PrismPrune: Decoupling Saliency and Diversity in Attention for Efficient Visual Token Pruning in VLMs

Supplementary Material

1. Experimental Setup Details

1.1. Evaluation Benchmark

We conduct experiments on a total of 13 widely used visual understanding benchmarks, including 10 image benchmarks and 3 video benchmarks.

1.1.1. Image Benchmark

We conducted experiments on these widely used visual understanding benchmarks. We follow the original configurations of inference settings and evaluation metrics in LLaVA-1.5 for these tasks.

MME. The MME[2] benchmark facilitates a comprehensive evaluation of multimodal models by assessing their perceptual and cognitive capabilities across 14 distinct subtasks. The perception-focused tasks are categorized into Optical Character Recognition (OCR), coarse-grained recognition, and fine-grained recognition. All subtasks are designed as binary judgment problems. We utilize the perception score, derived from a total of 2,374 image-question pairs, for performance evaluation.

TextVQA. The TextVQA[22] benchmark assesses a model's capacity for reading and understanding in-image text by integrating OCR and language comprehension. It utilizes images from Open Images v3[9] depicting text-heavy scenes (e.g., signs, packaging) and provides reference OCR tokens. Questions demand either direct text extraction or contextual reasoning. We evaluate performance on the 5,000 image-question pairs of the validation set.

ScienceQA. The ScienceQA[19] benchmark is designed to evaluate a model's zero-shot generalization capabilities on scientific topics through a multiple-choice question format. The dataset exhibits rich domain diversity, spanning three core subjects: natural science, language science, and social science. To provide a structured assessment, questions are hierarchically organized by topic, category, and skill, encompassing a total of 26 topics, 127 categories, and 379 distinct skills.

MM-Vet. The MM-Vet[28] benchmark is designed to evaluate the integration of core vision-language capabilities. It defines six fundamental capabilities: recognition, OCR, knowledge, language generation, spatial awareness, and mathematics. These capabilities are systematically combined to form 16 composite tasks. A key feature of MM-Vet is its use of a ChatGPT-based evaluator, which provides a unified metric for assessing complex, open-ended answers. The benchmark comprises 218 image-question pairs.

MMBench. The MMBench[18] benchmark provides a comprehensive evaluation of multimodal models through a three-tier hierarchical structure. This hierarchy begins at the highest level with two foundational abilities, perception and reasoning. These are subsequently divided into six specific capabilities at the second level, which culminate in 20 fine-grained tasks at the third level. The benchmark is available in both English and Chinese versions; the English set contains 4,377 image-question pairs, while the Chinese version (**MMBench-CN**) consists of 4,329 pairs. Our evaluation is conducted on both versions.

GQA. The GQA[7] benchmark evaluates a model's capacity for structured reasoning and scene understanding. A key feature of GQA is its use of scene graphs derived from the Visual Genome dataset[11], which provide structured representations of objects, their attributes, and inter-object relationships. The benchmark's questions are programmatically generated from these scene graphs, ensuring that each query is grounded in a specific semantic structure and reasoning path.

VQAv2. The VQAv2[5] benchmark is a standard for evaluating visual understanding through open-ended questions. It is constructed from 265,016 images from the MSCOCO dataset[15], with each image associated with at least three questions. A key characteristic of VQAv2 is its balanced pair design; each question is paired with at least two visually similar images that elicit different answers. This construction discourages models from relying on language priors and instead compels them to ground their reasoning in visual evidence.

POPE. The POPE[13] benchmark is designed to quantify object-level hallucination in large vision-language models. It formulates the evaluation as a binary classification task, posing questions about the presence or absence of specific objects within an image. The images are sourced from the MSCOCO validation set, and the questions are carefully constructed to probe for objects that are either present or contextually plausible but absent.

VizWiz. The VizWiz[6] benchmark is designed to evaluate a model's visual understanding capabilities on in-the-wild images captured by blind individuals. Each data point consists of an image and a corresponding spoken question about its content. This real-world collection process introduces practical challenges, such as variable image quality and instances where questions are unanswerable or not directly relevant to the visual evidence. For evaluation, each question is benchmarked against 10 ground-truth answers.

Method	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	VQA^{v2}	Avg. Rel (%)
*Vanilla, 576 Tokens (**100%**)*									
LLaVA-1.5-13B	1504.4 100.0%	61.3 100.0%	72.8 100.0%	63.2 100.0%	63.2 100.0%	67.7 100.0%	85.9 100.0%	80.0 100.0%	100.0%
LLaVA-1.5-7B	1510.7 100.4%	58.2 94.9%	69.5 95.5%	58.3 92.2%	62.0 98.1%	64.3 95.0%	85.9 100.0%	78.5 98.1%	96.8%
Retain 128 Tokens (↓ 77.8%)									
PrismPrune	1453.3 96.6%	59.4 96.9%	73.8 101.4%	62.8 99.4%	58.8 93.0%	67.5 99.7%	86.0 100.1%	77.4 96.8%	**98.0%**
PrismPrune†	1501.2 99.8%	59.2 96.6%	72.7 99.9%	61.8 97.8%	62.1 98.3%	67.1 99.1%	83.8 97.6%	77.6 97.0%	**98.3%**
Retain 64 Tokens (↓ 88.9%)									
PrismPrune	1406.0 93.5%	58.0 94.6%	73.3 100.7%	61.4 97.2%	57.6 91.1%	65.0 96.0%	81.5 94.9%	75.3 94.1%	**95.3%**
PrismPrune†	1468.5 97.6%	58.4 95.3%	72.8 100.0%	61.1 96.7%	60.4 95.6%	65.8 97.2%	82.2 95.7%	75.8 94.8%	**96.6%**
Retain 32 Tokens (↓ 94.4%)									
PrismPrune	1386.6 92.2%	56.1 91.5%	71.3 97.9%	58.6 92.7%	54.5 86.2%	63.8 94.2%	74.5 86.7%	71.6 89.5%	**91.4%**
PrismPrune†	1428.0 94.9%	56.9 92.8%	70.9 97.4%	59.0 93.4%	57.1 90.4%	64.4 95.1%	75.2 87.5%	72.1 90.1%	**92.7%**

Table 1. **Performance of PrismPrune on LLaVA-1.5-13B.** The vanilla number of visual tokens is 576. PrismPrune† indicates our model after a lightweight fine-tuning step. The first line of each method shows the raw benchmark accuracy, and the second line is the proportion relative to the upper limit. The final column shows the average performance relative to the original model, with the best result in each section highlighted in red.

1.1.2. Video Benchmark

We evaluate our proposed method on three prominent video question answering benchmarks. Our approach is implemented on the Video-LLaVA model. Following established works[14, 20], we employ a GPT-3.5-Turbo-based evaluator for automated scoring of the generated responses. Due to constraints on commercial API usage, and in line with the methodology of [4, 29], our experiments are conducted on the first 1,000 samples from each benchmark's test set.

TGIF-QA. The TGIF-QA[8] benchmark is designed to evaluate video understanding by extending static image-based VQA to the temporal domain. This task necessitates that models perform both spatial reasoning within individual frames and temporal reasoning across them. The benchmark is constructed from 72,000 animated GIFs from the TGIF dataset[12] and includes 165,000 question-answer pairs.

MSVD-QA. The MSVD-QA[26] benchmark is a standard dataset for video question answering, adapted from the MSVD dataset[3], which was originally developed for video captioning. The question-answer pairs are automatically generated from the ground-truth video descriptions, probing a model's comprehension of video content. The benchmark consists of 1,970 short video clips and approximately 50,500 question-answer pairs in total.

MSRVTT-QA. The MSRVTT-QA[26] benchmark is a widely-used dataset for video question answering, built upon the MSRVTT dataset[27]. The benchmark comprises 10,000 video clips and a total of approximately 243,000 question-answer pairs. Questions, categorized into what, who, how, when, and where, require models to process visual and temporal information.

1.2. Implementation Details

For inference, we utilized on 8 NVIDIA A100 GPUs for the GQA and VQAv2 benchmarks, while a single NVIDIA A100 GPU was used for all other datasets. To create our fine-tuned model, we employed a subset of the official LLaVA-1.5[16] instruction tuning data, specifically the VG[10], TextVQA[22], and GQA[7] datasets. The fine-tuning process was conducted on 8 NVIDIA A100 GPUs and completed in approximately 2 hours. Following the

Method	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	VQA^{v2}	Avg. Rel (%)
Vanilla, 2880 Tokens ***(100%)***									
LLaVA-1.6-7B	1513.8 100.0%	59.6 100.0%	70.2 100.0%	57.6 100.0%	62.9 100.0%	64.7 100.0%	86.3 100.0%	81.2 100.0%	100.0%
Retain 640 Tokens (↓ 77.8%)									
VisionZip	1468.4 97.0%	58.5 98.2%	68.1 97.0%	– –	61.3 97.5%	66.3 102.5%	86.3 100.0%	79.1 97.4%	98.5%
PrismPrune	1492.1 98.6%	60.4 101.3%	68.1 97.0%	57.9 100.5%	61.7 98.1%	65.7 101.5%	86.4 100.1%	79.2 97.5%	**99.3%**
Retain 320 Tokens (↓ 88.9%)									
VisionZip	1348.8 89.1%	55.3 92.8%	67.3 95.9%	– –	59.3 94.3%	63.1 97.5%	82.1 95.1%	76.2 93.8%	94.1%
PrismPrune	1432.6 94.6%	58.9 98.8%	68.0 96.9%	56.5 98.1%	59.9 95.2%	63.7 98.5%	83.3 96.5%	77.0 94.8%	**96.7%**
Retain 160 Tokens (↓ 94.4%)									
VisionZip	1239.7 81.9%	54.7 91.8%	68.3 97.3%	– –	55.5 88.2%	60.1 92.9%	74.8 86.7%	71.4 87.9%	89.5%
PrismPrune	1315.8 86.9%	57.2 96.0%	67.8 96.6%	52.9 91.8%	57.5 91.4%	61.5 95.1%	78.2 90.6%	74.1 91.3%	**92.5%**

Table 2. **Performance comparison with VisionZip on LLaVA-1.6-7B.** The vanilla number of visual tokens is 2880. The first line of each method shows the raw benchmark accuracy, and the second line is the proportion relative to the upper limit. PrismPrune consistently outperforms VisionZip across all token reduction levels.

LLaVA-1.5, we trained the model for a single epoch. During this stage, only the parameters of the cross-modality projector were updated, using a learning rate of 2e-5, while all other model components remained frozen.

1.3. Model Architectures

We apply our method to a variety of VLM architectures, including LLaVA-1.5[16], LLaVA-NeXT[17], Video-LLaVA[14], and Qwen2.5-VL[1]. Each model is described as follows.

LLaVA-1.5. LLaVA serves as a foundational paradigm for modern vision-language models, distinguished by its architectural simplicity and strong performance. The model's design features a pre-trained CLIP[21] vision encoder to extract visual features and a Vicuna[30] large language model for text generation and reasoning. A simple yet effective linear projection layer bridges these two modalities, mapping the visual features into the word embedding space of the language model. The LLaVA-1.5 variant builds upon this successful architecture by introducing key enhancements, including an increased input image resolution and a substantially expanded instruction-tuning dataset.

LLaVA-NeXT. LLaVA-NeXT (or LLaVA-1.6) builds upon the successful LLaVA-1.5 architecture but introduces several critical enhancements to address the limitations of prior models. A primary advancement is its support for dynamic high-resolution imagery, allowing the model to process visual inputs at multiple scales (up to 4x the resolution of LLaVA-1.5). This capability, combined with a substantially improved visual instruction tuning dataset, yields marked gains in fine-grained detail recognition and more robust logical reasoning.

Video-LLaVA. Video-LLaVA is a vision-language model designed around the central principle of a unified visual representation. It challenges the conventional approach of using separate encoders for images and videos, which can lead to feature space misalignment and suboptimal learning for the LLM. The core contribution of Video-LLaVA is to unify the tokenization of visual inputs, creating a common representational space for both images and videos that is directly compatible with the language model. By training a single model on a combined corpus of image and video data, Video-LLaVA fosters a synergistic learning process.

Qwen2.5-VL. Qwen2.5-VL is another widely used open-source vision-language model that introduces a native dynamic-resolution Vision Transformer and absolute time encoding. This architecture enables it to process variable-sized images and hour-long videos at their native scale and

Method	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	VQA^{v2}	Avg. Rel (%)
Vanilla, 2880 Tokens ***(100%)***									
LLaVA-1.6-13B	1580.1 100.0%	64.3 100.0%	73.5 100.0%	61.1 100.0%	65.4 100.0%	70.0 100.0%	86.2 100.0%	81.8 100.0%	100.0%
Retain 640 Tokens (↓ 77.8%)									
PrismPrune	1579.6 99.9%	62.2 96.7%	71.1 96.7%	62.8 102.8%	63.0 96.3%	67.9 97.0%	85.4 99.1%	80.0 98.0%	**98.3%**
Retain 320 Tokens (↓ 88.9%)									
PrismPrune	1508.7 95.5%	60.9 94.7%	70.6 96.1%	63.0 103.1%	61.4 93.9%	66.5 95.0%	83.5 96.9%	77.7 95.0%	**96.3%**
Retain 160 Tokens (↓ 94.4%)									
PrismPrune	1457.5 92.2%	58.7 91.3%	70.0 95.2%	61.9 101.3%	59.3 90.7%	65.7 93.9%	79.5 92.2%	74.5 91.1%	**93.5%**

Table 3. **Performance of PrismPrune on LLaVA-1.6-13B.** The vanilla number of visual tokens is 2880. The first line of each method shows the raw benchmark accuracy, and the second line is the proportion relative to the upper limit. The final column shows the average performance relative to the original model.

temporality. As a result, Qwen2.5-VL excels at fine-grained visual grounding, robust document parsing, and long-video comprehension, functioning as a powerful interactive agent for real-world tasks.

2. Additional Experiments

2.1. Performance on LLaVA at Different Scales

Results on LLaVA-1.5-13B. In the main paper, we demonstrate the effectiveness of our model on the 7B scale in Table 1; this section extends our analysis to the 13B model. As shown in Table 1, we evaluate our method across three vision token counts (128, 64, and 32). The results reveal that even when retaining only 32 visual tokens, our method preserves 91.4% of the baseline performance without requiring additional training. This figure increases to 92.7% in the efficient-tuning mode. Similarly, at 64 visual tokens, the model maintains 95.3% performance, which is further elevated to 96.6% with efficient-tuning. Furthermore, when retaining 128 tokens, our method enables the 13B model to outperform the 7B model.

Results on LLaVA-NeXT-13B. We demonstrate the effectiveness of PrismPrune on the LLaVA-NeXT-7B model across a comprehensive suite of benchmarks. For brevity, Table 2 in the main paper presents a curated summary of these results. A complete performance breakdown across all evaluated benchmarks is provided in Table 2. The findings consistently affirm that our proposed PrismPrune method delivers robust and high-quality performance.

Our evaluation is further extended to the more powerful and challenging LLaVA-1.6-13B baseline to demonstrate the generalizability of our method. As presented in Table 3, our approach shows remarkable performance preservation across different compression levels. At a modest 640 tokens, the model retains 98.3% of its original performance. This robustness is maintained at 320 tokens, achieving 96.3% of the baseline. Even under aggressive compression to just 160 tokens, the model still delivers an impressive 93.5% of its full capability, underscoring its high efficiency and robustness.

Method	MME	TextVQA	POPE	Avg. Rel.
LLaVA-1.6-7B	1513.8	59.6	86.3	100.0%
PrismPrune with varying Pool Size (M)				
M = 320 ($2\times K_{div}$)	1454.3	58.7	82.4	96.6%
M = 640 ($4\times K_{div}$)	1432.6	58.9	83.3	**96.7%**
M = 1280 ($8\times K_{div}$)	1432.7	59.0	82.9	96.6%
All (Global FPS)	1403.0	58.7	82.5	95.6%

Table 4. **Ablation study on the candidate pool size M.** All pruned variants are compared against the unpruned Vanilla model (100% baseline).

2.2. Additional Ablation Studies

Ablation study on the candidate pool size. To validate our "filter-then-sample" strategy, we conduct an ablation study on the candidate pool size M. For this experiment, we retain 320 tokens and the token allocation is fixed to $K_{sal} = 160, K_{div} = 160$. The results, presented in Table 4, are benchmarked against the unpruned LLaVA-1.6-7B model, which serves as the 100% baseline. Remarkably, all PrismPrune configurations retain over 96.6% of the vanilla model's average performance, demonstrating the ex-

Method	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	Avg. Rel (%)
*Vanilla, 576 Tokens **(100%)***								
LLaVA-1.5-7B	1510.7	58.2	66.8	58.3	62.0	64.3	85.9	100.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	
Retain 128 Tokens (↓ 77.8%)								
PrismPrune† (Original FT)	1481.6	57.4	68.0	57.7	60.9	63.8	84.2	**99.0%**
	98.1%	98.6%	101.8%	98.9%	98.2%	99.2%	98.0%	
PrismPrune† (1 Dataset FT)	1462.0	56.9	68.2	56.2	60.2	63.5	84.2	98.1%
	96.8%	97.8%	102.1%	96.4%	97.1%	98.8%	98.0%	
PrismPrune† (2 Datasets FT)	1477.5	56.6	67.6	57.1	60.8	63.7	84.1	98.5%
	97.8%	97.3%	101.2%	97.9%	98.1%	99.1%	97.9%	
Retain 64 Tokens (↓ 88.9%)								
PrismPrune† (Original FT)	1424.1	56.0	67.7	56.5	59.0	62.5	83.4	**96.9%**
	94.3%	96.2%	101.3%	96.9%	95.2%	97.2%	97.1%	
PrismPrune† (1 Dataset FT)	1420.2	55.4	67.6	55.8	58.6	61.9	83.3	96.3%
	94.0%	95.2%	101.2%	95.7%	94.5%	96.3%	97.0%	
PrismPrune† (2 Datasets FT)	1421.5	55.5	67.6	56.1	58.9	62.5	83.4	96.6%
	94.1%	95.4%	101.2%	96.2%	95.0%	97.2%	97.1%	
Retain 32 Tokens (↓ 94.4%)								
PrismPrune† (Original FT)	1386.6	55.0	68.0	52.3	56.0	60.7	80.3	**93.7%**
	91.8%	94.5%	101.8%	89.7%	90.3%	94.4%	93.5%	
PrismPrune† (1 Dataset FT)	1377.6	54.1	67.7	52.6	55.5	60.1	80.2	93.2%
	91.2%	92.9%	101.3%	90.2%	89.5%	93.5%	93.4%	
PrismPrune† (2 Datasets FT)	1387.5	54.2	67.6	52.2	56.1	60.4	80.4	93.4%
	91.8%	93.1%	101.2%	89.5%	90.5%	93.9%	93.6%	

Table 5. **Ablation study on the fine-tuning data for PrismPrune on LLaVA-1.5-7B.** We compare the original fine-tuning (Original FT) with versions fine-tuned on one dataset (1 Dataset FT) and two datasets (2 Datasets FT). The first line of each method shows the raw benchmark accuracy, and the second line is the proportion relative to the upper limit. The final column shows the average performance relative to the vanilla model.

K_{sal}	K_{div}	MME	TextVQA	POPE	Avg. Rel.
LLaVA-1.6-7B		1513.8	59.6	86.3	100.0%
PrismPrune with Total Tokens K=320					
240	80	1430.9	58.8	82.2	96.2%
160	160	1432.6	58.9	83.3	96.7%
80	240	1432.0	58.5	84.2	96.8%

Table 6. **Ablation on the ratio of salient to diverse tokens.** We fix the total token budget to $K = 320$ and vary the allocation. The results show that a balanced approach is critical, validating our hybrid strategy.

ceptional efficacy of our methodology.

The findings also reveal a clear trend of diminishing returns. As the pool size M increases, both a 2x ($M = 320$) and 4x ($M = 640$) candidate pool are shown to strike an excellent balance between performance and efficiency. While specific configurations may yield marginal gains on certain benchmarks, we establish the **4x** pool size as the default setting. This choice provides a robustly diverse candidate set for sampling, ensuring consistently high performance across various tasks without being computationally prohibitive. Crucially, all variants of our variance-based filtering strategy significantly outperform the Global FPS baseline, which only retains 95.6% of the original performance. This validates our core hypothesis: attention variance is a highly effective heuristic for curating a compact, high-potential candidate pool.

Ablation study on the token ratio. To determine the optimal allocation between salient (K_{sal}) and diverse (K_{div}) tokens, we conduct an ablation study with a fixed total budget of $K = 320$. The results, detailed in Table 6, reveal an important trade-off between different visual reasoning capabilities.

We observe that the performance is sensitive to the al-

Method	Memory	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	Avg. Rel (%)
Upper Bound, 2880 Tokens (100%)									
Vanilla 13B	36721Mb	1580.1 100.0%	64.3 100.0%	73.5 100.0%	61.1 100.0%	65.4 100.0%	70.0 100.0%	86.2 100.0%	100.0%
Vanilla 7B	18952Mb	1513.8 95.8%	61.3 95.3%	70.2 95.5%	57.6 94.3%	64.2 98.2%	67.9 97.0%	86.4 100.2%	**96.6%**
Retain 320 Tokens (↓ 88.9%)									
PrismPrune	29337Mb	1508.7 95.5%	60.9 94.7%	70.6 96.1%	63.0 103.1%	61.4 93.9%	66.5 95.0%	83.5 96.9%	**96.5%**
PrismPrune-8bit	17043Mb	1520.7 96.2%	61.0 94.9%	70.8 96.3%	63.0 103.1%	60.9 93.1%	66.6 95.1%	83.2 96.5%	**96.5%**
PrismPrune-4bit	10859Mb	1481.9 93.8%	60.3 93.8%	70.6 96.1%	61.4 100.5%	61.6 94.2%	66.6 95.1%	82.8 96.1%	**95.7%**

Table 7. **Performance and Memory Comparison of PrismPrune on LLaVA NeXT 13B with Quantization.** The vanilla number of visual tokens is 2880. The first line of each method shows the raw benchmark score, and the second line shows the performance relative to the Vanilla 13B baseline. The final column shows the average relative performance across all benchmarks. The memory refers to the practical CUDA memory usage on a single NVIDIA A800 GPU for SQA.

location strategy. For instance, the diversity-heavy configuration ($K_{sal} = 80, K_{div} = 240$) excels on the POPE benchmark, which is highly reliant on fine-grained details for hallucination detection. Conversely, configurations that better preserve saliency demonstrate stronger performance on general VQA tasks like MME and TextVQA.

This suggests that while the allocation ratio could be tuned as a task-specific hyperparameter to maximize performance on any single benchmark, a truly robust and generalizable method requires a universal configuration that performs well across the board. Therefore, we adopt the balanced **1:1** allocation as the default setting for LLaVA-NeXT. This choice is the most direct embodiment of our principle of balancing saliency and diversity. It establishes a strong, unified foundation that achieves consistently high performance across different task domains without the need for per-dataset tuning.

Ablation study on the fine-tuning dataset. To compensate for the representational shift induced by PrismPrune, we employ an efficient fine-tuning step that adapts the modality alignment module. We conduct an ablation study to assess the impact of fine-tuning data volume, comparing our original protocol (Original FT) with versions trained on significantly smaller subsets (one and two datasets). The results, presented in Table 5, reveal a crucial insight: increasing the fine-tuning data from one to two datasets yields only marginal gains, indicating that performance quickly saturates. Therefore, the performance improvement is not attributable to learning new information. Instead, it stems from a rapid recalibration, which realigns the model to effectively process the compressed visual representations.

2.3. Additional Efficiency Analysis

Efficiency analysis on the CUDA memory. To evaluate the combined effects of token pruning and model quantization, we conducted a comprehensive analysis on the LLaVA-NeXT 13B model. The results are detailed in Table 7. Applying PrismPrune alone, which reduces the visual token count to 320, yields notable efficiency gains. It lowers the CUDA memory footprint from 36721Mb to 29337Mb, a reduction of approximately 20%, while retaining 96.5% of the baseline's average performance.

Our analysis further demonstrates that PrismPrune integrates effectively with quantization to achieve additional CUDA memory savings. When applying 8-bit quantization, the memory requirement is reduced to 17043Mb, representing a total reduction of over 53% compared to the original model. Notably, this significant memory optimization is achieved with no measurable drop in average performance, as the model maintains the same 96.5% relative score. Proceeding with 4-bit quantization further enhances efficiency by decreasing the memory usage to 10859Mb, which represents a 70% reduction from the vanilla baseline. This is accompanied by a minor performance trade-off, with the model's average accuracy decreasing by less than one percentage point to 95.7%. These results validate that PrismPrune is compatible with and complementary to established quantization techniques.

Efficiency analysis on the inference time. To evaluate the trade-off between performance and inference speed, we applied PrismPrune to the LLaVA-1.6-13B model at three distinct visual token counts: 640, 320, and 160. The total inference time was benchmarked on the TextVQA dataset, with

Method	Total Time	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	VQA^{v2}	Avg. Rel (%)
*Vanilla, 2880 Tokens **(100%)***										
LLaVA-1.6-13B	3073s	1580.1 100.0%	64.3 100.0%	73.5 100.0%	61.1 100.0%	65.4 100.0%	70.0 100.0%	86.2 100.0%	81.8 100.0%	100.0%
LLaVA-1.6-7B	1955s	1513.8 95.8%	59.6 92.7%	70.2 95.5%	57.6 94.3%	62.9 96.2%	64.7 92.4%	86.3 100.1%	81.2 99.3%	95.8%
Retain 640 Tokens (↓ 77.8%)										
PrismPrune	1364s	1579.6 99.9%	62.2 96.7%	71.1 96.7%	62.8 102.8%	63.0 96.3%	67.9 97.0%	85.4 99.1%	80.0 97.8%	**98.3%**
Retain 320 Tokens (↓ 88.9%)										
PrismPrune	1084s	1508.7 95.5%	60.9 94.7%	70.6 96.1%	63.0 103.1%	61.4 93.9%	66.5 95.0%	83.5 96.9%	77.7 95.0%	**96.3%**
Retain 160 Tokens (↓ 94.4%)										
PrismPrune	942s	1457.5 92.2%	58.7 91.3%	70.0 95.2%	61.9 101.3%	59.3 90.7%	65.7 93.9%	79.5 92.2%	74.5 91.1%	**93.5%**

Table 8. **Performance of PrismPrune on LLaVA-1.6-13B.** The vanilla number of visual tokens is 2880. The first line of each method shows the raw benchmark accuracy, and the second line is the proportion relative to the LLaVA-1.6-13B. The final column shows the average performance relative to the original model. "Total Time" represents the actual testing time of the model on the TextVQA benchmark.

the results presented in Table 8.

The findings demonstrate that PrismPrune enables larger models to operate with superior efficiency. Notably, when retaining 640 tokens, the pruned 13B model not only maintains higher performance than the vanilla LLaVA-1.6-7B model, with an average relative accuracy of 98.3% compared to 95.8%, but also achieves significantly faster inference speeds, finishing in 1364s versus 1955s for the 7B model. As the pruning becomes more aggressive, the trade-off remains highly favorable. At a retention of 320 tokens, a reduction of nearly 89%, the model's average performance remains high at 96.3% and still surpasses the 7B baseline with faster inference speed. This is achieved with a 2.8x speedup over the original 13B model, as the inference time decreases from 3073s to 1084s. Even at the most extreme setting of 160 tokens, the model retains 93.5% of its performance while being 3.3x faster.

3. Theoretical Foundation for Attention Variance as a Diversity Metric

In our main paper, we introduce attention variance as a key component of PrismPrune for selecting diverse visual tokens. This section provides a rigorous theoretical justification for this choice. Our rationale is founded on a powerful analogy: we conceptualize the Multi-Head Attention (MHA) mechanism as an intrinsic ensemble system. Grounded in this perspective, we draw upon both the unified theory of ensemble learning and classical statistical analysis to demonstrate that attention variance serves as a principled and efficient proxy for diversity.

3.1. The Multi-Head Attention Mechanism as an Intrinsic Ensemble

Our theoretical framework is predicated on the conceptualization of the Multi-Head Attention mechanism as an intrinsic ensemble system[31]. This perspective is not merely a convenient analogy but is rooted in the foundational design of the Transformer architecture [24]. The MHA module consists of multiple parallel attention heads, each operating in a distinct representation subspace. This architectural choice inherently mirrors the core principle of ensemble learning: aggregating predictions from a diverse set of "experts" to achieve a more robust and comprehensive understanding than any single expert could alone.

In this framework, we establish a direct correspondence between the components of MHA and the elements of an ensemble. Each attention head functions as an ensemble member, providing an independent assessment of the input. The attention scores assigned by a head to the visual tokens are analogous to a member's predictions regarding token importance. Consequently, the mean attention score across all heads represents the ensemble's consensus, indicating tokens of unambiguous, salient importance. This principled view allows us to rigorously apply established theories from the field of ensemble learning to analyze and leverage the properties of MHA, particularly its diversity, as we detail in the following sections.

3.2. Diversity as Variance in Ensemble Theory

A foundational theoretical underpinning for our work is the Unified Theory of Diversity in Ensemble Learning [25]. This framework provides a loss-function-based definition of

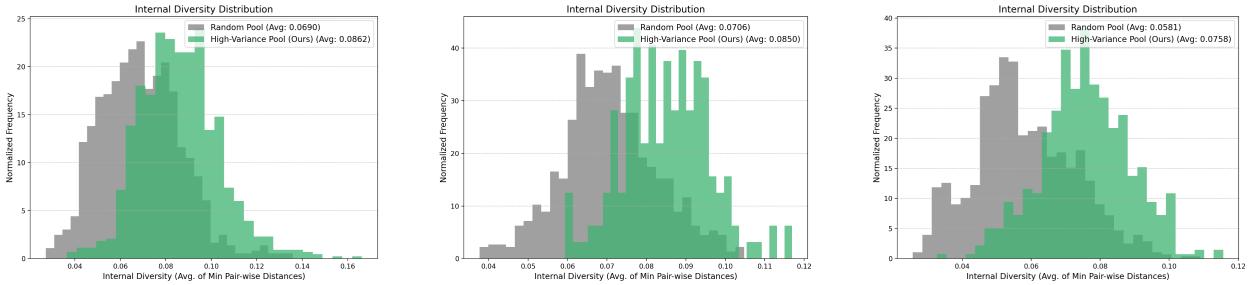

Figure 1. **Histogram of the internal diversity distribution of candidate token pools on different benchmarks.** From left to right, the plots show results on the TextVQA, POPE, and MME datasets. Internal diversity is quantified as the average of minimum pair-wise distances, where higher values are better.

diversity, moving it beyond a purely heuristic concept.

For an ensemble of H members, the diversity is universally defined as the expected average loss between each member's prediction, q_h, and the ensemble's consensus prediction, $\tilde{q}$:

$$\text{Diversity} = E_D \left[\frac{1}{H} \sum_{h=1}^{H} l(\tilde{q}, q_h) \right], \tag{1}$$

where l is the chosen loss function and E_D denotes the expectation over the entire distribution of possible training datasets.

A key result of this theory is its instantiation for the squared loss, where $l(a, b) = (a - b)^2$. In this case, the consensus $\tilde{q}$ is the arithmetic mean of the members' predictions. Substituting this into Eq. 1 yields:

$$\begin{aligned} \text{Diversity}_{\text{sq}} &= E_D \left[\frac{1}{H} \sum_{h=1}^{H} (q_h - \tilde{q})^2 \right] \\ &\equiv E_D \left[\text{Var}(\{q_1, \ldots, q_H\}) \right]. \end{aligned} \tag{2}$$

Eq. 2 provides a powerful and direct conclusion: under the squared loss, diversity is mathematically equivalent to the variance of the members' predictions.

This result provides a direct theoretical basis for our method. Within the analogy, the set of predictions $\{q_1, \ldots, q_H\}$ for a specific visual token i corresponds to the set of attention scores from the CLS token assigned to it by the H heads. Using the notation from our main paper, this set is $\{\mathbf{a}^{\text{cls}}_{1,i}, \ldots, \mathbf{a}^{\text{cls}}_{H,i}\}$.

While Eq. 2 defines diversity as an expectation over the data distribution, we apply this principle at the single-token level during inference. The justification for this instantiation is two-fold. First, the instantaneous variance across heads serves as a practical and efficient proxy for the expected diversity. Second, and more importantly, the disagreement among heads that were jointly trained is particularly informative. The very fact that these co-trained "experts" still produce divergent assessments for a specific token strongly suggests that this token encodes multifaceted, complex, or ambiguous information. Such tokens are precisely the ones that carry high diversity and are crucial to preserve for robust reasoning. Consequently, the variance term in Eq. 2 is precisely the attention variance score, $s_{\text{var},i}$, defined in our main paper:

$$\text{Var}(\{\mathbf{a}^{\text{cls}}_{1,i}, \ldots, \mathbf{a}^{\text{cls}}_{H,i}\}) = s_{\text{var},i}. \tag{3}$$

3.3. Variance as a Proxy for Decorrelation

A complementary justification arises from the classical statistical analysis of ensembles[23, 31]. The variance of an ensemble's mean prediction, $\text{Var}(\bar{Y})$, is given by the well-known formula:

$$\text{Var}(\bar{Y}) = \frac{\sigma^2}{H} + \frac{H-1}{H} \rho \sigma^2, \tag{4}$$

where H is the number of members (heads), σ^2 is the average variance of a single member, and ρ is the average correlation coefficient between any two members. (Note that Eq. 4 concerns the variance of the ensemble's *aggregated output*, which should be minimized for stable predictions, whereas our attention variance measures the *disagreement* among heads about individual tokens, which signals diversity worth preserving. These concepts operate at different conceptual levels.)

As indicated by Eq. 4, the key to improving model robustness and reducing ensemble variance is to foster high diversity among the model's members. This is achieved by minimizing their correlation ρ, as low correlation is synonymous with high diversity.

While ρ is a global, pairwise metric that is computationally expensive to track, the variance of predictions across members for a specific input serves as an efficient and effective local proxy for decorrelation. The relationship between this local variance and the global correlation ρ is inverse. Specifically, a low inter-head variance for a token indicates a strong consensus in the heads' assessments, reflecting high correlation and thus low diversity. Conversely, a high inter-head variance signifies substantial disagreement, revealing a divergence in their interpretations. This

divergence implies a low degree of local correlation and is therefore a strong signal of high diversity.

Thus, attention variance functions as a direct, computationally tractable, inverse proxy for correlation. By selecting tokens with high attention variance, PrismPrune actively favors tokens that induce decorrelation among the heads.

4. Visualization

Visualization of decoupled attention signals. Figure 2 provides a compelling visualization of our core insight: the intrinsic decoupling of saliency and diversity signals within the multi-head attention mechanism. A clear dichotomy emerges from the visual evidence. The tokens selected via mean attention (left column of each pair) consistently concentrate on the primary subjects and semantically salient regions—such as the people at the fruit stand, the double-decker bus, or the skier in the snow. This confirms its role in capturing the semantic core of the image.

In stark contrast, the tokens identified by our high-variance heuristic (right column) capture a rich tapestry of complementary information. They highlight fine-grained details, contextual elements, and structural edges that are often overlooked by mean attention. For instance, high-variance tokens successfully identify textual information on signs, the texture of the snow and surrounding terrain, and the intricate details of items in a cluttered room.

This visual evidence substantiates our central claim: relying solely on saliency is insufficient, as it leads to the loss of crucial contextual information required for complex reasoning. Our work leverages this natural decomposition, akin to a prism separating signals, to ensure that both the core subject and its rich context are preserved efficiently in the final token set.

Visualization of the internal diversity distribution. To validate the effectiveness of our variance-based filtering strategy, we analyzed the internal diversity of the candidate token pools on 1000 data samples from three distinct benchmarks: TextVQA, POPE, and MME. We quantify the internal diversity of a token subset by computing the average of the minimum pair-wise distances between its constituent feature vectors; a higher score indicates less redundancy and greater feature variation. For each sample, we compared the diversity score of the subset selected by our method against a baseline pool of randomly selected tokens.

The histograms of these diversity scores, presented in Figure 1, provide a clear and consistent validation of our approach. Across all three datasets, the distribution for the High-Variance Pool (Ours) is visibly shifted to the right compared to the Random Pool. This demonstrates that our method consistently selects candidate subsets where tokens are, on average, more dissimilar from one another. This higher internal diversity confirms that the tokens identified by our variance heuristic are less redundant and capture a richer, more varied set of visual features. This curated, high-potential candidate pool thus serves as a superior foundation for the subsequent diversity sampling stage, enabling more efficient and effective token pruning.

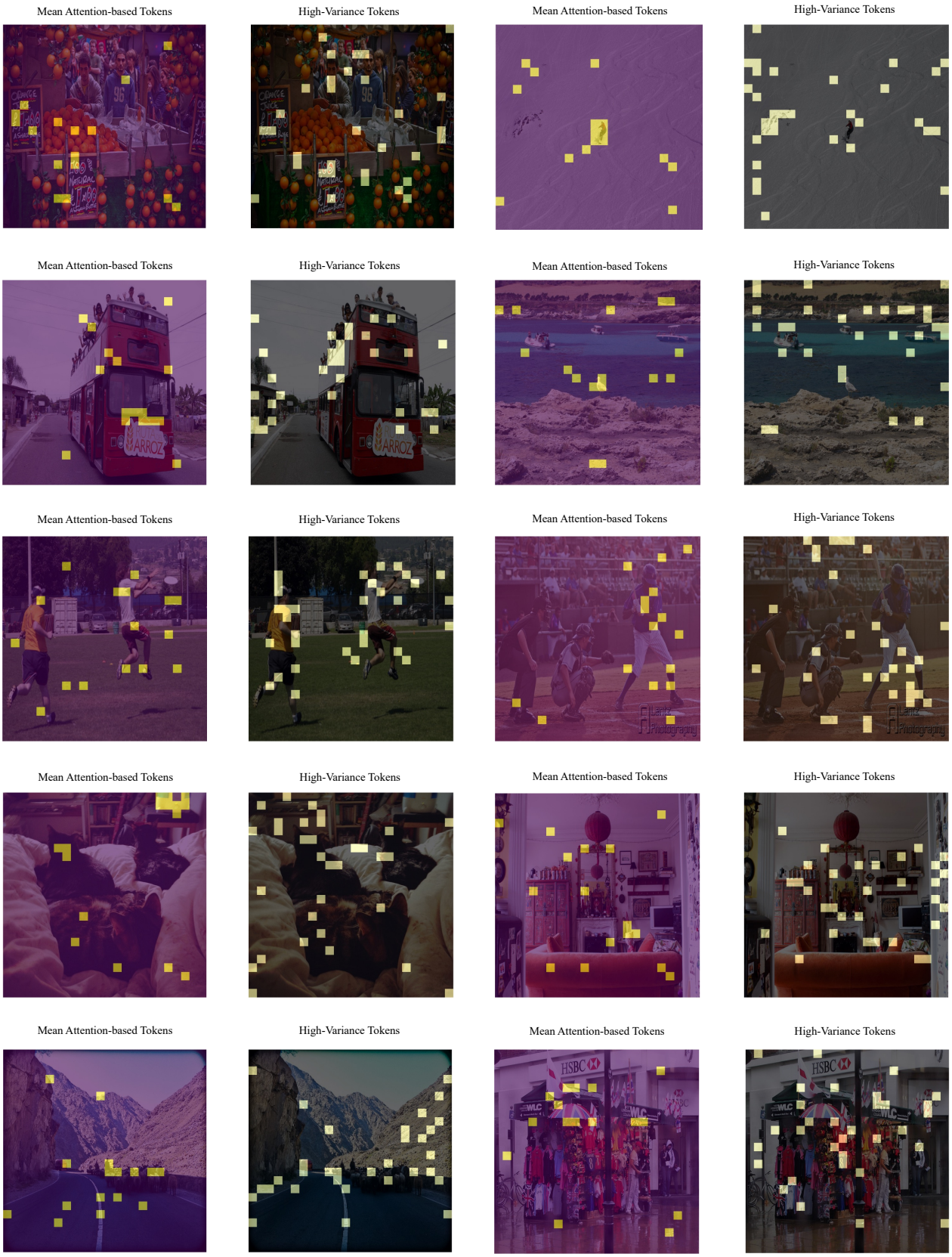

Figure 2. **More visualization of Decoupled Attention Signals.**

References

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

[2] Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3, 2023. 1

[3] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2

[4] 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, 2024. 2

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[6] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1

[7] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 2

[8] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 2

[9] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 1

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 2

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

[12] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 2

[13] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[14] 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. *arXiv preprint arXiv:2311.10122*, 2023. 2, 3

[15] 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*, pages 740–755. Springer, 2014. 1

[16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 2, 3

[17] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 3

[18] 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*, pages 216–233. Springer, 2024. 1

[19] 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. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 1

[20] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[22] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1, 2

[23] Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *International Conference on Neural Networks (ICNN'96)*, pages 90–95. IEEE, 1996. 8

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 7

[25] Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359):1–49, 2023. 7

[26] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2

[27] 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*, pages 5288–5296, 2016. 2

[28] 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. 1

[29] 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*, pages 20857–20867, 2025. 2

[30] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 3

[31] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012. 7, 8