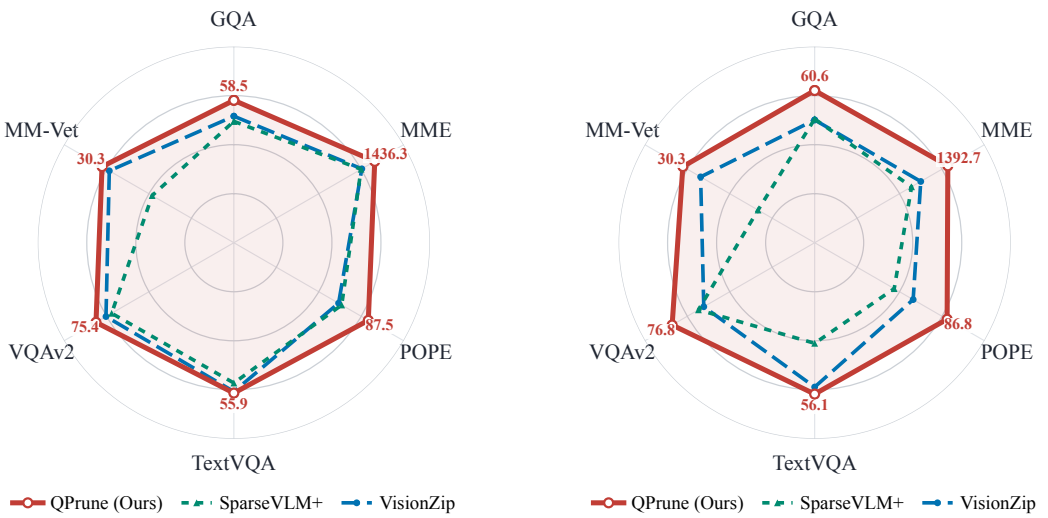
QPRUNE: QUANTUM-INSPIRED MIXED-STATE REPRESENTATION FOR VISUAL TOKEN PRUNING IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review



(a) LLaVA-1.5-7B (b) LLaVA-NeXT-7B

Figure 1: **Performance comparison of visual-token pruning methods across multiple benchmarks on LLaVA-1.5-7B and LLaVA-NeXT-7B.** QPrune achieves the best overall performance across benchmarks.

ABSTRACT

Multimodal large language models (MLLMs) process a large number of visual tokens, resulting in substantial inference overhead, which has motivated extensive research on visual token pruning. Existing methods commonly identify visual tokens based on attention scores, visual saliency, feature similarity, or cross-modal relevance. However, identifying informative visual tokens requires accounting for the interactions between visual and textual tokens, which together form a complex information state. Quantum information theory offers an alternative perspective for analyzing such a complex information state. In this work, we propose **QPrune**, a quantum-inspired visual token pruning method centered on mixed-state representation. QPrune represents the visual token set as a mixed state, where each visual token is represented as a pure state with an associated probability under the current image-text pair. Based on the mixed state, state-aware token pruning considers both token relevance and state overlap with the retained tokens to reduce redundancy. State-preserving token refinement further supplements local visual information that is insufficiently represented after pruning. Extensive experiments demonstrate that QPrune effectively reduces the number of visual tokens while maintaining strong task performance across different pruning ratios.

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.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved remarkable progress in vision-language understanding and reasoning (Liu et al., 2023; 2024a; Li et al., 2023a; Bai et al., 2023). These models typically encode an image into a sequence of visual tokens and process them together with textual tokens in a large language model. However, the large number of visual tokens introduces substantial computational and memory overhead during inference (Chen et al., 2024a; Xing et al., 2025; Zhang et al., 2025c; Khaki et al., 2025; Liang et al., 2025). Visual token pruning has therefore emerged as an effective approach to improving the inference efficiency of MLLMs by removing redundant or less informative visual tokens (Shang et al., 2025; Song et al., 2025; Yang et al., 2025b; Ye et al., 2025).

Existing visual token pruning methods commonly estimate token importance using attention scores or cross-modal relevance (Chen et al., 2024a; Zhang et al., 2026; Yang et al., 2025c; Khaki et al., 2025), while others reduce visual redundancy based on feature similarity, diversity, or token duplication (Yang et al., 2025b; Alvar et al., 2025; Wen et al., 2025; Dhouib et al., 2025; Yang et al., 2025a). Recent methods further explore adaptive token reduction and spatial information preservation (Ye et al., 2025; Zhao et al., 2025; Endo et al., 2025). However, in MLLMs, visual tokens interact with the textual input during multimodal processing, and their relevance therefore varies across image–text pairs. Meanwhile, different visual tokens may contain overlapping information, making the value of a token dependent not only on its individual relevance but also on the tokens that have already been retained. Together, these properties form a complex information state over the visual tokens, in which both token relevance and inter-token relations affect token selection.

Quantum information theory offers a natural perspective for representing such a complex information state (Wu et al., 2021). In quantum information theory, a mixed state describes an overall state as a probabilistic ensemble of pure states. Density-matrix-based formulations have been adopted in quantum-inspired semantic representation to model probabilistic states and semantic uncertainty (Zhang et al., 2018; Meyer & Lewis, 2020; Zhang et al., 2025a; Kankeu et al., 2025). Related quantum-inspired formulations have also been explored in multimodal representation learning (Li et al., 2023c; Qiao et al., 2024). This formulation is well suited to visual token pruning: each visual token can be represented as a pure state, while its associated probability characterizes its relevance under the current image–text pair. The overlap between token states further captures the relations and potential redundancy among visual tokens.

Motivated by this formulation, we propose **QPrune**, a quantum-inspired visual token pruning method that leverages mixed-state representation to capture both visual and textual information. By integrating the textual and visual states into a unified mixed-state representation, QPrune estimates the importance of each visual token based on its alignment with the input text. Based on this representation, **state-aware token pruning** jointly considers token relevance and state overlap with the retained tokens, thereby reducing redundant token representations during selection. Since aggressive pruning may leave local visual information insufficiently represented, **state-preserving token refinement** further supplements the retained set with complementary local information. Extensive experiments across multiple benchmarks and MLLMs demonstrate the effectiveness of QPrune. On LLaVA-1.5-7B (Liu et al., 2024a), QPrune retains only 11.1% of the original visual tokens while preserving 97.29% of the full-model performance, outperforming SparseVLM+ (Zhang et al., 2026) by 7.57 percentage points under the same token budget. Figure 1 further shows that QPrune retains a larger share of full-model performance on most benchmarks under aggressive pruning on both LLaVA-1.5-7B and LLaVA-NeXT-7B.

Our main contributions are summarized as follows:

- We introduce a quantum-inspired mixed-state representation for visual token pruning, where each token is assigned a probabilistic state conditioned on the image–text pair.
- We propose state-aware token pruning that jointly considers token relevance and state overlap with the retained tokens, reducing redundancy during token selection.
- We introduce state-preserving token refinement to supplement local visual information that may be insufficiently represented after pruning.

2 Related Work

2.1 Large Vision-Language Models

Large vision-language models (LVLMs) extend language models to image-grounded understanding and generation (Liu et al., 2023; Li et al., 2023a). Representative systems include LLaVA, LLaVA-1.5, BLIP-2, and Qwen-VL (Liu et al., 2023; 2024a; Li et al., 2023a; Bai et al., 2023), as well as Flamingo, MiniGPT-4, and InstructBLIP (Alayrac et al., 2022; Zhu et al., 2024; Dai et al., 2023). More recent examples include InternVL, Qwen2-VL, and LLaVA-OneVision (Chen et al., 2024b; Wang et al., 2024; Li et al., 2025). Qwen2-VL maps images of different resolutions to varying numbers of visual tokens, making the sequence processed by the language model resolution-dependent (Wang et al., 2024). InstructBLIP conditions visual feature extraction on the input instruction to obtain task-relevant representations (Dai et al., 2023). Together, these developments motivate efficient visual processing that preserves relevant evidence.

2.2 Visual Token Pruning

Visual-token reduction methods differ in pruning stage and selection criterion. Before language-model processing, ToMe, PruMerge, VisionZip, VisPruner, and TRIM reduce the visual sequence through merging or token selection (Bolya et al., 2023; Shang et al., 2025; Yang et al., 2025b; Zhang et al., 2025b; Song et al., 2025). FastV, PyramidDrop, PACT, SparseVLM, SparseVLM+, and ATP-LLaVA instead operate within the language model (Chen et al., 2024a; Xing et al., 2025; Dhouib et al., 2025; Zhang et al., 2025c; 2026; Ye et al., 2025). These methods reduce inference cost, but a token that scores highly on its own may add little information once similar evidence has been retained. Subset-oriented methods address token relations from different perspectives: DivPrune promotes diversity, DART measures duplication, and TopV incorporates spatial structure (Alvar et al., 2025; Wen et al., 2025; Yang et al., 2025a). Determinantal point processes likewise balance quality and diversity at the set level (Kulesza & Taskar, 2012). These approaches highlight why subset composition matters. For visual-token pruning, the remaining question is how to assess a candidate's contribution to the current image–text information state as the retained set grows. Quantum-inspired models use states and density matrices to represent semantic units and probabilistic mixtures (Zhang et al., 2018; 2025a). This perspective motivates QPrune's joint treatment of token contribution and overlap with the retained set.

3 Method

Visual tokens interact with the textual input during multimodal processing, forming a complex information state. Mixed states in quantum information theory provide a natural formulation for representing such a state through multiple token states and their associated probabilities. Motivated by this formulation, we propose **QPrune**, a quantum-inspired visual token pruning method based on mixed-state representation. QPrune represents the visual tokens under the current image-text pair as a mixed state. Based on this representation, state-aware token pruning constructs a main retained set while reducing redundant token representations. State-preserving token refinement further supplements local visual information that is insufficiently represented by the main retained set.

3.1 Mixed-State Representation

Let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ denote the N visual tokens produced by the vision encoder. Following the mixed-state formulation, each visual token is represented by a normalized state:

$$|\psi_i\rangle = \frac{x_i}{\|x_i\|_2}. \tag{1}$$

A mixed state associates each token state with a probability reflecting its contribution under the current image-text pair. Following the Born rule in quantum theory, the probability associated with a token state is computed from its overlap with a corresponding reference state:

$$P(|\psi_i\rangle|\phi) = |\langle\phi|\psi_i\rangle|^2, \tag{2}$$

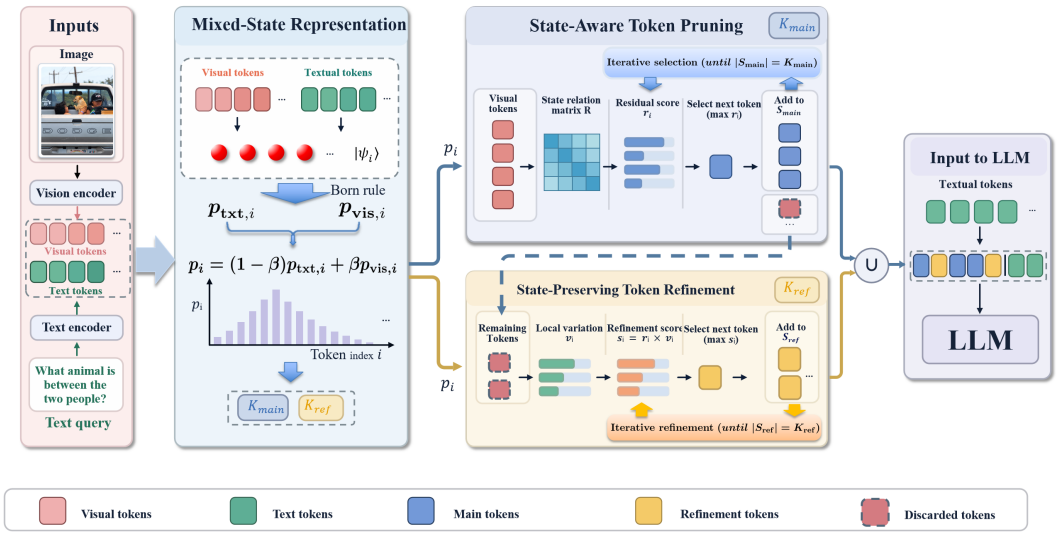


Figure 2: **Overview of QPrune.** QPrune consists of mixed-state representation, state-aware token pruning, and state-preserving token refinement. The mixed-state representation estimates mixed-state probabilities under the current image-text pair, state-aware token pruning reduces redundant tokens by considering token relevance and state overlap, and state-preserving token refinement supplements additional informative visual tokens.

where $|\phi\rangle$ denotes the reference state. The detailed formulation of the Born-rule-based probability estimation is provided in Appendix A.1.

In QPrune, two complementary reference states are constructed to capture image-text conditioned and visual evidence. Accordingly, the corresponding probabilities are computed as:

$$p_{\text{txt},i} = |\langle \phi_{I,T} | \psi_i \rangle|^2, \qquad p_{\text{vis},i} = |\langle \phi_I | \psi_i \rangle|^2, \tag{3}$$

where $|\phi_{I,T}\rangle$ and $|\phi_I\rangle$ denote the image-text conditioned and visual reference states, respectively. The detailed construction of these reference states is provided in Appendix A.1.

The two probabilities are adaptively combined as

$$p_i = (1-\beta)p_{\text{txt},i} + \beta p_{\text{vis},i}, \tag{4}$$

where β controls their relative contributions. The resulting p_i represents the mixed-state probability associated with token state $|\psi_i\rangle$ under the current image-text pair.

The mixed-state probability is used to allocate the retained token budget between the state-aware token pruning and state-preserving token refinement stages. Let K denote the total number of visual tokens to be retained after pruning, where K_{main} and K_{ref} denote the budgets allocated to the state-aware token pruning and state-preserving token refinement stages, respectively. Let $H(p)$ denote the normalized entropy of the mixed-state probabilities. The token budget for the state-aware token pruning stage is determined as:

$$K_{\text{main}} = \gamma(H(p))K, \qquad K_{\text{ref}} = K - K_{\text{main}}, \tag{5}$$

where $\gamma(\cdot)$ determines the budget proportion according to the entropy of the mixed-state probabilities. A more concentrated mixed-state probability distribution allocates more budget to the main pruning stage, while a more dispersed distribution reserves more budget for refinement.

3.2 State-Aware Token Pruning

The mixed-state probability p_i describes the contribution of token i, but selecting tokens only according to p_i may retain multiple tokens with similar states. The token probabilities and state representations obtained from the mixed-state representation are jointly considered to capture token contribution and state overlap.

For tokens i and j, we define their probability-weighted state relation as

$$R_{ij} = \sqrt{p_i p_j}\,|\langle\psi_i|\psi_j\rangle|^2 . \tag{6}$$

The probability term reflects the contributions of the two tokens, while the overlap term measures the similarity between their states. Since each state is normalized, $R_{ii} = p_i$ represents the initial contribution of token i.

The relation matrix R is used to update the residual scores during progressive token selection. For each candidate token, we maintain a residual score r_i, which measures the contribution that has not yet been represented by the retained tokens. Before any token is selected, the score is initialized by its own contribution, i.e., $r_i = R_{ii}$.

At each step, QPrune selects

$$j = \arg\max_{i \notin S} r_i, \tag{7}$$

where S denotes the current retained set.

After token j is retained, its relations with the remaining tokens, given by R, are used to remove the components that are already represented by the updated retained set. The residual scores are updated before the next selection. Consequently, tokens that strongly overlap with the retained states gradually receive smaller scores, while tokens that provide additional state information maintain larger scores. The procedure continues until K_{main} tokens are selected. The resulting tokens form the main retained set S_{main}.

3.3 State-Preserving Token Refinement

State-aware token pruning retains the major components of the mixed state by selecting S_{main}. However, aggressive pruning may leave some mixed-state information insufficiently represented by the selected tokens. The residual scores inherited from the pruning stage quantify such remaining information.

To preserve the residual information in the mixed state, we further refine the retained set by considering the local distinctiveness of remaining tokens. A candidate token is selected when it provides residual information that is not represented by the current set and complementary local visual details.

Specifically, we reuse the token states $|\psi_i\rangle$ defined in Sec. 3.2. Let $\mathcal{N}(i)$ denote the spatial neighborhood of token i in the original visual-token grid. The local variation of token i is measured by comparing its state with those of neighboring tokens:

$$v_i = 1 - \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} |\langle\psi_i|\psi_j\rangle|^2 . \tag{8}$$

The local variation v_i measures the difference between token i and its surrounding tokens. A larger v_i indicates that the token contains more distinctive local visual information, while a smaller v_i suggests that its information is already similar to nearby tokens.

For each remaining token, we combine its residual contribution and local variation:

$$s_i = r_i v_i, \tag{9}$$

where r_i is the residual score inherited from the state-aware pruning stage. Therefore, tokens with both residual state information and complementary local details obtain larger refinement scores.

Starting from S_{main}, QPrune selects refinement tokens according to

$$j_t = \arg\max_{i \notin S} s_i. \tag{10}$$

After each selection, the residual scores are updated using the same state relation R defined in Sec. 3.2. The process continues until K_{ref} additional tokens are selected, forming S_{ref}.

The final retained set is

$$S = S_{\text{main}} \cup S_{\text{ref}}, \qquad |S| = K. \tag{11}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmarks. To comprehensively evaluate QPrune, we conduct experiments on a diverse set of multimodal benchmarks spanning visual question answering, perception, reasoning, text understanding, and hallucination evaluation. These include GQA (Hudson & Manning, 2019), ScienceQA-IMG (Lu et al., 2022), TextVQA (Singh et al., 2019), VQAv2 (Goyal et al., 2017), MMBench (Liu et al., 2024b), MME (Fu et al., 2025), POPE (Li et al., 2023b) and MM-Vet (Yu et al., 2024). We follow the official evaluation protocol and metric of each benchmark.

Model Configurations. Our main experiments are conducted on LLaVA-1.5-7B (Liu et al., 2024a), which contains 576 visual tokens per image. We evaluate QPrune under three token budgets, $K \in \{128, 64, 32\}$. To further examine its generalization to higher-resolution visual inputs, we additionally evaluate QPrune on LLaVA-NeXT-7B with $K \in \{640, 320, 160\}$ retained visual tokens.

Comparison methods. We compare QPrune with representative visual token pruning methods, including FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2025c), SparseVLM+ (Zhang et al., 2026), VisionZip (Yang et al., 2025b), and DART (Wen et al., 2025). All methods use pretrained checkpoints without additional parameter updates, and share the same backbone model and final visual-token budget within each comparison.

Table 1: **Performance comparison of different visual token pruning methods on LLaVA-1.5-7B. Avg.** denotes the average performance across seven benchmarks, with MME divided by 20. **Rel.** denotes the average benchmark-wise performance ratio to the full-token Vanilla model, expressed as a percentage. Bold indicates the best result under each token budget.

Method	Tokens	GQA	MME	POPE	SQA^{IMG}	TextVQA	VQAv2	MM-Vet	Avg.	Rel. (%)
Vanilla	576	61.9	1506.5	85.9	69.5	58.2	78.5	31.3	65.80	100.00
$K = 128$										
FastV (ECCV 2024)	128	54.0	1368.9	68.2	69.2	56.4	71.0	27.0	59.18	90.10
SparseVLM (ICML 2025)	128	57.3	1399.3	83.1	69.0	56.3	75.1	29.7	62.92	95.54
VisionZip (CVPR 2025)	128	57.7	1433.2	83.2	68.9	56.8	75.6	**31.6**	63.64	97.03
DART (EMNLP 2025)	128	57.9	1408.7	80.4	69.1	56.3	74.7	30.9	62.82	95.81
SparseVLM+ (TPAMI 2026)	128	59.5	**1465.5**	85.0	69.2	**57.7**	**77.0**	31.1	64.68	98.36
QPrune (Ours)	128	**59.6**	1448.7	**87.6**	**69.5**	57.1	76.6	**31.6**	**64.92**	**98.72**
$K = 64$										
FastV (ECCV 2024)	64	46.0	973.5	35.5	**70.1**	51.6	55.9	18.9	46.67	71.63
SparseVLM (ICML 2025)	64	52.0	1190.4	69.7	69.2	52.1	66.9	24.4	56.26	85.20
VisionZip (CVPR 2025)	64	55.1	1363.3	77.0	69.0	55.6	72.4	29.4	60.95	92.87
DART (EMNLP 2025)	64	54.7	1365.1	73.8	69.3	54.7	71.3	26.5	59.79	90.58
SparseVLM+ (TPAMI 2026)	64	54.0	1359.1	78.1	69.5	53.9	70.7	24.1	59.75	89.72
QPrune (Ours)	64	**58.5**	**1436.3**	**87.5**	69.8	**55.9**	**75.4**	**30.3**	**64.17**	**97.29**
$K = 32$										
FastV (ECCV 2024)	32	41.5	884.6	32.5	42.6	42.5	43.4	20.7	38.20	59.91
SparseVLM (ICML 2025)	32	48.3	1046.7	67.9	57.3	46.1	58.6	18.6	49.88	74.61
VisionZip (CVPR 2025)	32	51.8	1247.4	68.7	68.8	53.1	67.1	25.5	56.77	86.23
DART (EMNLP 2025)	32	52.9	1273.3	69.1	69.3	52.2	67.1	25.0	57.04	86.45
SparseVLM+ (TPAMI 2026)	32	49.9	1152.7	59.6	**70.1**	51.2	63.6	21.2	53.32	80.59
QPrune (Ours)	32	**57.1**	**1369.2**	**87.9**	69.4	**53.2**	**73.6**	**26.1**	**62.25**	**93.41**

4.2 EXPERIMENTAL RESULTS

Performance Evaluation. QPrune achieves competitive performance across different token budgets and model configurations. As shown in Table 1, QPrune obtains the highest average score under all evaluated token budgets on LLaVA-1.5-7B. With 128 retained visual tokens, QPrune achieves an av-

erage score of 64.92 and preserves 98.72% of the full-model performance. When the token budget is reduced to 64, QPrune maintains an average score of 64.17 with 97.29% performance retention, outperforming SparseVLM+ by 4.42 points. Under the more aggressive setting of retaining only 32 visual tokens, QPrune still achieves an average score of 62.25 and preserves 93.41% of the full-model performance, exceeding SparseVLM+ by 8.93 points. The performance advantage becomes more evident under smaller token budgets. At $K = 32$, QPrune improves the average score over SparseVLM+ from 53.32 to 62.25, indicating that QPrune can better preserve useful visual token information under substantial token reduction. Furthermore, Table 2 presents the results on the higher-resolution LLaVA-NeXT-7B model. With 640 retained tokens, QPrune achieves an average score of 66.31, which is comparable to SparseVLM+ (66.25). When the token budget is reduced to 320, QPrune improves the average score from 63.26 to 65.31 compared with SparseVLM+, while achieving higher performance on six of the seven benchmarks. With only 160 retained tokens, QPrune maintains an average score of 64.02 and preserves 88.96% of the full-model performance, whereas SparseVLM+ retains 76.28%. These results show that QPrune maintains effective performance across different token budgets and higher-resolution visual inputs.

Table 2: **Performance comparison of different visual token pruning methods on LLaVA-NeXT-7B. Avg.** averages the seven benchmark scores, with MME divided by 20. **Rel.** denotes the mean benchmark-wise performance ratio relative to the full-token Vanilla model, expressed as a percentage. Bold indicates the best result within each token budget.

Method	Tokens	GQA	MME	POPE	SQA^{IMG}	TextVQA	VQAv2	MM-Vet	Avg.	Rel. (%)
Vanilla	2880	64.8	1498.0	86.7	72.8	65.7	82.2	47.3	70.63	100.00
$K = 640$										
VisionZip (CVPR 2025)	640	60.1	1467.8	82.4	68.2	**58.4**	79.2	32.5	64.88	90.49
SparseVLM+ (TPAMI 2026)	640	62.1	**1497.4**	86.6	68.4	57.8	**80.9**	**33.1**	66.25	92.29
QPrune (Ours)	640	**62.4**	1483.5	**87.7**	**68.9**	58.0	80.1	32.9	**66.31**	**92.34**
$K = 320$										
VisionZip (CVPR 2025)	320	58.1	1408.2	82.2	67.5	56.2	74.2	31.2	62.83	87.57
SparseVLM+ (TPAMI 2026)	320	59.4	1412.2	81.9	**68.5**	54.8	77.1	30.5	63.26	88.03
QPrune (Ours)	320	**61.1**	**1463.6**	**87.5**	67.9	**57.2**	**78.4**	**31.9**	**65.31**	**90.87**
$K = 160$										
VisionZip (CVPR 2025)	160	54.3	1239.7	74.8	67.8	54.7	67.3	28.1	58.43	81.50
SparseVLM+ (TPAMI 2026)	160	54.5	1187.7	68.1	**68.3**	46.0	68.9	21.0	55.17	76.28
QPrune (Ours)	160	**60.6**	**1392.7**	**86.8**	67.9	**56.1**	**76.8**	**30.3**	**64.02**	**88.96**

Table 3: **Efficiency comparison of visual token pruning methods.** E2E-128 includes token selection, prefill, and the generation of 128 tokens.

Method	Tokens	Prefill (ms) ↓	Decode (ms/tok) ↓	E2E-128 (ms) ↓	FLOPs (T) ↓	KV cache (MB) ↓	POPE ↑
Vanilla	576	156.20	81.92	11120.99	7.634	288.00	85.9
FastV (ECCV 2024)	64	54.17	60.85	10090.08	0.842	32.06	35.5
VisionZip (CVPR 2025)	64	70.86	83.60	11257.66	**0.831**	**32.00**	77.0
SparseVLM+ (TPAMI 2026)	64	967.50	38.77	6027.58	0.846	32.22	78.1
QPrune (Ours)	64	**34.33**	**23.97**	**3450.19**	**0.831**	**32.00**	**87.5**

Efficiency Analysis. We compare the runtime, computational cost, memory usage, and task performance of different pruning methods on LLaVA-1.5-7B with $K = 64$, using identical inference configurations and hardware. As shown in Table 3, QPrune introduces additional token selection computation but substantially reduces the cost of subsequent multimodal inference. Compared with the full-token model, QPrune reduces prefill latency from 156.20 ms to 34.33 ms, decoding latency from 81.92 ms/token to 23.97 ms/token, and the end-to-end time for generating 128 tokens from 11120.99 ms to 3450.19 ms. QPrune also reduces FLOPs and KV-cache usage by 89.1% and 88.9%, respectively, while achieving the highest POPE score among the compared methods.

These results demonstrate that QPrune effectively reduces visual-token computational costs while maintaining task performance.

4.3 Ablation Studies

We conduct ablation studies to investigate the effectiveness of the main components of QPrune, including mixed-state representation, state-aware token pruning, and state-preserving token refinement. All ablation experiments are conducted on LLaVA-1.5-7B and evaluated on six multimodal benchmarks, with Avg. denoting the average performance across these benchmarks.

Effect of Mixed-State Probability. For the mixed-state probability, we conduct ablation experiments on the two probability distributions used to construct it. Specifically, we remove p_{txt} or p_{vis} and compare the resulting variants with the complete QPrune. As shown in Table 4, using only p_{vis} or p_{txt} decreases the average score from 62.97 to 62.27 and 62.34, respectively. These results indicate that neither visual nor text information alone is sufficient to accurately estimate token contributions. Combining the two probability distributions provides complementary image-side and text-related evidence, leading to a more reliable mixed-state probability estimation.

Table 4: **Ablation study of probability sources in the mixed-state representation.**

Variant	GQA	MME	POPE	$\mathrm{SQA}^{\mathrm{IMG}}$	TextVQA	MM-Vet	Avg.
w/o p_{txt}	59.0	1438.4	85.7	68.7	56.9	31.4	62.27
w/o p_{vis}	59.3	1413.3	87.5	68.8	56.5	31.3	62.34
QPrune (Ours)	**59.6**	**1448.7**	**87.6**	**69.5**	**57.1**	**31.6**	**62.97**

Effect of State-aware Token Pruning. We compare QPrune with different token selection strategies. As shown in Table 5, text-relevance Top-K achieves an average score of 58.57, while mixture-probability Top-K improves it to 61.14, demonstrating the benefit of image-text conditioned token contribution. However, mixture probability ignores the redundancy among selected tokens. By considering token state overlap and dynamically updating residual contributions, QPrune further improves the average score to 61.71.

Table 5: **Ablation study of state-aware token pruning.**

Selection strategy	GQA	MMB	MME	POPE	TextVQA	MM-Vet	Avg.
Random selection	52.1	54.3	1312.3	61.8	51.9	21.2	51.15
Text-relevance Top-K	58.5	59.0	1362.1	83.4	53.7	28.7	58.57
Mixture-probability Top-K	59.4	60.4	1446.5	87.0	56.4	31.3	61.14
QPrune	**59.6**	**61.9**	**1448.7**	**87.6**	**57.1**	**31.6**	**61.71**

Effect of State-Preserving Token Refinement. We evaluate different token refinement strategies under the same final token budget. Specifically, we compare random refinement, local variation-based selection, fixed score-based selection, and the proposed iterative update strategy. Random refinement randomly selects additional tokens from the remaining tokens, while local variation-based selection chooses tokens according only to the visual differences between neighboring tokens. The fixed score strategy combines residual information and local variation using $s_i = r_i v_i$, but keeps the score fixed during refinement. As shown in Table 6, random refinement and local variation-based selection achieve average scores of 58.96 and 59.66, respectively. Combining residual information with local variation improves the average score to 59.93. QPrune further increases it to 60.70 through iterative residual updates and achieves the best performance across all six benchmarks. These results support updating token contributions as the retained set grows rather than relying on fixed refinement scores.

Table 6: **Ablation study of state-preserving token refinement.**

Refinement strategy	GQA	MMB	MME	POPE	TextVQA	MM-Vet	Avg.
Random refinement	57.9	58.7	1365.1	84.7	55.1	29.1	58.96
Local variation only (v_i)	58.4	59.1	1396.9	85.0	55.5	30.1	59.66
Fixed score ($s_i = r_i v_i$)	58.2	59.9	1411.1	85.2	55.7	30.0	59.93
QPrune	**58.5**	**60.2**	**1436.3**	**87.5**	**55.9**	**30.3**	**60.70**

4.4 Visualization Analysis

To investigate whether the selected tokens correspond to question-relevant visual regions, we visualize the spatial distribution of selected tokens from different pruning methods for comparison. As shown in Figure 3, QPrune retains more tokens around the glove, which directly corresponds to the queried object and provides visual evidence for identifying its color. In contrast, DART retains tokens more broadly across the image, with fewer tokens covering the glove, and incorrectly predicts its color as white rather than black. This example illustrates how QPrune's question-relevant token selection can preserve evidence needed for correct prediction under a limited token budget.

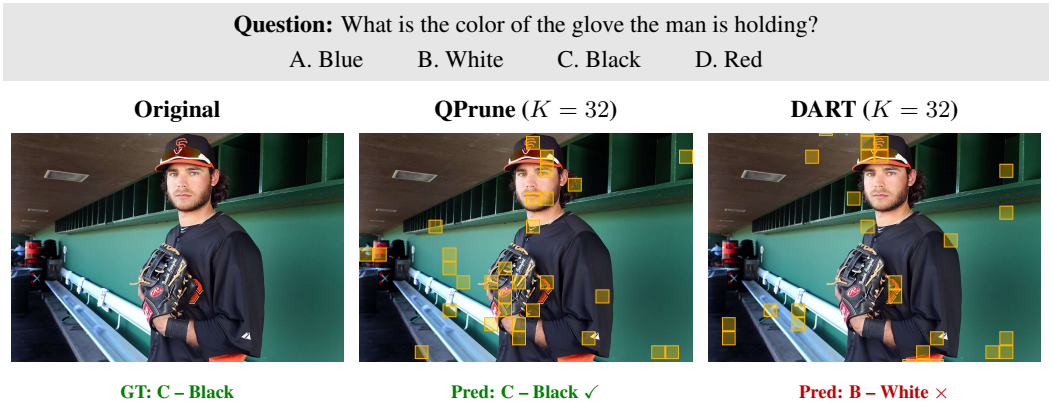


Figure 3: **Comparison of selected visual tokens between QPrune and DART at $K = 32$.** Yellow boxes indicate the selected visual token locations. For the given question about the color of the glove, the glove region contains the key visual evidence for prediction. QPrune selects more tokens around this region and correctly predicts black, whereas DART selects more spatially dispersed tokens and incorrectly predicts white.

5 Conclusion

In this work, we introduced **QPrune**, a quantum-inspired method for visual token pruning in multimodal large language models. QPrune addresses the complex information state formed by visual-textual interactions through a mixed-state representation, enabling token selection based on token relevance and state overlap. State-preserving token refinement further supplements local visual information after pruning. Extensive experiments across multiple benchmarks and model configurations demonstrate that QPrune maintains strong performance with fewer visual tokens. On LLaVA-1.5-7B, QPrune preserves 97.29% of the full-model performance while retaining only 11.1% of the original visual tokens.

AI USE STATEMENT

We used AI tools to assist with language editing, grammar checking, and checks for consistency in wording and reported numerical values. The authors reviewed all AI-assisted suggestions and take responsibility for the accuracy, integrity, and final content of this paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

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

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

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.

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*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven C. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267, 2023.

Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. PACT: Pruning and clustering-based token reduction for faster visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14582–14592, June 2025.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22826–22835, October 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Ivan Kankeu, Stefan Gerd Fritsch, Gunnar Schönhoff, Elie Mounzer, Paul Lukowicz, and Maximilian Kiefer-Emmanouilidis. Quantum-inspired embeddings projection and similarity metrics for representation learning, 2025.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N. Plataniotis, Yao Lu, Song Han, and Zhijian Liu. SparseVILA: Decoupling visual sparsity for efficient vlm inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23784–23794, October 2025.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023b.

Ziming Li, Yan Zhou, Yaxin Liu, Fuqing Zhu, Chuanpeng Yang, and Songlin Hu. QAP: A quantum-inspired adaptive-priority-learning model for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12191–12204. Association for Computational Linguistics, July 2023c.

Yinan Liang, Ziwei Wang, Xiuwei Xu, Jie Zhou, and Jiwen Lu. EfficientLLaVA: Generalizable auto-pruning for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9445–9454, June 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 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 (CVPR)*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024b.

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.

Francois Meyer and Martha Lewis. Modelling lexical ambiguity with density matrices. In *Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)*, pp. 276–290. Association for Computational Linguistics, November 2020.

Wenbo Qiao, Peng Zhang, and ZengLai Ma. A quantum-inspired matching network with linguistic theories for metaphor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 1435–1445. ELRA and ICCL, May 2024.

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

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

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7614–7623. Association for Computational Linguistics, 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, 2024.

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

Sixuan Wu, Jian Li, Peng Zhang, and Yue Zhang. Natural language processing meets quantum physics: A survey and categorization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3172–3182. Association for Computational Linguistics, November 2021.

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

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, and Bo Yuan. TopV: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19803–19813, June 2025a.

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

Sihan Yang, Runsen Xu, Chenhang Cui, Tai Wang, Dahua Lin, and Jiangmiao Pang. VFlowOpt: A token pruning framework for lmms with visual information flow-guided optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23924–23934, October 2025c.

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 (CVPR)*, pp. 24972–24982, June 2025.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 2024.

Chenchen Zhang, Qiuchi Li, Dawei Song, and Prayag Tiwari. Quantum-inspired semantic matching based on neural networks with the duality of density matrices. *Engineering Applications of Artificial Intelligence*, 140:109667, 2025a.

Peng Zhang, Jiabin Niu, Zhan Su, Benyou Wang, Liqun Ma, and Dawei Song. End-to-end quantum-like language models with application to question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857. PMLR, 2025c.

Yuan Zhang, Junpeng Ma, Qizhe Zhang, Chun-Kai Fan, Wenzhao Zheng, Kuan Cheng, Jiwen Lu, and Shanghang Zhang. SparseVLM+: Visual token sparsification with improved text-visual attention pattern. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N. Metaxas, and Licheng Yu. Accelerating multimodal large language models by searching optimal vision token reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29869–29879, June 2025.

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.

A Appendix

A.1 Born-Rule-Based Probability Estimation

This section provides the detailed formulation of the probability estimation used in the mixed-state representation.

As introduced in Sec. 3.1, QPrune associates each normalized visual-token state $|\psi_i\rangle$ with an observation probability under a corresponding reference state. In practice, the image–text conditioned reference state $|\phi_{I,T}\rangle$ and the image-side reference state $|\phi_I\rangle$ are represented implicitly through the observation distributions that they induce over the visual-token states, rather than being explicitly instantiated as additional feature vectors. Accordingly, $p_{\mathrm{txt},i}$ and $p_{\mathrm{vis},i}$ can be viewed as operational estimates of the Born-rule probabilities $|\langle\phi_{I,T}|\psi_i\rangle|^2$ and $|\langle\phi_I|\psi_i\rangle|^2$, respectively.

A.1.1 Image–Text Conditioned Observation Probability

Let the input text contain U valid semantic units $\{q_u\}_{u=1}^{U}$. We first obtain the visual and textual representations in a shared semantic space:

$$e_i = g_v(x_i), \qquad t_u = g_t(q_u), \tag{12}$$

where $g_v(\cdot)$ and $g_t(\cdot)$ denote the fixed visual and textual semantic mappings, respectively. The representations are normalized as

$$\hat{e}_i = \frac{e_i}{\max(\|e_i\|_2, \epsilon)}, \qquad \hat{t}_u = \frac{t_u}{\max(\|t_u\|_2, \epsilon)}. \tag{13}$$

For each visual token, we measure its relative response to different textual semantic units:

$$b_{iu} = \frac{\exp\left(\langle\hat{e}_i, \hat{t}_u\rangle/\tau_r\right)}{\sum_{v=1}^{U}\exp\left(\langle\hat{e}_i, \hat{t}_v\rangle/\tau_r\right)}, \tag{14}$$

where τ_r is the response temperature.

A visual token should receive a large observation probability when it is not only strongly associated with the textual input, but also exhibits a relatively unambiguous semantic correspondence. We therefore consider both the maximum semantic response

$$m_i = \max_u b_{iu} \tag{15}$$

and the entropy of its response distribution

$$H_i = -\sum_{u=1}^{U} b_{iu} \log(b_{iu} + \epsilon). \tag{16}$$

The resulting image–text conditioned evidence score is

$$s_i = m_i \,\mathrm{clip}\left(1 - \frac{H_i}{\log U}, 0, 1\right). \tag{17}$$

This formulation suppresses visual tokens whose responses are broadly distributed across unrelated textual units, while preserving tokens with strong and concentrated semantic responses.

We further normalize the scores across visual tokens:

$$\bar{s}_i = \frac{s_i - \min_j s_j}{\max_j s_j - \min_j s_j + \epsilon}. \tag{18}$$

The image–text conditioned observation probability is then obtained through temperature-controlled normalization:

$$p_{\mathrm{txt},i} = \frac{\exp(\bar{s}_i/\tau_s)}{\sum_{j=1}^{N} \exp(\bar{s}_j/\tau_s)}, \tag{19}$$

where τ_s controls the concentration of the probability distribution over visual tokens.

Therefore, $p_{\mathrm{txt},i}$ operationally represents the observation probability induced by the image–text conditioned reference state $|\phi_{I,T}\rangle$ in Eq. 3. When only one valid semantic unit is available, relative semantic discrimination among textual units is unavailable and we use a uniform distribution as the corresponding fallback.

A.1.2 Image-Side Observation Probability

The image–text conditioned probability emphasizes visual evidence that is directly related to the textual input. However, relying only on such cross-modal evidence may under-represent visually informative regions whose textual responses are relatively weak. We therefore construct a complementary image-side observation distribution from the vision encoder.

Let a_{hi} denote the attention response from the global visual token to the i-th patch token in attention head h. The image-side importance of token i is obtained by aggregating the responses across H_a attention heads:

$$\alpha_i = \frac{1}{H_a} \sum_{h=1}^{H_a} a_{hi}. \tag{20}$$

The corresponding image-side observation probability is

$$p_{\mathrm{vis},i} = \frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j + \epsilon}. \tag{21}$$

Unlike $p_{\mathrm{txt},i}$, which is conditioned on the current textual input, $p_{\mathrm{vis},i}$ captures the intrinsic visual importance of each token within the image. It therefore provides the operational estimate of the observation probability induced by the visual reference state $|\phi_I\rangle$ in Eq. 3.

A.1.3 Adaptive Combination of Observation Probabilities

The two observation distributions provide complementary evidence, but their relative reliability varies across different image–text pairs. When the text-conditioned evidence is consistent with the image-side visual prior, the visual prior can provide a stable complementary signal. In contrast, when the two distributions differ substantially, its contribution should be reduced.

We first define their mean distribution:

$$\bar{p} = \frac{1}{2}\left(p_{\mathrm{txt}} + p_{\mathrm{vis}}\right). \tag{22}$$

Their distributional discrepancy is measured by the Jensen–Shannon divergence:

$$D_{\mathrm{JS}} = \frac{1}{2} D_{\mathrm{KL}}\left(p_{\mathrm{txt}} \| \bar{p}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(p_{\mathrm{vis}} \| \bar{p}\right). \tag{23}$$

We convert the divergence into an agreement coefficient:

$$c = \mathrm{clip}\left(1 - \frac{D_{\mathrm{JS}}}{\log 2}, 0, 1\right), \tag{24}$$

where a larger c indicates stronger agreement between the two evidence sources.

The visual-evidence contribution is then determined adaptively:

$$\beta = \mathrm{clip}\left(\beta_0(0.5 + c), 0, 1\right), \tag{25}$$

where β_0 denotes the base contribution of the image-side prior.

Finally, the mixed-state observation probability used in the main text is

$$p_i = (1 - \beta) p_{\mathrm{txt},i} + \beta p_{\mathrm{vis},i}. \tag{26}$$

Since both p_{txt} and p_{vis} are non-negative normalized distributions, the resulting probabilities also satisfy

$$p_i \geq 0, \qquad \sum_{i=1}^{N} p_i = 1. \tag{27}$$

The task-conditioned mixed state can therefore be written as

$$\rho_x = \sum_{i=1}^{N} p_i |\psi_i\rangle\langle\psi_i|. \tag{28}$$

This representation jointly characterizes the visual-token states and their relative observation probabilities under the current image–text pair.

A.2 Entropy-Based Budget Allocation

To evaluate entropy-based budget allocation, we compare QPrune with fixed refinement ratios (Table 7). Fixed allocation is non-monotonic: a 20% refinement ratio raises the average score from 55.02 without refinement to 55.10, whereas increasing it to 30% lowers the score to 54.70. QPrune's adaptive allocation achieves the highest average score of 55.34 and leads on MMBench, MME, and TextVQA. These results indicate that adapting the refinement budget to each input is more effective than using a fixed ratio.

A.3 Additional Qualitative Analysis

We provide additional qualitative visualizations to further examine how QPrune preserves task-relevant visual evidence under different token budgets. Specifically, we visualize both the two-stage token selection process and the behavior of QPrune under increasingly aggressive pruning. These examples complement the quantitative results by showing which visual regions are retained by different stages of QPrune.

Table 7: **Ablation of token-budget allocation. Avg.** averages the five reported benchmarks.

Budget allocation	GQA	MMB	MME	TextVQA	MM-Vet	Avg.
Fixed refinement: 0%	**58.73**	60.12	1419.35	55.87	29.4	55.02
Fixed refinement: 10%	58.36	59.44	1416.56	55.72	29.9	54.85
Fixed refinement: 20%	58.42	60.12	1414.54	55.84	**30.4**	55.10
Fixed refinement: 30%	58.35	60.03	1404.87	55.66	29.2	54.70
QPrune (entropy-based allocation)	58.50	**60.20**	**1436.30**	**55.90**	30.3	**55.34**

Figure 4 illustrates the two-stage selection process. The state-aware token pruning stage first constructs the main retained set S_{main}, which mainly preserves visual evidence relevant to the current question. The refinement stage then selects an additional set S_{ref} from the remaining tokens to supplement local visual details that are insufficiently represented by the main set. In this example, QPrune allocates 51 tokens to the main set and 13 tokens to refinement, producing a final set of 64 visual tokens. The resulting representation preserves the visual evidence around both the dominoes and the interacting finger, leading to the correct prediction.

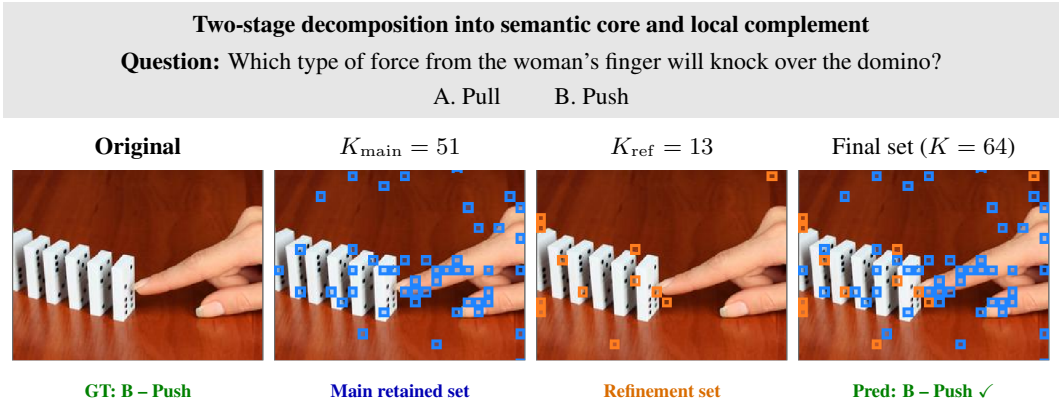


Figure 4: **Visualization of the two-stage token selection process.** State-aware token pruning constructs the main retained set S_{main}, while state-preserving token refinement supplements complementary visual evidence through S_{ref}. The final set contains 64 visual tokens and preserves the evidence needed for the correct prediction.

Figure 5 further visualizes QPrune under different token budgets. As the budget decreases from $K = 128$ to $K = 64$ and $K = 32$, the selected visual tokens become progressively more concentrated around regions relevant to the question. Despite the substantial reduction in visual tokens, evidence around the person and the helmet remains represented, and QPrune produces the correct answer under all three budgets. This observation is consistent with the quantitative results, showing that QPrune can preserve task-relevant evidence even under aggressive visual-token compression.

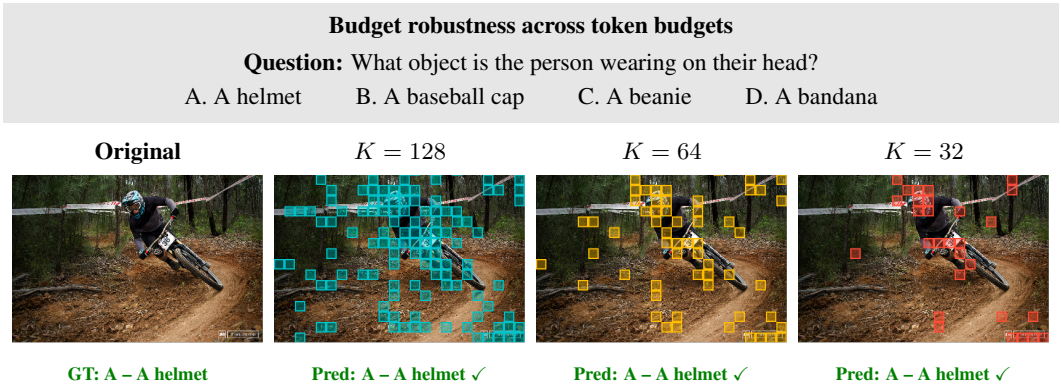


Figure 5: **Qualitative comparison across token budgets.** QPrune retains task-relevant visual evidence as the budget decreases from $K = 128$ to $K = 32$, maintaining the correct prediction under all three budgets.