 2024; Zhang et al., 2025b; Ma et al., 2026; Shang et al., 2024; Ye et al., 2025; Wang et al., 2025; Xing et al., 2025) approaches. Attention-based methods, such as FastV (Chen et al., 2024a), estimate token importance from language-model attention scores and retain tokens relevant to the query or generation process. Despite their effectiveness, their reliance on attention distributions can introduce positional bias and make them less compatible with efficient attention implementations (Dao et al., 2022; Dao, 2024; Shazeer, 2019). Representation-based methods instead reduce redundancy through feature relations. PyramidDrop (Xing et al., 2024) removes locally similar tokens at successive LLM layers, whereas ApET and ResPrune construct reconstruction subspaces from sampled or progressively selected tokens (Ma et al., 2026; Li et al., 2026). SVD-Prune applies a global SVD but retains high-leverage tokens that preserve dominant variance (Apedo et al., 2026). In contrast, LRCP estimates one PCA subspace from all tokens in the current input, treats this shared structure as reusable background, and retains the residual variations that it cannot explain.

3 Method

3.1 Preliminaries: LVLMs

Most current large vision-language models (LVLMs) follow a ViT-Projector-LLM architecture (Dosovitskiy et al., 2021; Radford et al., 2021; Vaswani et al., 2017; Touvron et al., 2023). Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ or a video $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{T}$, the vision encoder first extracts visual features, yielding image-level embeddings $\mathbf{E} \in \mathbb{R}^{N\times D}$ or video-level embeddings $\mathbf{E} \in \mathbb{R}^{T\times N\times D}$. The projector then maps these visual embeddings into visual tokens aligned with the language space, denoted by $\mathbf{F}^v \in \mathbb{R}^{M\times D'}$ for images and $\mathbf{F}^v \in \mathbb{R}^{T\times M\times D'}$ for videos. These visual tokens are concatenated with text tokens $\mathbf{F}^t$ corresponding to the input instruction and passed into the LLM decoder. The decoder processes the multimodal context in a single prefilling pass, then autoregressively generates the response $\mathbf{Y} = \{y_j\}_{j=1}^{L}$. The conditional generation probability can be written as

$$p(\mathbf{Y} \mid \mathbf{F}^v, \mathbf{F}^t) = \prod_{j=1}^{L} p\left(y_j \mid \mathbf{F}^v, \mathbf{F}^t, \mathbf{Y}_{1:j-1}\right), \tag{1}$$

where $\mathbf{Y}_{1:j-1}$ denotes the sequence of previously generated tokens before step j.

3.2 Low-Rank Structure of Visual Tokens

We characterize the low-rank structure of visual tokens by performing PCA (Jolliffe, 2002) on each layer's output across three models and three datasets. We measure effective dimensionality as the minimum rank required to reach a given fraction of explained variance:

$$\mathrm{Rank}_l@v = \min\left\{ d \in \mathbb{Z}^+ : \sum_{j=1}^{d} \lambda_l^j \geq \frac{v}{100} \right\}. \tag{2}$$

Here, λ_l^j denotes the variance contribution of the j-th normalized principal component at layer l. We primarily report Rank@90% and Rank@95% as quantitative indicators. Since Qwen2.5-VL adopts a dynamic-resolution encoding mechanism, we fix its input visual token count to 1024 for consistent cross-model comparison.

Figure 2 provides cross-model Rank@90% and Rank@95% results on POPE. Across all three models, both indicators remain far below the full feature dimension, confirming that the low-rank structure is a consistent property rather than an isolated phenomenon. This implies that visual token redundancy is largely rooted in the shared low-rank structure, as most tokens concentrate near a common subspace and thus carry highly overlapping information. The plot also reveals a clear layer-wise trend: shallow layers are more low-rank, whereas deeper layers are more dispersed, likely because shallow layers primarily encode shared patterns while deeper layers capture finer semantic details.

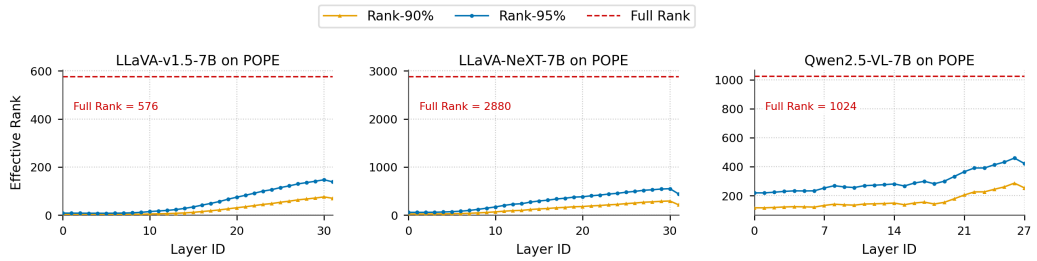

Figure 2: Layer-wise Rank@90% and Rank@95% on POPE for LLaVA-v1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B. Lower values indicate stronger low-rank compressibility. Additional results are in Appendix A.2.

Furthermore, we examine the stability of the dominant subspace under token reduction by measuring the principal-angle-based similarity (Björck & Golub, 1973) between the subspaces estimated from the full token set and from the remaining tokens after random dropout. As shown in Figure 3, subspace similarity remains high even after dropping 80% of tokens across all three models, indicating that the dominant low-rank subspace originates from the shared structure of visual representations rather than depending on specific tokens. We further verify that this stability also holds under importance-based pruning (i.e., the actual token selection performed by LRCP), with detailed results provided in Appendix A.2.

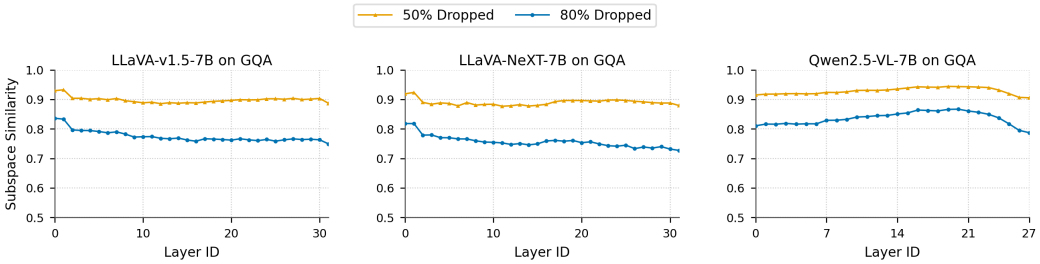

Figure 3: Low-rank subspace stability on GQA for LLaVA-v1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B. High similarity after randomly dropping 50% and 80% of tokens confirms that the dominant subspace is a collective property. Additional results are in Appendix A.2.

Overall, these results suggest three key properties: (i) visual tokens exhibit a pronounced low-rank structure that is consistent across models and datasets; (ii) visual token redundancy is largely rooted in this shared low-rank structure; (iii) the dominant low-rank subspace is highly stable under token reduction. Table 7 additionally verifies that the high-residual tokens selected under this view align with annotated task-relevant regions. Together, these results motivate us to leverage the global low-rank structure for visual token compression.

3.3 A Low-Rank Compressibility Analysis

Based on the empirical observations above, we view visual token compression as a lossy compression problem under a stable low-rank background. At every compression location, we mean-center and normalize the token matrix independently for each input; neither the mean nor the PCA basis is shared across samples. Let $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times D}$ denote the resulting token matrix. The goal is to select a subset S with $|S| = K$ tokens to retain. The selection should ensure that the compressed representation preserves as much task-relevant information as possible. Formally, this objective can be written as

$$S^\star = \arg \min_{S:\,|S|=K} \mathcal{D}(X, \widehat{X}_S),$$

where $\widehat{X}_S$ denotes the compressed representation induced by the retained tokens, and $\mathcal{D}(X, \widehat{X}_S)$ measures the distortion between the original visual representation and its compressed counterpart. In practice, the true task-level distortion is difficult to compute directly, since the contribution of

each visual token depends on the multimodal projector, cross-token interactions in the LLM, and the specific downstream task. We therefore introduce a tractable surrogate objective based on squared reconstruction error to characterize the geometric information loss caused by token selection.

Let $P_r \in \mathbb{R}^{D\times D}$ denote the orthogonal projection matrix onto the r-dimensional dominant subspace estimated from X. We decompose X into a low-rank component and a residual component:

$$X = XP_r + X(I - P_r).$$

Here, XP_r captures the shared low-rank component, while $X(I-P_r)$ contains the residual variations. Let $M_S \in \mathbb{R}^{N\times N}$ denote the diagonal row mask associated with the retained set S:

$$(M_S)_{ii} = \begin{cases} 1, & i \in S, \\ 0, & i \notin S. \end{cases}$$

The empirical analysis in Section 3.2 shows that the dominant subspace remains close to the original one even after a large fraction of tokens is randomly removed. This suggests that the dominant low-rank structure is not determined by a few individual tokens, but rather forms a shared background across the token set. Based on this observation, our surrogate analysis treats the shared low-rank component as approximately preserved after compression and focuses on characterizing the selection-dependent loss in the residual component. Specifically, we use the following surrogate representation:

$$\widehat{X}_S(P_r) = XP_r + M_S X(I - P_r).$$

The corresponding surrogate compression loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{LR}}(S; P_r) &= \|X - \widehat{X}_S(P_r)\|_F^2 \\ &= \|X - [XP_r + M_S X(I - P_r)]\|_F^2 \\ &= \|(I - M_S)X(I - P_r)\|_F^2 \\ &= \sum_{i\notin S} \|x_i(I - P_r)\|_2^2. \end{aligned}$$

Thus, under this surrogate objective, the selection-dependent loss is exactly the sum of the squared projection residuals of the discarded tokens with respect to the dominant subspace. This gives a direct surrogate criterion for pruning: tokens with larger projection residuals should be retained first, since removing them would incur a larger surrogate compression loss.

3.4 Low-Rank-Based Pruning Method

Based on the above analysis, we use PCA to estimate the dominant low-rank subspace and score each token by its projection residual, which serves as a surrogate measure of token importance. We propose Low-Rank Compressibility Guided Pruning (LRCP). LRCP can be seamlessly inserted after the visual encoder output or into intermediate layers of the LLM. It does not rely on attention weights, and is therefore compatible with efficient attention implementations such as FlashAttention (Dao et al., 2022; Dao, 2024). Specifically, LRCP consists of three stages: low-rank subspace construction, residual-based token ranking, and token merging. The complete pseudocode is provided in Algorithm 1 (Appendix A.3).

Low-rank subspace construction. Given the preprocessed token matrix $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N\times D}$ and a retention budget K, we set the PCA subspace dimension to r. We perform randomized low-rank PCA (Halko et al., 2011) and take the top r principal components to span the dominant subspace. Let the basis of this subspace be

$$U_r \in \mathbb{R}^{D\times r}, \tag{3}$$

then the corresponding projection matrix is $P_r = U_r U_r^\top$.

Projection residual computation. After obtaining U_r, we compute the projection residual of each token with respect to the dominant subspace and use it as the importance score:

$$s_i = \|x_i(I - P_r)\|_2^2. \tag{4}$$

Based on the projection residuals, we rank the visual tokens by importance and discard tokens with lower scores.

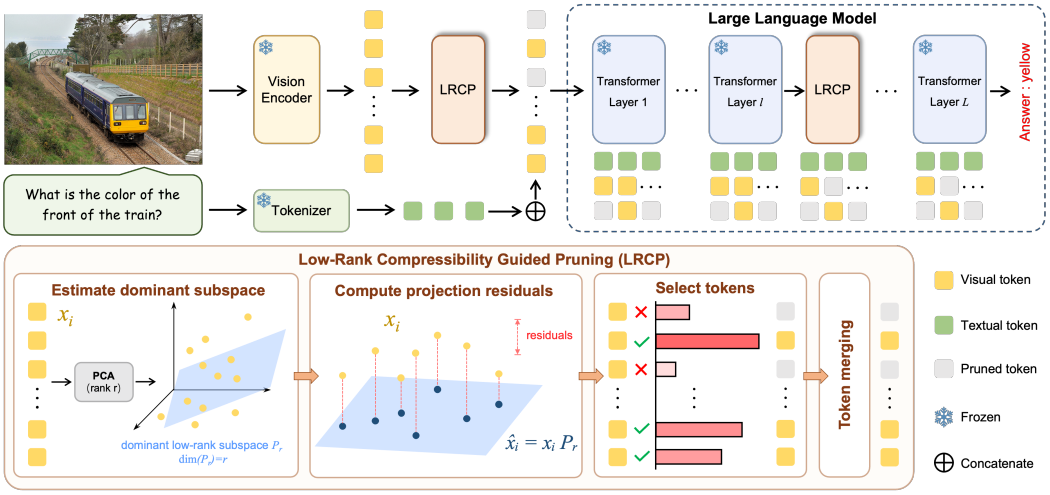

Figure 4: Overview of LRCP. The method estimates the dominant low-rank subspace via PCA, retains tokens with large projection residuals, and merges discarded tokens into their nearest retained neighbors.

Token merging. To reduce the information loss caused by directly discarding tokens, we adopt a cosine-similarity-based token merging strategy inspired by prior work on token merging (Bolya et al., 2023). For each token to be discarded, we assign it to the most similar retained token. Then, each retained token is updated by averaging it with all discarded tokens assigned to it, producing an updated feature representation.

4 Experiments

4.1 Experimental Setup

Models and baselines. To evaluate the effectiveness of the proposed method, we conduct experiments on four representative vision-language models. For image understanding, we use LLaVA-v1.5-7B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025); for video understanding, we use Video-LLaVA-7B (Lin et al., 2024). We compare against FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2024b), PDrop (Xing et al., 2024), VisionZip (Yang et al., 2024), V2Drop (Chen et al., 2025), and ApET (Ma et al., 2026). The image benchmarks are GQA, MMBench, MMBench-CN, MME, POPE, ScienceQA, VQAv2, VQA^{Text}, SEED, MMVet, and LLaVA-B (Hudson & Manning, 2019; Liu et al., 2024c; Fu et al., 2023; Li et al., 2023b; Lu et al., 2022; Goyal et al., 2017; Singh et al., 2019; Li et al., 2024b; Yu et al., 2023; Liu et al., 2024a). The video benchmarks are TGIF (Jang et al., 2017), MSVD (Chen & Dolan, 2011), and MSRVTT (Xu et al., 2016).

Implementation details. We follow the default inference configurations from the official implementations of all models and preserve each baseline's native pruning schedule while matching the average number of visual tokens per LLM layer. LRCP applies compression at the vision-encoder output and at an intermediate LLM layer: Layer 16 for LLaVA models and Layer 14 for Qwen2.5-VL. For example, at the LLaVA-v1.5 average-64 setting, LRCP retains 96 tokens before Layer 16 and 32 afterwards, yielding the same average budget as direct 64-token pruning; all settings are detailed in Appendix A.4. Qwen2.5-VL uses dynamic-resolution encoding, so we fix the minimum and maximum number of pixels to 256 and 2048, respectively. Video-LLaVA encodes each video into 8 frames with 256 visual tokens per frame, yielding 2048 tokens. All experiments use NVIDIA A800 GPUs.

Table 1: Performance comparison on 11 image understanding benchmarks with LLaVA-v1.5-7B (576 visual tokens). The last column reports the average retention ratio relative to the uncompressed upper bound. The best result in each setting is in **bold** and the second-best is underlined.

Method	GQA	MMB	MMB^{CN}	MME	POPE	SQA	VQA^{V2}	VQA^{Text}	SEED	MMVet	LLaVA-B	Avg.
Upper Bound, 576 Tokens **(100%)**												
Vanilla	61.9	64.7	58.1	1862	85.9	69.5	78.5	58.2	59.3	31.1	66.9	100%
Retain 192 Tokens (↓ 66.7%)												
FastV	52.7	61.2	57.0	1612	64.8	67.3	67.1	52.5	57.1	27.7	49.4	88.3%
SparseVLM	57.6	62.5	53.7	1721	83.6	69.1	75.6	56.1	55.8	31.5	66.1	96.2%
PDrop	57.3	63.2	56.8	1766	82.3	69.0	75.1	56.1	54.7	30.5	65.8	96.2%
VisionZip	59.3	63.0	57.3	1778	85.2	68.7	76.6	57.3	56.4	**31.7**	**67.7**	98.1%
V2Drop	58.5	**63.7**	56.6	1796	85.1	**69.3**	74.9	55.6	56.4	28.8	66.5	96.7%
ApET	60.2	63.4	**57.9**	1808	86.3	68.5	76.2	54.4	56.8	28.7	66.1	97.1%
LRCP	**60.3**	**63.7**	57.5	**1823**	**86.4**	68.8	**76.9**	**57.4**	**57.2**	31.2	67.5	**98.8%**
Retain 128 Tokens (↓ 77.8%)												
FastV	49.6	56.1	56.4	1490	59.6	60.2	61.8	50.6	**55.9**	28.1	52.0	84.4%
SparseVLM	56.0	60.0	51.1	1696	80.5	67.1	73.8	54.9	53.4	30.0	62.7	92.8%
PDrop	57.1	61.1	56.6	1644	82.3	68.4	72.9	54.8	53.3	30.8	61.9	94.1%
VisionZip	57.6	62.0	56.2	1759	83.2	**68.9**	75.6	**56.8**	54.9	**32.1**	**64.8**	96.6%
V2Drop	56.3	61.8	54.5	1712	80.9	68.8	72.1	53.8	53.8	29.3	62.9	93.5%
ApET	**58.9**	62.3	56.4	1801	86.1	68.7	75.1	53.9	54.7	29.0	64.2	95.8%
LRCP	**58.9**	**63.1**	**57.1**	**1807**	**87.0**	68.5	**75.8**	**56.8**	55.4	31.8	**64.8**	**97.7%**
Retain 64 Tokens (↓ 88.9%)												
FastV	46.1	48.0	52.7	1256	48.0	51.1	55.0	47.8	51.9	25.8	46.1	75.3%
SparseVLM	52.7	56.2	46.1	1505	75.1	62.2	68.2	51.8	51.1	23.3	57.5	84.7%
PDrop	47.5	58.8	50.5	1092	55.9	**69.2**	69.2	45.9	40.0	30.7	59.2	81.8%
VisionZip	55.1	60.1	**55.3**	1687	77.0	69.0	72.4	55.5	52.2	**31.5**	62.9	93.4%
V2Drop	50.5	55.2	49.7	1470	75.1	68.9	71.2	51.8	51.4	29.1	62.4	88.3%
ApET	56.9	**61.2**	54.4	1714	84.4	68.9	72.5	53.0	**52.4**	29.4	63.0	93.6%
LRCP	**57.1**	60.9	54.9	**1721**	**85.3**	68.5	**73.4**	**55.6**	52.2	31.3	**63.1**	**94.7%**

4.2 Main Results

Image understanding results. Table 1 reports the main results on LLaVA-v1.5-7B. LRCP achieves the best overall retention across different compression ratios, and the advantage widens as the token budget decreases. Under the most aggressive setting of retaining only 64 tokens, LRCP still preserves strong overall performance, exceeding the second-best method by 1.1% in average retention. This suggests that explicitly modeling the global structure of the visual token set is more stable under extremely low token budgets.

High-resolution and dynamic-resolution results. LLaVA-NeXT-7B splits each input image into four local views and one rescaled global view, resulting in a much larger visual token budget than LLaVA-v1.5-7B. Table 2 shows that LRCP achieves the best overall results under all three token budgets of 640, 320, and 160, surpassing the second-best method by 0.7%, 1.1%, and 1.7% in average performance retention, respectively. Table 3 shows a similar pattern on Qwen2.5-VL-7B. Under 20% and 10% average retention ratios, LRCP outperforms existing baselines by 1.2% and 2.0%, respectively, showing that the method adapts well to variable-length visual sequences.

Video understanding results. Table 4 summarizes the video understanding results on Video-LLaVA-7B under a unified budget of 256 retained tokens (an 87.5% reduction from the original 2048 video tokens). LRCP achieves 97.8% average accuracy retention and 99.3% average score retention, surpassing the second-best method (VisionZip) by 3.4% and 1.1%, respectively. This demonstrates that LRCP generalizes well beyond image understanding to temporal visual sequences.

4.3 Ablation Studies

Effect of the subspace dimension r. The subspace dimension r controls the boundary between shared structure and discriminative residuals. Table 5 shows that LRCP remains stable across a broad range of r values, although the best setting differs by model. This is expected, as models with

Table 2: Performance comparison on high-resolution image benchmarks with LLaVA-NeXT-7B (2880 visual tokens).

Method	GQA	MMB	MMB^{CN}	MME	POPE	VQA^{V2}	VQA^{Text}	Avg.
Upper Bound, 2880 Tokens **(100%)**								
Vanilla	64.2	67.4	60.6	1851	86.5	81.8	61.3	100%
Retain 640 Tokens (↓ 77.8%)								
SparseVLM	60.3	65.8	58.5	1773	84.2	77.1	57.8	95.7%
VisionZip	61.3	66.2	59.1	1787	85.9	79.1	**60.2**	97.4%
ApET	63.0	65.3	59.3	1815	**87.2**	79.2	57.9	97.6%
LRCP	**63.2**	**66.3**	**60.1**	**1827**	86.3	**79.6**	58.9	**98.3%**
Retain 320 Tokens (↓ 88.9%)								
SparseVLM	59.3	**64.2**	55.9	1690	83.3	75.7	58.8	93.7%
VisionZip	59.3	63.1	56.3	1702	82.1	76.2	**58.9**	93.6%
ApET	61.0	63.5	56.6	**1783**	**85.6**	75.8	54.4	94.2%
LRCP	**61.1**	63.9	**58.4**	1759	84.8	**76.7**	57.4	**95.3%**
Retain 160 Tokens (↓ 94.4%)								
SparseVLM	51.2	52.1	48.6	1542	72.7	66.3	46.4	80.2%
VisionZip	55.5	60.1	**55.1**	1628	74.8	71.4	**56.2**	88.6%
ApET	58.4	60.8	52.3	1680	82.6	72.7	53.8	90.1%
LRCP	**58.6**	**61.6**	54.5	**1693**	**83.3**	**74.6**	55.8	**91.8%**

Table 3: Performance comparison on dynamic-resolution image benchmarks with Qwen2.5-VL-7B.

Method	GQA	POPE	SQA	MME	MMB	Avg.
Upper Bound, 256–2048 Tokens **(100%)**						
Vanilla	60.5	86.2	76.7	2327	83.3	100%
Retain 20% Tokens in Average (↓ 80%)						
SparseVLM	54.7	73.6	71.6	2063	76.0	89.8%
PDrop	55.1	78.4	70.9	2117	77.3	91.6%
VisionZip	56.8	82.4	**76.3**	2134	**79.3**	95.2%
V2Drop	55.8	83.0	74.3	2196	77.6	94.6%
ApET	57.0	**83.6**	75.5	2211	77.4	95.5%
LRCP	**57.6**	83.5	76.0	**2283**	78.5	**96.7%**
Retain 10% Tokens in Average (↓ 90%)						
SparseVLM	51.3	71.9	68.2	1849	71.7	84.5%
PDrop	52.0	74.8	69.7	1886	73.6	86.6%
VisionZip	52.4	78.9	74.1	2003	**75.6**	90.3%
V2Drop	52.5	78.7	73.6	2058	74.5	90.4%
ApET	53.4	**79.3**	74.0	2030	73.8	90.5%
LRCP	**54.7**	79.1	**74.4**	**2168**	74.9	**92.5%**

Table 4: Video understanding results on Video-LLaVA-7B (2048 visual tokens). All methods retain 256 tokens (↓ 87.5%).

Method	TGIF		MSVD		MSRVTT		Avg.	
	Acc	Scr	Acc	Scr	Acc	Scr	Acc	Scr
Vanilla	46.9 100%	3.34 100%	69.8 100%	3.91 100%	57.1 100%	3.49 100%	100%	100%
PDrop	40.3 85.9%	3.21 96.1%	61.5 88.1%	3.74 95.7%	41.8 73.2%	3.19 91.4%	82.4%	94.4%
VisionZip	44.3 94.5%	3.29 98.5%	65.2 93.4%	3.83 97.9%	54.5 95.4%	3.43 98.3%	94.4%	98.2%
LRCP	**46.5** **99.1%**	**3.34** **100.0%**	**67.7** **97.0%**	**3.87** **99.0%**	**55.6** **97.4%**	**3.45** **98.9%**	**97.8%**	**99.3%**

Table 5: Ablation on subspace dimension r. Token retention ratios are 11.1% for LLaVA-v1.5-7B and 10% for Qwen2.5-VL-7B.

Model	Setting	GQA	POPE	SQA	MME
LLaVA-v1.5-7B	$r = 4$	57.1	**85.3**	**68.5**	**1721**
LLaVA-v1.5-7B	$r = 8$	**57.2**	84.9	68.1	1696
LLaVA-v1.5-7B	$r = 12$	56.5	84.2	67.4	1692
LLaVA-v1.5-7B	$r = 16$	56.6	84.1	67.8	1694
Qwen2.5-VL-7B	$r = 4$	54.1	77.1	74.2	2128
Qwen2.5-VL-7B	$r = 8$	**54.7**	**79.1**	74.4	**2168**
Qwen2.5-VL-7B	$r = 12$	54.2	77.5	**74.5**	2132
Qwen2.5-VL-7B	$r = 16$	53.9	77.1	74.1	2136

different degrees of low-rank compressibility require appropriately sized subspaces to capture the shared background. We use $r = 4$ for LLaVA-based models and $r = 8$ for Qwen-based models. Furthermore, when r becomes too large, the dominant subspace absorbs more fine-grained token variations, making the residual scores less discriminative.

Scoring criterion and task relevance. Table 6 compares four scoring variants on LLaVA-v1.5-7B with 64 retained tokens: *Low-rank projection norm* ($\|x_i P_r\|_2^2$ with merging, descending), *Small residual first* ($\|x_i(I - P_r)\|_2^2$ with merging, ascending), *Residual only* (projection-residual scoring without merging), and *LRCP full* (projection-residual scoring with merging). The full LRCP variant consistently outperforms the alternatives, indicating that projection-residual-based scoring provides an effective measure of token importance and that token merging further improves performance. To test whether large residuals reflect useful evidence rather than artifacts, we map selected tokens back

Table 6: Ablation of token scoring criteria on LLaVA-v1.5-7B with 64 retained visual tokens. All variants use the same compression location.

Setting	GQA	POPE	SQA	MME	VQA^{Text}
Low-rank projection norm	56.1	81.0	67.4	1598	53.9
Small residual first	54.7	80.6	66.1	1532	54.2
Residual only	56.8	82.3	68.4	1612	55.3
LRCP full	**57.1**	**85.3**	**68.5**	**1721**	**55.6**

Table 7: Task-relevance validation. A selected patch is counted as overlapping when it intersects an annotated informative region.

Setting	Samples	Overlap
POPE object regions	200 images	93.4%
TextVQA OCR regions	200 images	94.2%
Complex-motion videos	200 frames	90.5%
Noisy or artifact images	200 images	88.3%

to image patches under the main-experiment budgets. A patch is counted as relevant if it intersects an annotated informative region. Table 7 shows 93.4% overlap on POPE, 94.2% on TextVQA, and 90.5% on complex-motion video frames; even deliberately noisy or artifact-rich images retain 88.3% overlap.

Efficiency analysis. As shown in Table 8, LRCP achieves the highest inference speedup on both LLaVA-v1.5-7B and LLaVA-NeXT-7B. On LLaVA-v1.5-7B, LRCP with 64 retained visual tokens achieves a $1.48\times$ overall speedup and a $1.59\times$ prefilling speedup. On LLaVA-NeXT-7B, the advantage further increases to $2.96\times$ and $3.00\times$, respectively. LRCP also yields the lowest theoretical FLOPs in both settings. All timings include the full overhead of each compression method; for LRCP this covers mean centering, PCA, projection-residual scoring, token selection, and merging. Detailed budget matching and complexity accounting are provided in Appendix A.4.

Table 8: Efficiency analysis on LLaVA-v1.5-7B (left) and LLaVA-NeXT-7B (right). We report total inference time, prefilling time, and TFLOPs. Δ denotes the speedup relative to the uncompressed baseline. All timings include the full overhead of each compression method and are measured on a single A800 GPU.

Methods	Token	Total Time↓	Δ ↑	Prefilling Time↓	Δ ↑	TFLOPs
LLaVA-v1.5-7B	576	17:05	1.00×	102ms	1.00×	8.82
+ SparseVLM	64	15:57	1.07×	90.1ms	1.13×	2.31
+ PDrop	64	12:56	1.32×	72.5ms	1.41×	2.16
+ VisionZip	64	12:10	1.40×	69.2ms	1.47×	2.03
+ V2Drop	64	12:24	1.38×	70.1ms	1.46×	2.12
+ ApET	64	11:46	1.45×	65.2ms	1.56×	2.09
+ LRCP	64	**11:31**	**1.48**×	**64.0ms**	**1.59**×	**2.02**

Methods	Token	Total Time↓	Δ ↑	Prefilling Time↓	Δ ↑	TFLOPs
LLaVA-NeXT-7B	2880	35:16	1.00×	206ms	1.00×	31.03
+ SparseVLM	160	30:26	1.16×	136ms	1.51×	7.62
+ PDrop	160	13:45	2.56×	79.5ms	2.59×	6.78
+ VisionZip	160	12:32	2.81×	69.9ms	2.95×	4.72
+ V2Drop	160	12:49	2.75×	73.5ms	2.80×	5.22
+ ApET	160	12:03	2.93×	69.1ms	2.98×	4.74
+ LRCP	160	**11:56**	**2.96**×	**68.7ms**	**3.00**×	**4.67**

5 Conclusion

We propose LRCP, a training-free visual token compression method guided by low-rank compressibility for efficient LVLM inference. Our analysis shows that visual tokens in LVLMs exhibit a stable low-rank structure across models, datasets, and layers, suggesting that visual token redundancy can be viewed as a consequence of shared low-rank representations. LRCP selects tokens according to their projection residuals with respect to the dominant low-rank subspace, aligning token selection with the global representation structure of visual tokens. Additionally, it is fully compatible with efficient attention implementations like FlashAttention. Extensive experiments show that LRCP achieves a favorable balance between performance preservation and computational efficiency.

AI use statement

Generative AI was used to assist with the LaTeX formatting and template migration of this submission. It was also used to polish the manuscript for clarity and readability. The authors reviewed every resulting change and take responsibility for the final content of this work, including all text, claims, and artifacts.

Reproducibility statement

Section 4.1 specifies the model configurations, compression layers, hyperparameters, datasets, and compute platform used in the experiments. Algorithm 1 and the implementation details in the appendix provide the full LRCP procedure and its complexity analysis; the appendix also reports additional analyses and layer-wise ablations.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Yvon Apedo, Martyna Poreba, Michal Szczepanski, and Samia Bouchafa. Beyond attention scores: SVD-based vision token pruning for efficient vision-language models. *arXiv preprint arXiv:2604.11530*, 2026.

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

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

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

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

David Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 190–200, 2011.

Junjie Chen, Xuyang Liu, Zichen Wen, Yiyu Wang, Siteng Huang, and Honggang Chen. Variation-aware vision token dropping for faster large vision-language models. *arXiv preprint arXiv:2509.01552*, 2025.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022.

Alexei Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

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*, pp. 6904–6913, 2017.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

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*, 2019.

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*, pp. 2758–2766, 2017.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

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

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*, pp. 19730–19742, 2023a.

Xu Li, Yi Zheng, Yuxuan Liang, Zhe Liu, Xiaolei Chen, Haotian Chen, Rui Zhu, and Xiangyang Xue. ResPrune: Text-conditioned subspace reconstruction for visual token pruning in large vision-language models. *arXiv preprint arXiv:2603.21105*, 2026.

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*, 2023b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

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

Jizhihui Liu, Feiyi Du, Guangdao Zhu, Niu Lian, Jun Li, and Bin Chen. Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models. *arXiv preprint arXiv:2508.00553*, 2025a.

Xuyang Liu, Zichen Wen, Shaobo Wang, Junjie Chen, Zhishan Tao, Yubo Wang, Xiangqi Jin, Chang Zou, Yiyu Wang, Chenfei Liao, et al. Shifting AI efficiency from model-centric to data-centric compression. *arXiv preprint arXiv:2505.19147*, 2025b.

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

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

Qiankun Ma, Ziyao Zhang, Haofei Wang, Jie Chen, Zhen Song, and Hairong Zheng. Apet: Approximation-error guided token compression for efficient vlms. *arXiv preprint arXiv:2602.19870*, 2026.

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

OpenAI. GPT-4V(ision) system card. *OpenAI Technical Report*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

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

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Haicheng Wang, Zhemeng Yu, Gabriele Spadaro, Chen Ju, Victor Quétu, Shuai Xiao, and Enzo Tartaglione. Folder: Accelerating multi-modal large language models with enhanced performance. *arXiv preprint arXiv:2501.02430*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14593–14603, 2025.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016.

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.

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

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Yaqi Xie, Katia Sycara, Haitao Mi, and Dong Yu. Vscan: Rethinking visual token reduction for efficient large vision-language models. *arXiv preprint arXiv:2505.22654*, 2025b.

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

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

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

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

A Appendix

A.1 Qualitative Visualization

To further illustrate the effectiveness of LRCP, we provide qualitative visualizations of token selection results in Figure 5. We compare LRCP with a representation-based method, where we choose ApET (Ma et al., 2026) as the baseline. LRCP more consistently retains text-bearing patches and fine-grained objects that are critical for answering the question, whereas the representation-based method tends to discard such discriminative tokens. These visualizations confirm that modeling the global structure of visual tokens enables more precise retention of irreplaceable visual information, leading to more robust token reduction.

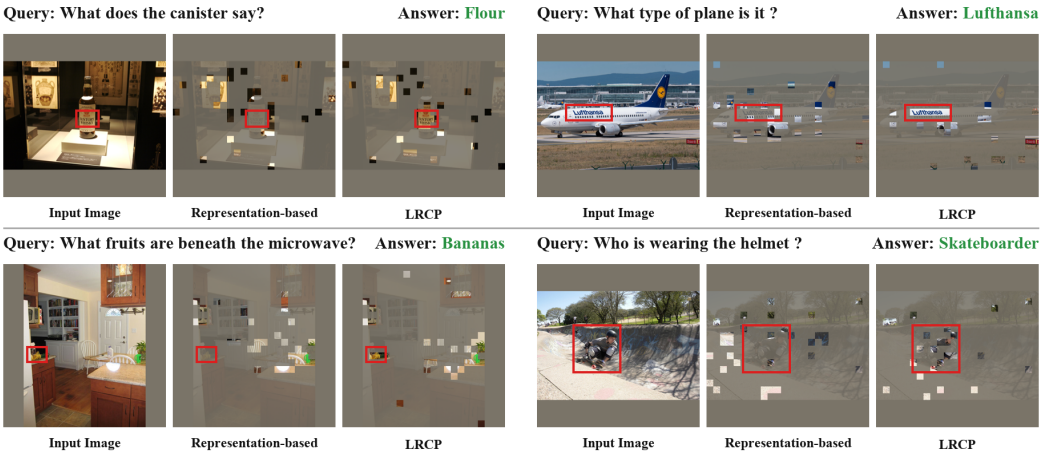

Figure 5: Qualitative comparison of retained-token distributions on VQA^{Text} (top) and GQA (bottom). Retained tokens are highlighted; red boxes indicate answer-relevant regions. LRCP preserves more tokens in semantically informative areas.

A.2 Additional Low-Rank Analysis on Other Datasets

We extend the effective-dimensionality analysis to GQA and VQA^{Text}, and the subspace stability analysis to POPE and VQA^{Text}, verifying that the observed low-rank structure generalizes across datasets.

Effective dimensionality on additional datasets. Figures 6 and 7 show the layer-wise Rank@90% and Rank@95% curves on GQA and VQA^{Text}, respectively. The trends are highly consistent with the POPE results reported in the main text. (i) The effective rank remains substantially below the full feature dimension across all layers and models, reaffirming the pervasive low-rank compressibility of visual tokens. (ii) Shallow layers consistently exhibit lower effective rank than deeper layers, reflecting the transition from shared low-frequency visual patterns to more diverse, semantically rich representations. (iii) The cross-model ordering is preserved: LLaVA-v1.5-7B and LLaVA-NeXT-7B display similar rank profiles due to their shared ViT-based architecture, while Qwen2.5-VL-7B shows a slightly different trajectory owing to its dynamic-resolution encoding and distinct visual backbone.

Notably, VQA^{Text} images contain dense textual content and complex visual layouts, which might be expected to require higher-dimensional representations. Nevertheless, the rank statistics remain comparable, suggesting that the low-rank structure is an intrinsic property of how LVLMs encode visual information, rather than an artifact of image simplicity.

Subspace stability on additional datasets. The main text reports subspace stability on GQA (Figure 3). We now extend this analysis to POPE and VQA^{Text} (Figures 8 and 9). Following the same protocol, we measure the principal-angle-based similarity (Björck & Golub, 1973) between the dominant subspace estimated from the full visual token set and the subspace re-estimated after randomly dropping 50% and 80% of the tokens. Across all models and both datasets, the subspace

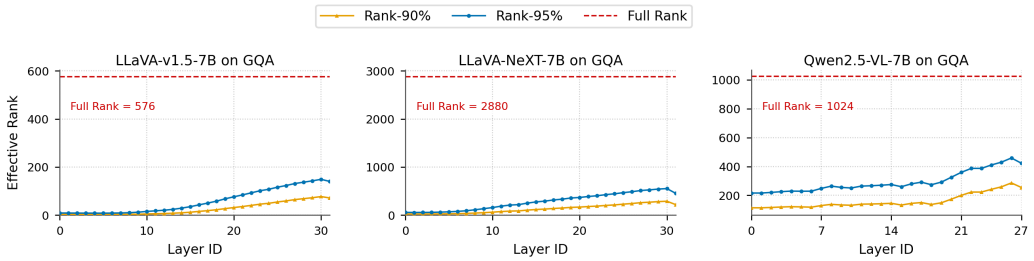

Figure 6: Layer-wise Rank@90% and Rank@95% statistics on GQA for LLaVA-v1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B. The low-rank structure is consistent with the POPE results in the main text.

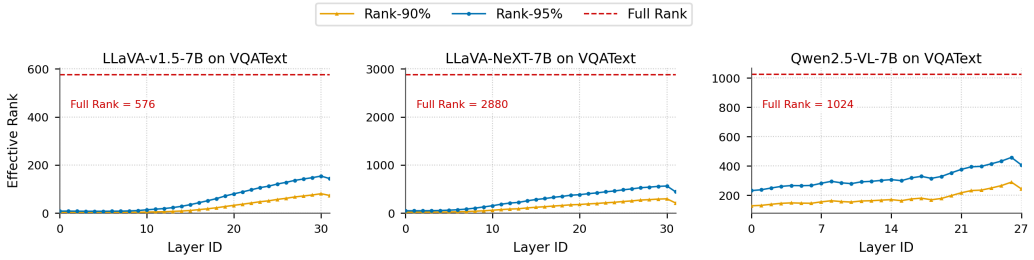

Figure 7: Layer-wise Rank@90% and Rank@95% statistics on VQA^{Text} for LLaVA-v1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B. Despite the higher visual complexity of text-centric images, the low-rank property persists.

similarity remains consistently high (above 0.9 for most layers), even under the extreme 80% dropout ratio. This corroborates the main-text finding that the dominant low-rank subspace arises from the collective structure of the visual token population rather than being anchored to a few specific tokens.

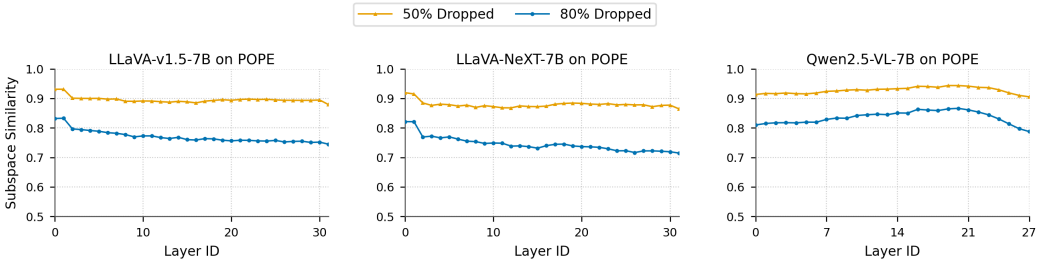

Figure 8: Low-rank subspace stability on POPE for LLaVA-v1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B. Subspace similarity remains high even after removing 80% of tokens.

Subspace stability under importance-based pruning. The stability analyses above employ random token dropout as a stress test. Since LRCP preferentially retains tokens that *deviate* from the dominant subspace, a natural follow-up question is whether the subspace remains stable under this non-random, importance-based selection. To investigate this, we compute the principal-angle-based subspace similarity between the full-set subspace and the subspace estimated from only the LRCP-retained tokens after each pruning stage.

As shown in Figure 10, the dominant subspace estimated from the retained tokens remains highly similar to the original across all measured stages. This observation can be explained as follows: tokens with large projection residuals still possess substantial components within the dominant subspace, which are sufficient to reconstruct the subspace with high fidelity. This stability under importance-based pruning provides a post-hoc justification for a key assumption in our surrogate analysis (Section 3.3), namely that the low-rank component is approximately preserved after compression.

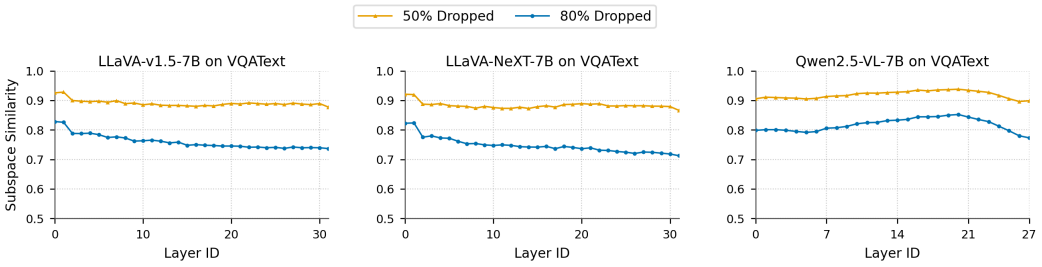

Figure 9: Low-rank subspace stability on VQA^{Text} for LLaVA-v1.5-7B, LLaVA-NeXT-7B, and Qwen2.5-VL-7B. The stability is maintained across both text-dense and natural-scene images.

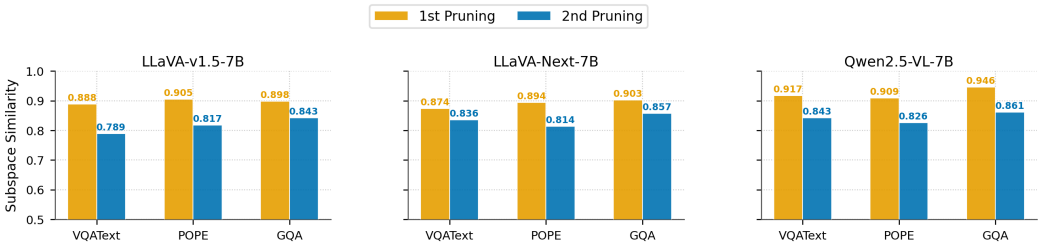

Figure 10: Subspace stability under importance-based pruning by LRCP. We compare the dominant subspace estimated from the full token set with the subspace estimated from the retained tokens at each pruning stage. The consistently high similarity validates the assumption that LRCP preserves the low-rank structure.

A.3 Algorithm Pseudocode

We provide the complete pseudocode of LRCP in Algorithm 1. The procedure takes as input the visual token matrix $X \in \mathbb{R}^{N \times D}$, a retention budget K, and a subspace dimension r. It outputs the compressed token set $X' \in \mathbb{R}^{K \times D}$.

Algorithm 1 Low-Rank Compressibility Guided Pruning (LRCP)

Require: Visual tokens $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N \times D}$, token budget K, subspace rank r
Ensure: Compressed tokens $X' \in \mathbb{R}^{K \times D}$

1: Mean-center and normalize X along the token dimension
2: Compute the top-r PCA basis $U_r \in \mathbb{R}^{D \times r}$ of X
3: $s_i \leftarrow \|x_i(I - U_r U_r^\top)\|_2^2, \quad i = 1, \ldots, N$
4: $\mathcal{S} \leftarrow \text{TopK}_i(s_i, K), \quad \mathcal{R} \leftarrow \{1, \ldots, N\} \setminus \mathcal{S}$
5: **for** $j \in \mathcal{S}$ **do**
6: $\quad \mathcal{A}_j \leftarrow \{i \in \mathcal{R} : j = \arg\max_{k \in \mathcal{S}} \cos(x_i, x_k)\}$
7: $\quad x'_j \leftarrow \dfrac{x_j + \sum_{i \in \mathcal{A}_j} x_i}{1 + |\mathcal{A}_j|}$
8: **end for**
9: **return** $X' = [x'_j]_{j \in \mathcal{S}}^\top$

Computational complexity. The dominant cost of LRCP lies in the truncated PCA decomposition (Line 2), which computes the top-r principal components of X. Using randomized SVD (Halko et al., 2011), this step and residual scoring each require $\mathcal{O}(NDr)$ operations, while top-K selection takes $\mathcal{O}(N \log K)$. Cosine merging costs $\mathcal{O}((N-K)KD)$ in the worst case. The concrete settings and accounting used for Table 8 are provided next.

A.4 Implementation and Budget Details

At every compression location, we process each sample independently: we subtract its mean token feature, normalize along the token dimension, and call `torch.pca_lowrank` with two subspace iterations and random seed 42. The decomposition is re-estimated from the current tokens at the second pruning stage; hence neither batch size nor other samples affect a sample's score.

For a fair comparison, each baseline follows its original pruning schedule while all methods are matched by the average number of visual tokens per LLM layer. Table 9 gives the absolute two-stage budgets for LLaVA-v1.5-7B. For the average-64 setting, the two stages use $576 \to 96$ and $96 \to 32$ tokens at Layer 16. This schedule has $16 \times 96 + 16 \times 32 = 2048$ visual token-layers, identical to direct 64-token pruning over 32 layers.

Table 9: Two-stage budgets for LLaVA-v1.5-7B. The reported setting is the average number of tokens across all 32 LLM layers.

Average budget	After Stage 1	Stage 2 layer	After Stage 2	Final retain
192	288	16	96	16.7%
128	192	16	64	11.1%
64	96	16	32	5.6%

For the average-64 configuration, Stage 1 decomposes 576×1024 features and Stage 2 decomposes 96×4096 features. With $r = 4$ and two randomized-PCA iterations, we conservatively estimate the complete two-stage module at 0.40 GFLOPs, versus 0.24 GFLOPs for direct $576 \to 64$ compression. Including the attention difference gives 0.694 GFLOPs (0.034% of the reported 2.02 TFLOPs), so both schedules round to 2.02 TFLOPs while LRCP retains more early-layer visual evidence.

A.5 Additional Ablations and Baseline Comparison

Dominant-subspace extraction. We fix the residual scoring rule and token budget, and compare four strategies for constructing the dominant subspace. PCA achieves the best overall results, indicating that leading directions of variation produce the most reliable residual scores.

Table 10: Subspace-extraction ablation on LLaVA-v1.5-7B with 64 retained visual tokens.

Subspace	GQA	POPE	SQA	MME	VQA^{Text}
Random directions	51.1	82.5	64.2	1485	52.3
Coordinate variance	55.7	<u>84.3</u>	67.1	1579	53.9
Cluster centers	**57.6**	84.0	<u>67.7</u>	<u>1706</u>	<u>54.8</u>
PCA	<u>57.1</u>	**85.3**	**68.5**	**1721**	**55.6**

Comparison with SVD-Prune. SVD-Prune (Apedo et al., 2026) selects tokens that contribute most to the dominant low-rank subspace. Under the 64-token LLaVA-v1.5-7B setting, LRCP's residual criterion produces stronger results on all four shared metrics (Table 11).

Table 11: Comparison with SVD-Prune on LLaVA-v1.5-7B at 64 tokens.

Method	GQA	POPE	VQA^{Text}	MME
SVD-Prune	53.77	83.87	55.14	1575
LRCP	**57.10**	**85.30**	**55.60**	**1721**

A.6 Compression Layer Ablation

LRCP applies compression at two stages: the vision encoder output (fixed) and an intermediate LLM layer (a design choice; default: Layer 16 for LLaVA, Layer 14 for Qwen2.5-VL). Here we ablate the LLM compression layer while keeping the first stage unchanged. We sweep over layers {8, 12,

14, 16, 18, 20} for LLaVA-v1.5-7B (32 layers, 64 retained tokens) and {8, 10, 12, 14, 16, 18} for Qwen2.5-VL-7B (28 layers, 10% average retention). Results are shown in Tables 12 and 13.

Table 12: Ablation on the intermediate compression layer for LLaVA-v1.5-7B with an average of 64 retained visual tokens.

Layer	GQA	POPE	SQA	MME	VQA^{Text}
Layer 8	56.4	82.9	67.0	1619	55.3
Layer 12	**57.1**	83.8	67.6	1667	55.4
Layer 14	57.0	83.9	68.1	1702	55.4
Layer 16	**57.1**	**85.3**	**68.5**	**1721**	55.6
Layer 18	56.9	83.1	68.2	1625	**55.8**
Layer 20	56.6	85.2	67.3	1709	55.0

Table 13: Ablation on the intermediate compression layer for Qwen2.5-VL-7B with 10% average token retention.

Layer	GQA	POPE	SQA	MME	MMB
Layer 8	54.4	78.3	74.1	2128	74.6
Layer 10	54.1	80.1	73.8	2090	73.3
Layer 12	54.1	**81.0**	73.7	2028	74.0
Layer 14	**54.7**	79.1	**74.4**	**2168**	**74.9**
Layer 16	54.4	78.1	74.3	2139	74.8
Layer 18	54.4	78.3	74.1	2128	74.6

Analysis. Several observations emerge from Tables 12 and 13. First, for LLaVA-v1.5-7B, compressing at very early layers (e.g., Layer 8) consistently underperforms later layers across all benchmarks. This is expected: early LLM layers are still integrating cross-modal information from the visual encoder, and premature token removal disrupts this process. As the compression layer moves deeper, performance generally improves, with Layers 14–20 exhibiting comparable and near-optimal accuracy. We select Layer 16 as the default for LLaVA-based models, since it provides the best overall balance.

Second, for Qwen2.5-VL-7B, performance is notably more stable across compression layers, with narrow score ranges on all benchmarks. This robustness likely stems from Qwen2.5-VL's dynamic-resolution architecture, which may enable stronger cross-layer feature alignment. Layer 14 achieves the best results on most metrics and is therefore adopted as the default.

Overall, LRCP demonstrates low sensitivity to the compression layer choice in the middle-to-deep range, reducing the need for extensive hyperparameter tuning across architectures.

A.7 Per-Stage Pruning Ratios

The main text reports token budgets as the average number of visual tokens per layer. Table 14 provides the detailed per-stage retention ratios for all experimental settings.

Across all models, the second-stage retention ratio is kept fixed (33.3% for LLaVA-based models, 50% for Qwen2.5-VL), while the first-stage ratio is adjusted to meet the target budget. This design concentrates the compression control in the first stage, keeping the second stage as a uniform refinement step.

A.8 Supplementary Visualization

To complement the qualitative analysis in Figure 5, we provide additional visualization examples illustrating the token selection behavior of LRCP across different image types and compression levels.

Table 14: Per-stage retention ratios for each model. The compression layers are Layer 16 for LLaVA-based models (32 LLM layers) and Layer 14 for Qwen2.5-VL (28 LLM layers). "Final Retain" is the product of the two stages; "Avg. Retain" is the average retention ratio across all LLM layers, corresponding to the token budgets reported in the main text.

Model	Setting	Stage 1 Retain	Stage 2 Retain	Final Retain	Avg. Retain
LLaVA-v1.5-7B	Avg 192	50.0%	33.3%	16.7%	33.3%
	Avg 128	33.3%	33.3%	11.1%	22.2%
	Avg 64	16.7%	33.3%	5.6%	11.1%
LLaVA-NeXT-7B	Avg 640	33.3%	33.3%	11.1%	22.2%
	Avg 320	16.7%	33.3%	5.6%	11.1%
	Avg 160	8.3%	33.3%	2.8%	5.6%
Qwen2.5-VL-7B	Avg 20%	26.7%	50%	13.3%	20.0%
	Avg 10%	13.3%	50%	6.7%	10.0%

Figure 11 shows LRCP's retained-token distributions under three different retention ratios. As the token budget decreases, LRCP progressively discards tokens in visually homogeneous regions while consistently preserving semantically salient areas such as object boundaries, fine-grained details, and regions relevant to the question. This is consistent with our theoretical motivation: background regions lie close to the dominant low-rank subspace and are compressed first, whereas discriminative regions with large projection residuals are preferentially retained.

Figure 12 presents examples from VQA^{Text}, where text-bearing regions occupy a small fraction of the image yet carry critical information. We compare LRCP with a representation-based baseline: the baseline spreads retained tokens across the image according to feature-level diversity, which may miss concentrated text regions. In contrast, LRCP assigns high importance scores to text regions because their fine-grained, high-frequency features deviate strongly from the low-rank visual background. As highlighted by the red boxes, LRCP more reliably retains tokens covering the relevant textual content, leading to more accurate answers.

A.9 Broader Impact

This work proposes a training-free visual token compression method for large vision-language models. We discuss both positive and potential negative societal implications below.

Positive impacts. By significantly reducing the number of visual tokens processed during inference, LRCP lowers the computational cost (FLOPs), memory footprint, and wall-clock latency of LVLM deployment. These efficiency gains have several positive downstream effects: (i) *Reduced energy consumption*: Large-scale LVLM serving is energy-intensive; token compression directly reduces the carbon footprint per inference call, contributing to more sustainable AI deployment. (ii) *Broader accessibility*: Lower resource requirements enable deployment on edge devices, mobile platforms, and in resource-constrained environments, democratizing access to multimodal AI capabilities. (iii) *Scalability*: As LVLMs continue to grow in parameter count and context length, efficient token management becomes critical for maintaining practical inference throughput in real-world applications.

Potential negative impacts. We do not foresee direct negative societal impacts from this work, as LRCP does not alter the pretrained model weights or introduce new training data; it only accelerates existing models at inference time. However, two indirect considerations merit discussion: (i) As with any efficiency improvement, reduced deployment cost may lower the barrier to using LVLMs in applications that could raise ethical concerns (e.g., surveillance, content generation without consent). (ii) Aggressive token pruning may inadvertently remove tokens encoding information about underrepresented visual content (e.g., small objects, fine-grained text, minority-class instances), potentially introducing or amplifying distributional biases in model outputs. We encourage practitioners to evaluate compression methods on diverse and representative benchmarks before deployment.

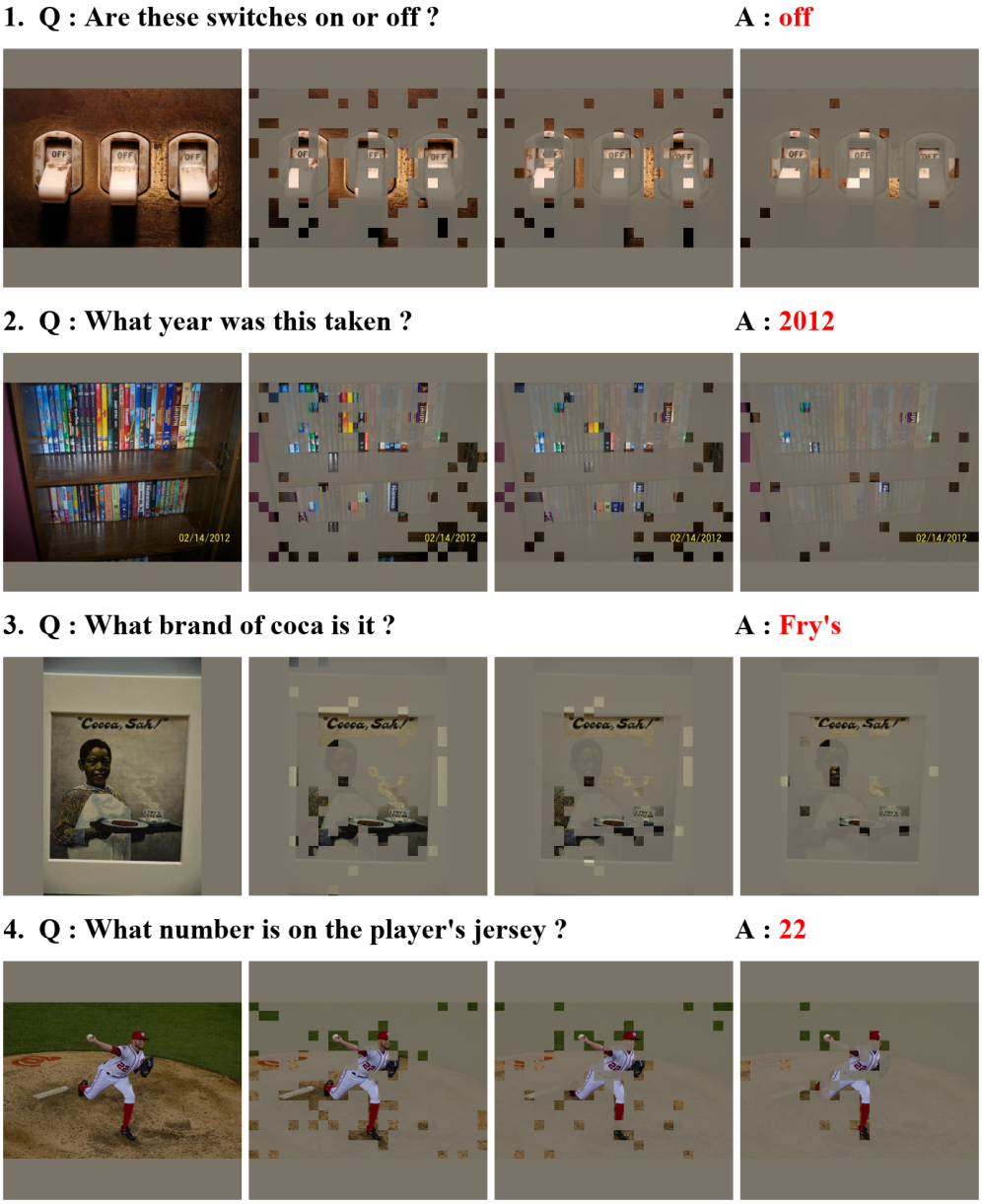

Figure 11: Retained-token distributions of LRCP under three retention ratios. As the budget decreases, tokens in homogeneous backgrounds are progressively removed while semantically informative regions are preserved.

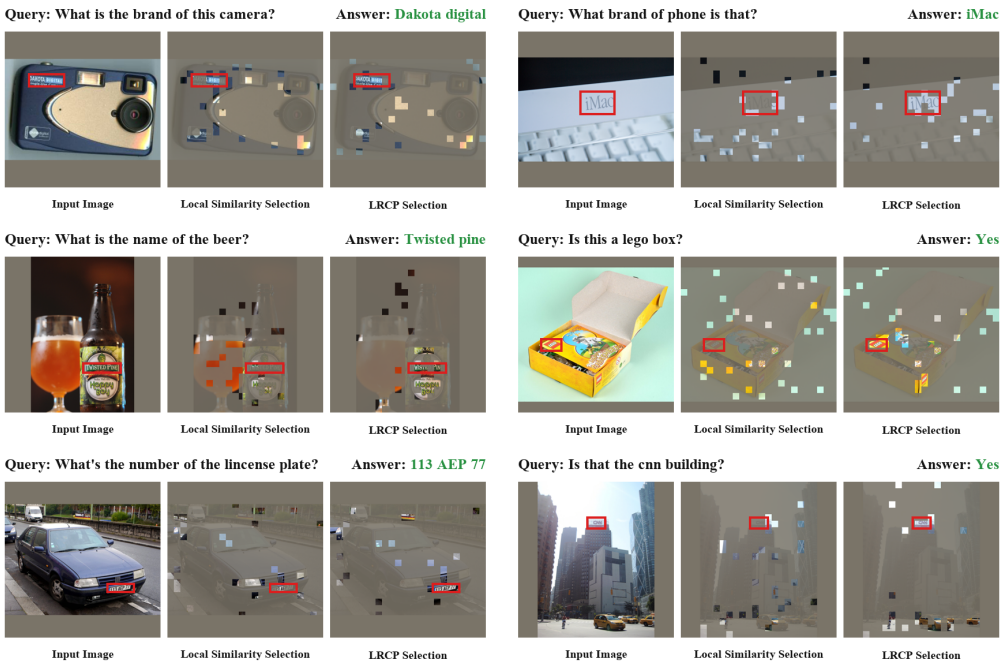

Figure 12: Qualitative examples on VQA^{Text}. For each case, we show the input image, the representation-based selection result, and the LRCP selection result. Red boxes indicate answer-relevant regions. LRCP more effectively preserves text-bearing regions.