FlashVLM: Text-Guided Visual Token Selection for Large Multimodal Models

Kaitong Cai[1,*], Jusheng Zhang[1,*,†],
Sizhuo Ma[2], Bingqian Lu[2], Jian Wang[2], and Keze Wang[1,†]

[1] Sun Yat-sen University
{jushengzhang88889,kezewang}@gmail.com
[2] Snap Inc.

[*] Equal contribution. [†] Corresponding authors.

Abstract. Large Vision-Language Models (VLMs) typically process hundreds or even thousands of visual tokens per image or video frame, incurring quadratic attention costs and significant information redundancy. Existing token-reduction methods either ignore the textual query or rely on deep attention maps, which become unstable under aggressive pruning and often lead to degraded semantic alignment. To address this, we introduce **FlashVLM**, a text-guided visual token selection framework that dynamically adapts visual inputs to the given query. Rather than relying on noisy attention weights, FlashVLM computes an explicit cross-modal similarity between projected image tokens and normalized text embeddings within the LLM space. It then fuses this extrinsic text-image relevance with intrinsic visual saliency using log-domain weighting and temperature-controlled sharpening. Furthermore, a diversity-preserving partitioning mechanism retains a minimal yet representative set of background tokens to preserve global context. Evaluated under identical token budgets and protocols, FlashVLM achieves *beyond-lossless* compression—slightly surpassing the unpruned baseline while discarding up to 77.8% of visual tokens on LLaVA-1.5. Remarkably, it maintains 92.8% performance even under an aggressive 94.4% compression rate. Extensive experiments across 14 image and video benchmarks demonstrate that FlashVLM establishes a new state-of-the-art in efficiency-performance trade-offs, exhibiting strong robustness and broad generalization across mainstream VLMs.

Keywords: Vision-Language Models · Token Pruning · Efficient Deep Learning · Cross-Modal Alignment

1 Introduction

Large Vision Language Models (VLMs), e.g., LLaVA [25, 26, 37] and GPT-4V [32], have achieved remarkable success by coupling powerful vision encoders with Large Language Models (LLMs) [20, 28, 37, 39, 40, 42, 48, 54, 56, 58]. The prevailing paradigm converts each image or video frame into a long sequence of visual tokens—often hundreds (e.g., $24 \times 24 = 576$) or even thousands (e.g.,

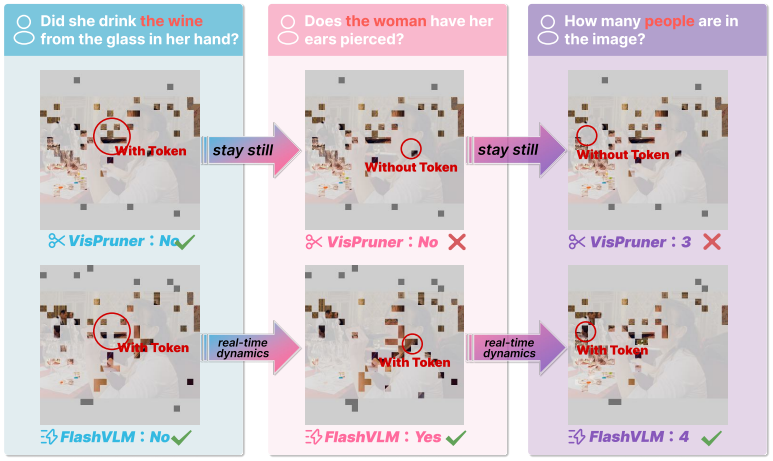

Fig. 1: VisPruner keeps its attention static and insensitive to question semantics, while FlashVLM dynamically selects key tokens based on textual cues, achieving more accurate region focus and stronger semantic alignment.

over 2K tokens for multi-frame videos)—and then concatenates them with text tokens. However, this "feed-all-tokens" design is fundamentally *token-inefficient*: the self-attention in LLMs scales quadratically ($O(L^2)$) with the total sequence length, so every additional visual token inflates both computational and memory costs [7, 8, 33, 40, 53, 55]. This inefficiency is particularly acute for high-resolution inputs and video understanding, which are known to contain substantial spatial and temporal redundancy [2, 35, 56].

Beyond raw computation, this holistic representation also creates a *semantic redundancy* problem. In most real-world scenarios, the user's query is highly localized, focusing on a small subset of objects or regions (e.g., "What is the dog on the left doing?"). Yet existing VLMs must still propagate and attend over all visual tokens, including hundreds of query-irrelevant patches from backgrounds, skies, or distractor objects. These tokens not only waste the attention budget but also dilute focus and introduce noise into the reasoning process. Empirically, we observe that judiciously filtering such noisy tokens can lead to a counter-intuitive phenomenon: *"beyond-lossless"* compression, where performance can slightly exceed that of the unpruned baseline under identical evaluation protocols [10, 13, 16, 49, 50, 57], precisely because irrelevant distractors are suppressed.

Problem and Desiderata. We therefore study the following challenge: given a sequence of visual tokens $\mathbf{v} = \{v_1, \ldots, v_N\}$ from a frozen vision encoder and a text query x, how can we select a compact subset $\mathbf{v}_S$ to feed into the LLM such that task performance is preserved or improved, the selection procedure adds negligible overhead, and the mechanism remains architecture-agnostic without modifying internal transformer layers? In particular, we focus

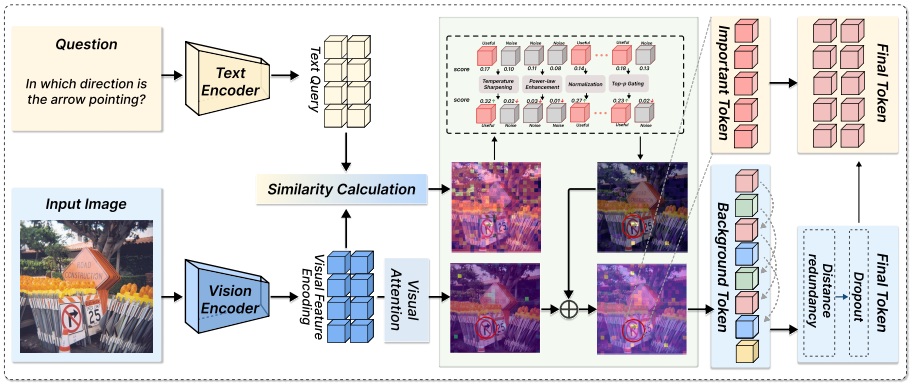

Fig. 2: The model encodes text and image features, then combines cross-modal similarity with visual attention to partition visual tokens into important, background, and diverse tokens. A semantic reweighting and redundancy elimination module selects the final key tokens from these groups, enabling efficient and semantically aligned visual reasoning.

on a *single-shot* selection scheme at the encoder–decoder interface, a far more deployable approach than per-layer sparsification.

Existing visual token reduction techniques struggle to satisfy these desiderata simultaneously. They generally fall into three families:

Text-agnostic pruning. Methods such as ToMe [3], VisionZip [44], and VisPruner [59] rely solely on intrinsic visual cues to merge or drop tokens. While efficient, they are inherently query-agnostic: subtle yet query-relevant regions may be removed, causing brittle performance under heavy compression.

Text-guided pruning via attention. Approaches like FastV [5] and SparseVLM [60] use cross-/self-attention maps to derive relevance, but such signals are often unstable under aggressive pruning due to head sparsity and positional bias. Many also operate inside the LLM, performing layer-wise sparsification that increases overhead and breaks compatibility with optimized kernels such as FlashAttention.

Structural merging and layer-wise sparsification. A third line compresses tokens within each transformer block via merging or clustering. However, this requires architectural modifications, complicating deployment and limiting portability across VLM backbones.

Designing a token selector that is simultaneously query-aware, stable, efficient, and non-intrusive thus remains an open challenge. In this paper, we introduce **FlashVLM**, a text-guided visual token selection framework that dynamically and efficiently tailors the visual input to the LLM. Unlike prior relevance estimators that rely on deep attention maps, FlashVLM computes an *explicit cross-modal similarity* between projected image tokens and normalized text embeddings in the LLM space. This attention-light design yields a stable and interpretable query signal. Concretely, FlashVLM fuses an "extrinsic" query–token relevance score with "intrinsic" visual saliency. By balancing se-

mantic relevance with structural priors, FlashVLM robustly preserves not only the directly queried anchor objects but also the implicitly related contextual cues and numerical instances required for complex multi-hop reasoning and counting tasks. A diversity-preserving selection step further mitigates local redundancy by retaining a non-redundant subset of background tokens to preserve essential global context. Crucially, the entire procedure is performed once at the encoder–decoder boundary, requires no modifications to transformer blocks, and remains fully compatible with FlashAttention.

We evaluate FlashVLM on 14 multimodal datasets spanning image and video QA. Under a unified evaluation protocol with identical token budgets, FlashVLM delivers exceptional efficiency–performance trade-offs: on LLaVA-1.5 it achieves **100.60%** relative accuracy while pruning 77.8% of visual tokens, slightly outperforming the unpruned model. It also remains highly robust under 94.4% compression, with consistent gains across LLaVA, Qwen-VL, InternVL, and CogVLM for both image and video tasks.

Our main contributions are threefold:

- **FlashVLM.** We propose a lightweight text-guided token selector that fuses intrinsic saliency with embedding-based query relevance, offering a stable, attention-independent signal with minimal overhead.
- **State-of-the-Art Compression.** FlashVLM achieves "beyond-lossless" performance at moderate budgets and maintains structural integrity under extreme pruning, all evaluated under strictly identical protocols.
- **Broad Generality.** Our method consistently improves diverse VLM backbones across image and video tasks without modifying transformer layers or requiring task-specific tuning.

2 Related Work

Visual Token Reduction and Efficiency in VLMs. Recent Vision-Language Models (VLMs) [20, 28, 39, 42, 45, 51, 52] encode high-resolution images or multi-frame videos into dense patch embeddings, resulting in thousands of visual tokens and quadratic attention cost. To address this, a series of token compression and merging strategies has been explored. Early works such as ToMe [3] and VisionZip [44] compress tokens through intra-image redundancy analysis, while VisPruner [59] introduces saliency-guided pruning for adaptive visual sparsity. However, these approaches are largely text-agnostic and fail to consider the semantic relevance of tokens to the input query, leading to the removal of semantically critical regions. Other compression efforts [23, 33] focus on attention-based clustering, but their performance deteriorates under aggressive pruning due to instability in attention weights and limited cross-modal guidance.

Text-Guided Token Selection and Cross-Modal Relevance. Recent advances have incorporated textual cues into token selection to improve query awareness. FastV [5] and SparseVLM [60] adopt attention-based cross-modal

relevance estimation to retain query-related regions. Despite improved alignment, attention-derived signals are inherently unstable—affected by head sparsity, query-length sensitivity, and noise accumulation in deep layers. Other approaches leverage gradient-based saliency [30,36] or CLIP-based similarity [9,34] to integrate linguistic guidance. However, these methods require additional backpropagation or suffer from semantic drift when vision–language features are not co-embedded. In contrast, FlashVLM introduces an explicit and attention-light similarity formulation between projected visual tokens and normalized text embeddings in the LLM space, enabling stable, interpretable, and low-cost relevance estimation even under heavy compression.

Balancing Efficiency and Semantic Fidelity. Beyond pruning, a concurrent research line explores maintaining global context and semantic coherence after compression. PruMerge+ [37] and VisionZip [44] attempt to preserve holistic representations via clustering or token regrouping, while VisPruner [59] employs lightweight reweighting to stabilize saliency maps. However, these approaches often exhibit *local collapse*, i.e., focusing too narrowly on high-response regions while losing background cues critical for reasoning. FlashVLM mitigates this issue through diversity-preserving partitioning, ensuring that both important and representative background tokens are maintained.

3 Methodology

Our **FlashVLM** introduces a dynamic, text-guided mechanism for selecting visual tokens during inference. Given a frozen vision encoder and an LLM, FlashVLM replaces the full visual token sequence with a compact, query-aware subset [9, 19, 21]. This strategy aims to preserve task performance while significantly reducing computational overhead. By discarding query-irrelevant tokens before they enter the LLM, FlashVLM reduces the quadratic attention cost and suppresses visual noise. As shown in our experiments (Sec. 4), the method remains empirically robust even under high compression ratios. As illustrated in Fig. 3, our method comprises two principal stages:

Query-Guided Relevance Fusion (Sec. 3.1). We derive a fused relevance score S_{fused} for each visual patch by combining *intrinsic* visual saliency (query-agnostic) with *extrinsic* query relevance (query-conditioned) through log-domain fusion. While our formulation is compatible with various similarity measures [15, 31], we employ dot-product similarity by default for its efficiency and stability.

Diversity-Preserving Partitioning (Sec. 3.2). Given S_{fused}, we partition tokens into an "important" set and a "residual" set. The important set contains top-scoring tokens, while an iterative pruning procedure over the residual set retains a small, non-redundant subset of background tokens to maintain global context.

Crucially, FlashVLM operates in an *attention-light* manner: all relevance signals are computed directly from feature embeddings without modifying or repeatedly sparsifying internal attention maps.

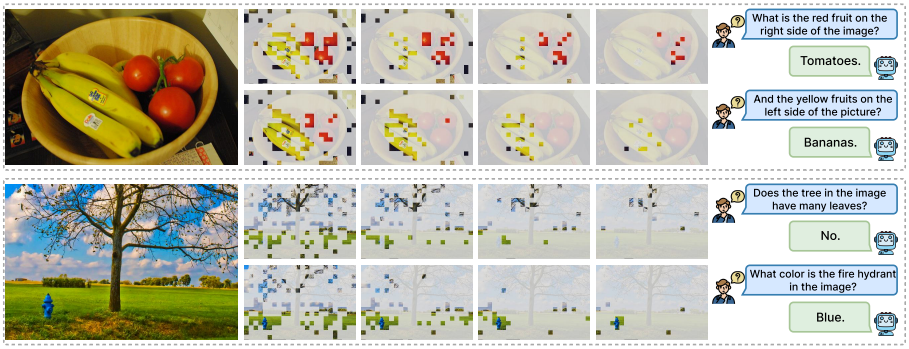

Fig. 3: The model's focus shifts according to each question: in the fruit image, it attends to tomatoes when asked about "the red fruit" and switches to bananas for "the yellow fruits"; in the outdoor scene, it focuses on the tree crown for leaf-related queries and moves to the blue fire hydrant when asked about its color, demonstrating strong question-dependent dynamic adaptation.

3.1 Query-Guided Relevance Fusion

Let N visual patch features $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^{N} \in \mathbb{R}^{N \times D_v}$ be extracted from the vision tower, and L_t text token embeddings $\mathbf{T}_{\text{raw}} = \{\mathbf{t}_j\}_{j=1}^{L_t} \in \mathbb{R}^{L_t \times D_{\text{llm}}}$ be generated by the LLM's embedder. D_v and D_{llm} represent the visual and LLM feature dimensions, respectively.

Intrinsic Visual Saliency ($S_{\text{intrinsic}}$) quantifies the query-agnostic importance of each patch. We derive this signal, A, from the vision encoder's attention mechanism (e.g., from the final layer) by averaging attention weights across all heads. To normalize the scores, we apply min–max normalization N_{minmax}:

$$S_{\text{intrinsic}} = N_{\text{minmax}}(A) = \frac{A - \min(A)}{\max(A) - \min(A) + \epsilon}, \tag{1}$$

where ϵ is a small constant (e.g., 10^{-6}) for numerical stability. This score reflects inherent visual prominence independent of the current query. Crucially, because it is derived from the vision encoder's dense self-attention, it inherently captures local context and multi-object relationships. This prevents the model from collapsing its focus solely onto explicitly queried anchor objects, thereby preserving the indirectly relevant visual cues necessary for complex multi-hop reasoning.

Extrinsic Query Relevance ($S_{\text{extrinsic}}$) measures how well each patch aligns with the text query in the shared LLM space, providing a complementary cross-modal signal.

Projection and Contextual Query Gating. We first project visual features into the LLM space using the multimodal projector W_{proj} and apply ℓ_2-normalization:

$$\mathbf{V}_{\text{proj}} = N_{L_2}\big(W_{\text{proj}}(\mathbf{V})\big) \in \mathbb{R}^{N \times D_{\text{llm}}}. \tag{2}$$

Rather than relying on the raw vector magnitude to estimate token importance—which overlooks the highly contextualized nature of LLM embeddings—we

extract the inherent text-to-text self-attention weights A_{text} from the LLM's early layers. This allows the model to naturally gate tokens based on syntactic and semantic dependencies. For each token $\mathbf{t}_j$, the context gate g_j is defined as its aggregated attention sink score:

$$g_j = \frac{\sum_{k=1}^{L_t} A_{\text{text}}[k, j]}{\max_m \sum_{k=1}^{L_t} A_{\text{text}}[k, m] + \epsilon}. \tag{3}$$

The gated embeddings $\mathbf{T}_{\text{gated}} = \{g_j \cdot \mathbf{t}_j\}_{j=1}^{L_t}$ are subsequently ℓ_2-normalized to yield $\mathbf{T}_{\text{norm}} = N_{L_2}(\mathbf{T}_{\text{gated}})$. This dynamic modulation weights token contributions according to the model's own semantic parsing rather than arbitrary heuristics. In the absence of a text query, $S_{\text{extrinsic}}$ is set to zero, and the method gracefully degrades to attention-only pruning.

Similarity Aggregation. We compute a cross-modal similarity matrix $S_{\text{cross}} \in \mathbb{R}^{N \times L_t}$ via dot product:

$$S_{\text{cross}} = \mathbf{V}_{\text{proj}} \cdot \mathbf{T}_{\text{norm}}^{\top}. \tag{4}$$

For each visual patch $\mathbf{v}_i$, we then aggregate its similarity to all text tokens into a single score, $s_{\text{text}}[i]$, using a temperature-controlled weighted sum:

$$w_{ij} = \text{softmax}_j\big(S_{\text{cross}}[i, j]/\tau_{\text{agg}}\big), \quad s_{\text{text}}[i] = \textstyle\sum_{j=1}^{L_t} w_{ij}\, S_{\text{cross}}[i, j], \tag{5}$$

where $\tau_{\text{agg}} = 0.05$ is chosen empirically to sharpen intra-query attention (see ablations in Sec. 4.3).

Adaptive Relevance Filtering. To separate query-relevant patches from low-signal backgrounds without relying on rigid hyperparameter heuristics, we apply an adaptive, parameter-free statistical filter to the aggregated similarity scores s_{text}. We calculate the batch mean μ and standard deviation σ of the scores, and use standardization to suppress values that fall below a dynamically calculated threshold:

$$s_{\text{filtered}}[i] = \max\left(0, \frac{s_{\text{text}}[i] - \mu(s_{\text{text}})}{\sigma(s_{\text{text}}) + \epsilon}\right). \tag{6}$$

This approach naturally adjusts to the varying score distributions of different queries and images. The filtered result is then min–max normalized to produce the final extrinsic score:

$$S_{\text{extrinsic}} = N_{\text{minmax}}(s_{\text{filtered}}). \tag{7}$$

Log-Domain Fusion. To ensure that selected tokens are both visually salient and text-relevant, we fuse $S_{\text{intrinsic}}$ and $S_{\text{extrinsic}}$ using a weighted geometric mean implemented in the log domain:

$$S_{\text{fused}} = N_{\text{minmax}}\Big(\exp\Big((1-\eta)\,\log(S_{\text{intrinsic}} + \epsilon) + \eta\,\log(S_{\text{extrinsic}} + \epsilon)\Big)\Big), \tag{8}$$

where $\eta = 0.5$ (by default) balances the contributions of intrinsic and extrinsic signals. This fusion strategy prioritizes tokens that score high under both modalities, acting as a soft geometric compromise between visual saliency and semantic relevance.

3.2 Diversity-Preserving Token Partitioning

Given a total token budget T_{keep}, we partition the visual tokens into T_{imp} *important* tokens and T_{div} *diverse background* tokens. By default, we use a balanced allocation where $T_{\text{imp}} = T_{\text{div}} = 0.5 \times T_{\text{keep}}$. This bifurcated design ensures we preserve the most semantically relevant patches while maintaining a complementary set of background tokens for global contextual coverage.

Important Token Selection. We first sort all patches based on their fused relevance score S_{fused} in descending order. The top T_{imp} indices are exclusively selected to form the important token set:

$$I_{\text{sorted}} = \underset{\text{desc}}{\text{argsort}}(S_{\text{fused}}), \quad I_{\text{imp}} = I_{\text{sorted}}[: T_{\text{imp}}], \quad I_{\text{candidate}} = I_{\text{sorted}}[T_{\text{imp}} :]. \tag{9}$$

Diversity-Preserving Residual Selection. The remaining candidate pool, $I_{\text{candidate}}$, often contains redundant or near-duplicate background patches. To enforce diversity without incurring the prohibitive $O(N^2)$ cost of full pairwise similarity computation, we perform an iterative bipartite pruning over the ℓ_2-normalized features $\mathbf{V}_{\text{cand}} = N_{L_2}(\mathbf{V}[I_{\text{candidate}}])$.

To ensure reproducibility and spatial uniformity, we deterministically split the candidate set into two disjoint subsets, $\mathbf{V}_{\text{even}}$ and $\mathbf{V}_{\text{odd}}$, based on their flattened 2D spatial sequence indices. At each iteration, we compute the cross-subset cosine similarity matrix between $\mathbf{V}_{\text{even}}$ and $\mathbf{V}_{\text{odd}}$. We identify the specific token pair exhibiting the maximum similarity (i.e., highest redundancy) and drop the token possessing the lower fused relevance score S_{fused}. This greedy removal dynamically updates the subset. The iterative pruning terminates strictly when the total number of remaining tokens in the candidate pool reaches the target budget T_{div}. This produces a compact, structurally diverse background representation with minimal overhead.

4 Experiments

Datasets. We evaluate the effectiveness of FlashVLM through comprehensive benchmarking on 14 authoritative multimodal datasets. The evaluation covers 10 image-based benchmarks, including VQAv2 [12], GQA [17], VizWiz [14], ScienceQA-IMG [30], TextVQA [38] for visual question answering, as well as POPE [22], MME [11], MMBench [29], MMBench-CN, and MMVet [46] for holistic multimodal assessment. In addition, we benchmark on 4 video-based question answering datasets: TGIF-QA [18], MSVD-QA [43], MSRVTT-QA, and

Method	VQAv2	GQA	VizWiz	SQA-IMG	TextVQA	POPE	MME	MMB	MMB-CN	MMVet	ACC
Upper Bound	**78.5**	**62**	**50**	**66.8**	**58.2**	**85.9**	**1510.7**	**64.3**	**58.3**	**31.1**	**100.0%**
Keep 128 Tokens (Prune 77.8%)											
ToMe	63	52.4	50.5	59.6	49.1	62.8	1088.4	53.3	48.8	27.2	83.9%
FastV	61.8	49.6	51.3	60.2	50.6	59.6	1208.9	56.1	51.4	28.1	85.4%
SparseVLM	73.8	56.0	51.4	67.1	54.9	80.5	1376.2	60.0	51.1	30.0	94.4%
PruMerge+	74.7	57.8	52.4	67.6	54.3	81.5	1420.5	61.3	54.7	28.7	95.8%
VisionZip	75.6	57.6	52.0	68.9	56.8	83.2	1432.4	62.0	56.7	32.6	98.4%
VisPruner	75.8	58.2	52.7	69.1	57.0	84.6	1461.4	62.7	57.3	33.7	99.7%
FlashVLM	**76.4** (+0.6)	**58.9** (+0.7)	**53.1** (+0.4)	**69.5** (+0.4)	**57.6** (+0.6)	**85.1** (+0.5)	**1483.2** (+21.8)	**63.2** (+0.5)	**57.6** (+0.3)	**34.1** (+0.4)	**100.6%** (+0.9)
Keep 64 Tokens (Prune 88.9%)											
ToMe	57.1	48.6	50.2	50.0	45.3	52.5	922.3	43.7	38.9	24.1	73.9%
FastV	55.0	46.1	50.8	51.1	47.8	48.0	1019.6	48.0	42.7	25.8	75.9%
SparseVLM	68.2	52.7	50.1	62.2	51.8	75.1	1221.1	56.2	46.1	23.3	86.4%
PruMerge+	67.4	54.9	52.9	68.6	53.0	77.4	1198.2	59.3	51.0	25.9	90.6%
VisionZip	72.4	55.1	52.9	69.0	55.5	77.0	1365.6	60.1	55.4	31.7	95.6%
VisPruner	72.7	55.4	53.3	69.1	55.8	80.4	1369.9	61.3	55.1	32.3	96.6%
FlashVLM	**73.6** (+0.9)	**56.1** (+0.7)	**53.6** (+0.3)	**69.3** (+0.2)	**56.1** (+0.3)	**81.7** (+1.3)	**1383.4** (+13.5)	**61.8** (+0.5)	**55.8** (+0.7)	**32.9** (+0.6)	**97.9%** (+1.3)
Keep 32 Tokens (Prune 94.4%)											
ToMe	46.8	43.6	51.3	41.4	38.3	39.0	828.4	31.6	28.1	17.3	61.4%
FastV	43.4	41.5	51.7	42.6	42.5	32.5	884.6	37.8	33.2	20.7	64.1%
SparseVLM	58.6	48.3	51.9	57.3	46.1	67.9	1046.7	51.4	40.6	18.6	77.9%
PruMerge+	54.9	51.1	52.8	68.5	50.6	70.9	940.8	56.8	47.0	21.4	83.0%
VisionZip	67.1	51.8	52.9	68.8	53.1	68.7	1247.4	57.7	50.3	25.5	89.0%
VisPruner	67.7	52.2	53.0	69.2	53.9	72.7	1271.0	58.4	52.7	28.8	91.5%
FlashVLM	**69.3** (+1.6)	**52.8** (+0.6)	**53.2** (+0.2)	**69.4** (+0.2)	**54.6** (+0.7)	**74.5** (+1.8)	**1286.5** (+15.5)	**59.3** (+0.9)	**53.4** (+0.7)	**29.6** (+0.8)	**92.8%** (+1.3)

Table 1: Under different token budgets (128/64/32), FlashVLM consistently achieves the best or near-best performance across 12 benchmarks, maintaining high accuracy even under aggressive token reduction. This highlights its superior robustness and efficiency compared with prior pruning and compression baselines.

ActivityNet-QA [47]. For all benchmarks, we strictly follow the official evaluation protocols and metrics. Additional details are provided in the supplementary material.

Model Architecture. To validate the generality of FlashVLM, we apply it to a diverse set of visual-language model architectures. This includes the LLaVA family (LLaVA-1.5 [25] [27] for image and Video-LLaVA [24] [61] for video understanding), demonstrating its cross-modal applicability. Moreover, we extend FlashVLM to other mainstream VLMs such as Qwen-VL [1], InternVL [6], and CogVLM [41], with detailed results reported in Appendix. All models follow their respective original inference configurations to ensure fair comparison.

Baselines. For a comprehensive comparison, we include visual token pruning baselines from two major paradigms. The first category consists of *text-irrelevant* methods, including ToMe [4], LLaVA-PruMerge [37], VisionZip [44], and VisPruner [59]. The second category includes *text-relevant* pruning methods, represented by FastV [5] and SparseVLM [60]. This selection ensures broad coverage across pruning paradigms and pruning stages within VLM pipelines. For fair comparison, performance numbers for all baseline models are directly taken from the results reported in the VisPruner paper. Performance of FlashVLM is reported as the average over five independent runs to ensure statistical stability.

4.1 Different Visual Token Configurations

We first deploy and evaluate FlashVLM on the widely adopted LLaVA-1.5-7B model, followed by comprehensive comparisons against existing pruning methods. Table 1 reports the performance of different pruning strategies under three visual token budgets (128, 64, and 32 tokens).

The results clearly demonstrate that FlashVLM achieves state-of-the-art performance across all evaluated pruning configurations. Under moderate pruning

Method	TGIF-QA		MSVD-QA		MSRVTT-QA		Average	
	Acc (%)	Score	Acc (%)	Score	Acc (%)	Score	Acc (%)	Score
Upper Bound	**19.8**	**2.53**	**70.5**	**3.93**	**57.5**	**3.50**	**49.3**	**3.32**
Keep 455 Tokens (Prune 77.8%)								
FastV	19.2	2.50	69.1	3.91	54.4	3.42	47.6	3.28
VisPruner	18.4	2.49	70.2	3.95	56.7	3.50	48.4	3.31
FlashVLM	**18.9**	**2.52**	**70.3**	**3.97**	**57.0**	**3.51**	**48.7**	**3.33**
Keep 227 Tokens (Prune 88.9%)								
FastV	14.3	2.42	68.9	3.90	53.0	3.40	45.4	3.24
VisPruner	15.9	2.41	69.3	3.92	55.6	3.45	46.9	3.26
FlashVLM	**16.0**	**2.46**	**69.6**	**3.94**	**55.8**	**3.47**	**47.1**	**3.29**
Keep 114 Tokens (Prune 94.4%)								
FastV	10.6	2.29	64.1	3.78	52.4	3.39	42.4	3.15
VisPruner	14.1	2.35	65.4	3.79	54.1	3.41	44.5	3.18
FlashVLM	**14.3**	**2.37**	**65.6**	**3.80**	**54.3**	**3.44**	**44.7**	**3.20**

Table 2: The consistently superior or competitive accuracy/scores of our FlashVLM under all compression levels (455/227/114 tokens), maintaining strong performance even under aggressive pruning, compared with FastV and VisPruner.

(128 tokens retained, corresponding to 77.8% pruning), FlashVLM reaches an average accuracy of 100.60%, surpassing even the reported performance of the unpruned (upper-bound) model. This seemingly counterintuitive result highlights the effectiveness of our token selection mechanism: FlashVLM not only performs lossless compression but can also bring slight performance gains by filtering redundant or noisy visual tokens. Moreover, FlashVLM exhibits strong robustness under aggressive pruning. With 88.9% and 94.4% compression (64 and 32 tokens retained), FlashVLM maintains the highest accuracy scores of 97.90% and 92.80%, respectively. Notably, under the extreme 32-token configuration, FlashVLM significantly outperforms the next-best method VisPruner (91.50%), while methods such as ToMe suffer substantial performance degradation (61.40%). Finally, the superiority of FlashVLM is consistent and general. As indicated by the red highlights in Table 1, FlashVLM achieves the best results across all 10 sub-benchmarks under each of the three token retention settings.

4.2 FlashVLM in Video Understanding Scenarios

Video understanding represents another typical setting characterized by extremely high visual redundancy. To further validate the generalizability and robustness of FlashVLM, we extend and deploy it on the Video-LLaVA model. We follow similar experimental configurations and conduct evaluations on three widely used VideoQA benchmarks (TGIF-QA, MSVD-QA, and MSRVTT-QA), using ChatGPT-Assistant as the unified evaluator. In this setup, Video-LLaVA processes 8 video frames at 224 resolution, resulting in a total of 2048 visual tokens. As shown in Table 2, FlashVLM consistently achieves strong performance across all three benchmarks. Under all pruning levels (77.8%, 88.9%, and 94.4%), FlashVLM attains the highest average accuracy scores of 48.7%, 47.1%,

Model	Attention Distance ↓	Score Map Entropy ↓	Token–Box IoU ↑
FastV	2.86	1.52	0.21
SpareVLM	2.43	1.36	0.27
VisPruner	1.63	0.91	0.35
FlashVLM	**1.08**	**0.76**	**0.46**

Table 3: Effectiveness comparison of visual token selection strategies on VQAv2 (500 annotated samples). Lower Attention Distance and Score Entropy indicate more accurate spatial alignment and less attention dispersion; higher Token–Box IoU indicates more precise semantic grounding.

and 44.7%, respectively, outperforming both FastV and VisPruner. Notably, under the moderate compression setting (retaining 455 tokens), FlashVLM achieves an accuracy of 48.7%, and its average score (3.33) slightly surpasses the reported score of the unpruned upper-bound model (3.32). This again demonstrates that FlashVLM's dynamic token selection mechanism can effectively filter temporal redundancy and noise in video sequences, enabling "beyond lossless" compression. Moreover, under the extreme pruning ratio of 94.4% (retaining only 114 tokens), FlashVLM still maintains the highest average accuracy of 44.7%, and achieves the best score across all three sub-task benchmarks. In contrast, FastV exhibits more severe performance degradation (42.4%). These results further substantiate the robustness and performance advantages of FlashVLM in highly compressed video scenarios.

4.3 Effectiveness Analyses

To systematically evaluate the advantage of FlashVLM's lightweight text-guided token selection over conventional vision-language attention strategies, we conduct a comparative study on 500 randomly sampled VQAv2 instances with manually annotated ground-truth region boxes. We compare four representative methods, i.e., FlashVLM, FastV, SpareVLM, and VisPruner, and measure three key metrics: **Attention Distance** (quantifying the spatial deviation of the final focus region), **Score Map Entropy** (assessing the concentration of the score distribution and the degree of redundancy suppression), and **Token-Box IoU** (evaluating the overlap between the selected visual tokens and the annotated region, i.e., semantic relevance). These metrics respectively reflect **spatial alignment**, **distribution sparsity**, and **semantic matching accuracy**, enabling a comprehensive assessment of cross-modal grounding quality and visual redundancy reduction across different token selection paradigms.

Table 3 reflects the differences in cross-modal focusing quality across different models. **(1) Attention Distance.** This metric captures the RoPE-induced "proximity bias". FastV and SpareVLM exhibit substantial spatial drift (2.86 / 2.43) due to the absence of corrective mechanisms, while VisPruner partially mitigates this effect (1.63), but its focusing remains primarily driven by visual saliency rather than semantic grounding. In contrast, **FlashVLM** performs

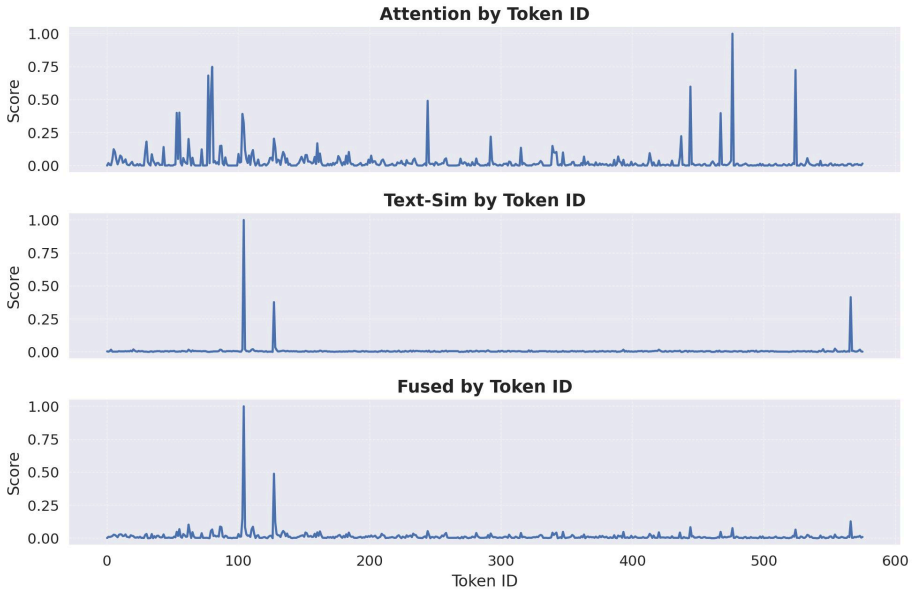

Fig. 4: The attention signal contains many noisy peaks, making key tokens hard to identify. In contrast, the text-similarity signal is sparse and semantically focused, effectively filtering out noisy attention activations. The fused score combines both, yielding more accurate and stable key-token localization.

lightweight text-guided semantic reweighting over visual tokens, steering the focus from positional proximity to semantic relevance, thus significantly reducing the spatial deviation (1.08). **(2) Score Map Entropy.** FlashVLM produces a low-entropy and semantically coherent concentration (0.76), effectively avoiding the "high-entropy, low-peak" dispersion problem. FastV and SpareVLM show higher entropy values (1.52 / 1.36), failing to form a stable salient region, while VisPruner may exhibit concentration in favorable cases but tends to revert to saliency-driven dispersion in complex scenes lacking a dominant subject (0.91).**(3) Token–Box IoU.** FlashVLM achieves the highest degree of semantic grounding (0.46), considerably outperforming FastV (0.21), SpareVLM (0.27), and VisPruner (0.35), demonstrating that its focused regions are not only concentrated but also accurately aligned with the task-relevant visual object.

As shown in Figure 4, FlashVLM's log-space fusion integrates the complementary strengths of the two unimodal signals. The *visual-attention* score (top) exhibits low-amplitude, noisy responses dominated by local textures without semantic grounding. The *text-similarity* score (middle) shows sparse, sharp peaks—strong semantic cues but poor spatial continuity—leading to unstable focus when used alone.

Method	Tokens Kept	FLOPs (T)	KV Cache (MB)	GPU Memory (GB)	Latency (ms)
LLaVA-NeXT-7B	2880	43.6	1440	17.0	313
FastV	640	13.5	380	16.9	148
VisPruner	640	11.5	360	14.8	117
FlashVLM	**640**	**11.3**	**356**	**14.6**	**115**
FastV	160	6.3	95	16.9	112
VisPruner	160	3.8	80	14.7	78
FlashVLM	**160**	**3.6**	**78**	**14.7**	**74**

Table 4: Runtime efficiency under identical token budgets. FlashVLM performs *single-shot* semantic token selection at the encoder–decoder interface, requiring no modification to internal transformer blocks. This design introduces negligible overhead and preserves full compatibility with FlashAttention, enabling the lowest FLOPs, KV-cache usage, and latency among all methods.

Method	VQAv2	POPE	MMB	MMVet	GQA
Keep 128 Tokens (Pruning Ratio 77.8%)					
Full Model	**76.4**	**85.1**	**63.2**	**34.1**	**58.9**
w/o Sextrinsic (no text guidance)	75.9	84.6	62.8	33.7	58.3
w/o Sintrinsic (no visual saliency)	75.3	84.1	62.6	33.5	58.1
w/o Diversity Reserve	75.9	84.8	63.0	33.8	58.1
w/o Log-Domain Fusion ($\rightarrow$ Linear)	76.1	84.9	63.1	33.9	58.6
w/o Text Sharpening	75.7	84.8	62.9	33.6	58.5
w/o Query Gating	75.8	84.4	62.7	33.7	58.3
Keep 64 Tokens (Pruning Ratio 88.9%)					
Full Model	**73.6**	**81.7**	**61.8**	**32.9**	**56.1**
w/o Sextrinsic (no text guidance)	72.7	80.6	61.3	32.4	55.4
w/o Sintrinsic (no visual saliency)	72.5	80.7	61.1	32.1	55.2
w/o Diversity Reserve	72.4	81.3	61.5	32.2	55.6
w/o Log-Domain Fusion ($\rightarrow$ Linear)	73.2	80.2	61.2	32.5	55.8
w/o Text Sharpening	73.0	80.9	61.2	32.3	55.6
w/o Query Gating	72.8	81.0	61.4	32.3	55.3

Table 5: Removing individual components under different token budgets (128 / 64) consistently degrades performance. The largest drops occur when removing text guidance (w/o Sextrinsic) or diversity preservation (w/o Diversity Reserve), highlighting their essential roles in maintaining robust cross-task performance.

4.4 Efficiency Evaluation

To assess the efficiency advantages of **FlashVLM**, we conduct a comprehensive comparison against two representative token-reduction baselines, i.e., **FastV** (text-aware but attention-dependent) and **VisPruner** (text-agnostic saliency-guided pruning), on the **LLaVA-NeXT-7B** model. These two methods represent the strongest prior approaches from the two dominant pruning paradigms and are widely used as primary efficiency references in recent VLM compression studies. Following the measurement protocol of VisPruner, we report four key runtime indicators: **FLOPs**, **KV-cache size**, **GPU memory consumption**, and **end-to-end CUDA latency**. All measurements are performed with batch size 1, FP16 inference, and a single NVIDIA A100 GPU, ensuring strict fairness. We evaluate two commonly used compression budgets, reducing the original 2880 visual tokens to **640** and **160** tokens.

As shown in Table 4, FlashVLM achieves the lowest FLOPs and latency under both token budgets. At 640 tokens, FlashVLM reduces computation to **11.3T FLOPs** and **115 ms** latency—slightly outperforming VisPruner and substantially faster than FastV. Under aggressive pruning (160 tokens), FlashVLM further improves to **3.6T FLOPs** and **74 ms** latency. These results highlight the core advantage of FlashVLM: Unlike FastV or SparseVLM, which repeatedly sparsify and recompute attention maps within the transformer layers, FlashVLM operates entirely outside the transformer stack, preserving dense attention and maintaining compatibility with FlashAttention. This yields both strong runtime efficiency and stable semantic relevance, enabling FlashVLM to scale naturally to future high-resolution or long-context VLMs.

4.5 Ablation Study

To assess the contribution of each component in FlashVLM, we evaluate six ablation settings across five benchmarks (VQAv2, POPE, MMB, MMVet, GQA). (1) **w/o Text Guidance ($S_{\text{extrinsic}}$)**: remove query-conditioned similarity to test localization without semantic grounding. (2) **w/o Visual Saliency ($S_{\text{intrinsic}}$)**: keep only text-based relevance to examine reliance on structural visual priors. (3) **w/o Diversity Reserve**: use only Top-K tokens to measure how aggressive redundancy removal affects contextual completeness. (4) **w/o Log-Fusion (Linear Fusion)**: replace log-domain fusion with linear summation to test the role of geometric combination. (5) **w/o Text Sharpening**: remove sparsity-enhancing sharpening to analyze its effect on emphasizing discriminative regions. (6) **w/o Query Gating**: disable query-norm gating to study the impact of weak or irrelevant textual signals on cross-modal alignment. As shown in Table 5, removing any component reduces performance under both 128- and 64-token settings. Under 64 tokens, removing *query-guided relevance* (w/o $S_{\text{extrinsic}}$) or *visual saliency* (w/o $S_{\text{intrinsic}}$) leads to drops of **0.7** and **0.8**, showing that semantic cues and structural priors jointly support reliable relevance estimation. Notably, the degradation without $S_{\text{intrinsic}}$ confirms that purely text-driven similarity is insufficient; inherent visual attention is necessary to capture the inter-object relations crucial for multi-hop reasoning.

5 Conclusion and Limitations

We introduced FlashVLM, a simple, query-aware token selection method operating at the encoder–decoder boundary. By fusing intrinsic saliency with an attention-free cross-modal similarity signal, it retains only relevant visual tokens while preserving essential context. Experiments across 14 benchmarks demonstrate beyond-lossless accuracy with over 75% token reduction. FlashVLM remains stable under extreme pruning and generalizes across diverse VLMs, offering an efficient solution for scalable multimodal reasoning.

Limitations. FlashVLM's performance depends on projected visual embedding quality, and highly fine-grained tasks may require larger token budgets. While

norm-based gating mitigates complex compositional queries, pure logical negation without explicit parsing remains challenging. Finally, our single-shot selection lacks multi-step refinement or temporal feedback, presenting avenues for future work to enhance adaptability.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by the Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grants 23hytd006 and 23hytd006-2, and in part by the Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), https://arxiv.org/abs/2308.12966
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? (2021), https://arxiv.org/abs/2102.05095
3. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster (2023), https://arxiv.org/abs/2210.09461
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your ViT but faster. In: International Conference on Learning Representations (2023)
5. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models (2024), https://arxiv.org/abs/2403.06764
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks (2024), https://arxiv.org/abs/2312.14238
7. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning (2023), https://arxiv.org/abs/2307.08691
8. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness (2022), https://arxiv.org/abs/2205.14135
9. Du, Y., Liu, Z., Li, J., Zhao, W.X.: A survey of vision-language pre-trained models (2022), https://arxiv.org/abs/2202.10936
10. Frankle, J., Carbin, M.: The lottery ticket hypothesis: Finding sparse, trainable neural networks (2019), https://arxiv.org/abs/1803.03635

11. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), https://arxiv.org/abs/2306.13394
12. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering (2017), https://arxiv.org/abs/1612.00837
13. Graesser, L., Evci, U., Elsen, E., Castro, P.S.: The state of sparse training in deep reinforcement learning (2022), https://arxiv.org/abs/2206.10369
14. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people (2018), https://arxiv.org/abs/1802.08218
15. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization (2019), https://arxiv.org/abs/1808.06670
16. Hooker, S., Courville, A., Clark, G., Dauphin, Y., Frome, A.: What do compressed deep neural networks forget? (2021), https://arxiv.org/abs/1911.05248
17. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering (2019), https://arxiv.org/abs/1902.09506
18. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering (2017), https://arxiv.org/abs/1704.04497
19. Jin, Y., Li, J., Liu, Y., Gu, T., Wu, K., Jiang, Z., He, M., Zhao, B., Tan, X., Gan, Z., Wang, Y., Wang, C., Ma, L.: Efficient multimodal large language models: A survey (2024), https://arxiv.org/abs/2405.10739
20. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Comput. Surv. **54**(10s) (Sep 2022). https://doi.org/10.1145/3505244, https://doi.org/10.1145/3505244
21. Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants (2023), https://arxiv.org/abs/2309.10020
22. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models (2023), https://arxiv.org/abs/2305.10355
23. Liang, J.C., Cui, Y., Wang, Q., Geng, T., Wang, W., Liu, D.: Clusterformer: Clustering as a universal visual learner (2023), https://arxiv.org/abs/2309.13196
24. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
28. Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., He, Z.: A survey of visual transformers. IEEE Transactions on Neural Networks and Learning Systems **35**(6), 7478–7498 (2024). https://doi.org/10.1109/TNNLS.2022.3227717

29. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), https://arxiv.org/abs/2307.06281
30. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS) (2022)
31. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019), https://arxiv.org/abs/1807.03748
32. OpenAI: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
33. Papa, L., Russo, P., Amerini, I., Zhou, L.: A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 7682–7700 (Dec 2024). https://doi.org/10.1109/tpami.2024.3392941, http://dx.doi.org/10.1109/TPAMI.2024.3392941
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
35. Ryoo, M.S., Piergiovanni, A., Arnab, A., Dehghani, M., Angelova, A.: Tokenlearner: What can 8 learned tokens do for images and videos? (2022), https://arxiv.org/abs/2106.11297
36. Sanh, V., Wolf, T., Rush, A.M.: Movement pruning: Adaptive sparsity by fine-tuning (2020), https://arxiv.org/abs/2005.07683
37. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. ICCV (2025)
38. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read (2019), https://arxiv.org/abs/1904.08920
39. Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient transformers: A survey. ACM Comput. Surv. **55**(6) (Dec 2022). https://doi.org/10.1145/3530811, https://doi.org/10.1145/3530811
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762
41. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models (2024), https://arxiv.org/abs/2311.03079
42. Wu, J., Gan, W., Chen, Z., Wan, S., Yu, P.S.: Multimodal large language models: A survey (2023), https://arxiv.org/abs/2311.13165
43. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM International Conference on Multimedia. p. 1645–1653. MM '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3123266.3123427, https://doi.org/10.1145/3123266.3123427
44. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. arXiv preprint arXiv:2412.04467 (2024)

45. Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J., Molchanov, P.: A-vit: Adaptive tokens for efficient vision transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10799–10808 (2022). https://doi.org/10.1109/CVPR52688.2022.01054
46. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities (2024), https://arxiv.org/abs/2308.02490
47. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering (2019), https://arxiv.org/abs/1906.02467
48. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5625–5644 (2024). https://doi.org/10.1109/TPAMI.2024.3369699
49. Zhang, J., Cai, K., Fan, Y., Liu, N., Wang, K.: MAT-agent: Adaptive multi-agent training optimization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=YDWRTYgR79
50. Zhang, J., Cai, K., Fan, Y., Wang, J., Wang, K.: CF-VLM: Counterfactual vision-language fine-tuning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=0qGtaRTsCo
51. Zhang, J., Cai, K., Fan, Y., Wang, J., Wang, K.: Cf-vlm:counterfactual vision-language fine-tuning (2025), https://arxiv.org/abs/2506.17267
52. Zhang, J., Cai, K., Zeng, Q., Liu, N., Fan, S., Chen, Z., Wang, K.: Failure-driven workflow refinement (2025), https://arxiv.org/abs/2510.10035
53. Zhang, J., Fan, Y., Cai, K., Sun, X., Wang, K.: Osc: Cognitive orchestration through dynamic knowledge alignment in multi-agent llm collaboration (2025), https://arxiv.org/abs/2509.04876
54. Zhang, J., Fan, Y., Cai, K., Yang, J., Yao, J., Wang, J., Qu, G., Chen, Z., Wang, K.: Why keep your doubts to yourself? trading visual uncertainties in multi-agent bandit systems (2026), https://arxiv.org/abs/2601.18735
55. Zhang, J., Fan, Y., Lin, W., Chen, R., Jiang, H., Chai, W., Wang, J., Wang, K.: GAM-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=EKJhU5ioSo
56. Zhang, J., Fan, Y., Wen, Z., Wang, J., Wang, K.: Tri-MARF: A tri-modal multi-agent responsive framework for comprehensive 3d object annotation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=YGIbwfNWot
57. Zhang, J., Guo, X., Cai, K., Lv, Q., Fan, Y., Chai, W., Wang, J., Wang, K.: Hybridtoken-vlm: Hybrid token compression for vision-language models (2025), https://arxiv.org/abs/2512.08240
58. Zhang, J., Huang, Z., Fan, Y., Liu, N., Li, M., Yang, Z., Yao, J., Wang, J., Wang, K.: KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=AKvy9a4jho
59. Zhang, Q., Cheng, A., Lu, M., Zhang, R., Zhuo, Z., Cao, J., Guo, S., She, Q., Zhang, S.: Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. arXiv preprint arXiv:2412.01818 (2025)
60. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., Zhang, S.: Sparsevlm: Visual token sparsification for efficient vision-language model inference (2025), https://arxiv.org/abs/2410.04417

61. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852 (2023)