SPOT: Sparsification with Attention Dynamics via Token Relevance in Vision Transformers

Supplementary Material

6. Token Relevance Module

6.1. Architecture

The token identification module first employs the following layers to generate z_{local} and z_{global} token representations:

$$\text{LayerNorm} \rightarrow \text{Linear}(d_{\text{remap}}, \tfrac{d_{\text{remap}}}{2}) \rightarrow \text{GELU},$$

where GELU refers to a Gaussian Error Linear Unit [22].

In addition to these base features, it incorporates attention-derived statistics from the attention maps in relevant heads and layers, $D^A_{l,h}$. The token and attention components are then concatenated into a unified representation.

Finally, the module applies

$$\text{Linear}(E, \tfrac{E}{2}) \rightarrow \text{GELU} \rightarrow \text{Linear}(\tfrac{E}{2}, \tfrac{E}{4}) \rightarrow \text{GELU} \rightarrow \text{Linear}(\tfrac{E}{4}, 2) \rightarrow \text{Softmax},$$

which transforms this concatenated vector into a probability distribution from which redundant tokens are identified.

The embedding dimension E depends on the inclusion of attention-derived features, as detailed in Sections 3.2 and 4.3. The design further allows incorporating statistics from earlier processing stages, which may include mean-based descriptors, variance-based descriptors, or both. Similarly, these statistics can be computed separately for each head or once across all heads, with the latter reducing the resulting feature dimensionality.

A key distinction of our module design lies in its capacity to process and utilize compact attention map representations for token relevance prediction. This approach enables more effective identification of contextually redundant tokens and delivers notable performance improvements compared to sparsification techniques that depend solely on token embedding or attention-derived information.

Given *SPOT*'s modular design and integration into existing blocks, the framework is in principle applicable to other Transformer architectures and tasks. For models with massive embedding dimensions, utilizing the lightweight variants outlined in Section 4.3 can effectively mitigate computational overhead by minimizing reliance on token-derived information. For dense prediction tasks like semantic segmentation, where preserving spatial resolution is critical, *SPOT*'s relevance scores can guide masking or merging operations rather than token removal, or serve as a guiding signal for decoding heads. Similarly, in multi-modal architectures such as Vision-Language Models (VLMs), the predictor input can be augmented with cross-attention statistics to capture the alignment between visual and textual tokens. Such adaptations can drastically lower the predictor's parameter count while preserving the rich contextual cues provided by attention dynamics or enhance performance in other domains. Other adaptation considerations include optimizing module placement, potentially adjusting the uniform spacing to align with architecture-specific depth and bottlenecks.

Model	Method	Accuracy (%)	GFLOPS
DeiT-T	Baseline [43]	44.8	1.3
	SPOT (ours)	45.2 (+0.4) ▲	0.8 (-0.5) ▼
DeiT-S	Baseline [43]	57.8	4.6
	SPOT (ours)	58.0 (+0.2) ▲	3.0 (-1.6) ▼
LV-ViT-T	Baseline [24]	50.8	2.9
	SPOT (ours)	49.6 (-1.2) ▼	2.0 (-0.9) ▼
LV-ViT-S	Baseline [24]	60.3	6.6
	SPOT (ours)	61.0 (+0.7) ▲	4.5 (-2.1) ▼

Table 5. Results of robustness evaluation on DeiT and LV-ViT models over ImageNet-C. *SPOT* not only achieves substantial computational savings but also improves accuracy in most cases compared to the baselines, demonstrating both efficiency and stability under perturbations.

6.2. Implementation Details

6.2.1. General Experimental Setup

The following implementation details apply to all experiments unless otherwise specified.

All models were optimized using AdamW [35]. The network was fine-tuned using eight NVIDIA GeForce RTX 4090 GPUs with a learning rate of 0.001, while the prediction modules used a rate of 0.01 to facilitate enhanced adaptation. Unless otherwise stated, all model instances underwent fine-tuning without a distillation token for 80 epochs with a batch size of 512. When integrating our approach with other methods, training durations were adjusted to align with their corresponding recommended fine-tuning protocols for fair comparison. For all experiments, we used a patch size of 16×16 pixels for processing. For the ImageNet [13] experiments, the input image resolution was 224×224 pixels.

While DeiT [43] preserves a structure comparable to standard ViTs, LV-ViT [24] introduces token labeling as an auxiliary training objective and incorporates convolutional operations into its design, requiring similarity between the refined tokens and those of a pre-trained model. Hence, the loss coefficients λ_1 and λ_2 in Equation 14 were set to 2 and

0.5, respectively, across all experiments, while λ_3 was set to 0 with DeiT models and 10 with LV-ViT models. The dimensionality of remapped token representations, d_{remap}, was set to the original dimensionality of the token embedding space, d_k, for each respective model.

In accordance with established practices in prior literature [28, 39, 48], token relevance prediction and subsequent sparsification were applied one layer after each quarter mark of the network depth, for the first three quarters, resulting in $K = 3$ sparsification stages. Specifically, for models with 12 layers, sparsification modules were plugged into layers 4, 7, and 10, whereas for the 16-layer LV-ViT-S, they were analogously positioned at layers 5, 9, and 13. Each sparsification iteration employs a separate module, as detailed in Section 3.2, with a token retention rate ρ_k.

6.2.2. Cross-Domain Ablation Setup

For the ablation study concerning additional benchmarks in Section 4.3, the following specific fine-tuning protocol was used. For each dataset, each model classification head was fine-tuned for 80 epochs with a learning rate of 0.025, while freezing the remaining components. This fine-tuning configuration is designed to encourage adaptation on smaller-scale datasets rather than subtle adjustments of pre-trained weights. No teacher supervision is applied in these downstream tasks, as the teacher model was trained solely on ImageNet; its classification head is specifically configured for ImageNet class labels and could bias adaptation dynamics. Although the DeiT models used in this setup were not trained exactly following the original recipe [43], both the baseline and *SPOT*-augmented models share identical architectural and training conditions. This ensures that performance differences reflect *SPOT*'s true effect.

7. Additional Experiments

In this section, we aim to rigorously isolate and characterize the intrinsic robustness behavior of the proposed *SPOT* framework, independent of architectural or training modifications introduced by other methods. Unlike the comparisons against state-of-the-art sparsification and token selection approaches presented in the main manuscript, the following evaluations focus on *SPOT*'s behavior relative to standard, well-established baselines. While prior works on token sparsification have primarily emphasized performance on ImageNet-1k, aspects such as robustness to perturbations and cross-domain transferability have received comparatively limited attention. In contrast, our evaluations explicitly address these aspects, enabling observed differences in robustness, accuracy, and computational efficiency to be directly attributed to *SPOT* itself, whose design facilitates the early detection and suppression of less informative or contextually redundant tokens. This setup provides a controlled and reproducible assessment that serves as a complementary and orthogonal analysis to the state-of-the-art comparisons presented in the main manuscript.

Dataset	Training Images	Test Images	Categories	Domain
ImageNet-1K [13]	1,281,167	50,000	1,000	Natural images
ImageNet-C [21]	–	950,000	1,000	Corrupted images
CIFAR-100 [30]	50,000	10,000	100	Pixelated objects
Food-101 [4]	75,750	25,250	101	Food items
DTD [10]	3,760	1,880	47	Textures
EuroSAT [20]	21,600	5,400	10	Satellite images

Table 6. Summary of datasets used for evaluation. The listed numbers reflect our usage of these datasets in this work. ImageNet-C was used solely for evaluation.

7.1. Robustness under Perturbations

To further assess the transferability and stability of *SPOT*, we evaluate its performance under challenging, corrupted visual conditions using the ImageNet-C benchmark [21]. This dataset consists of ImageNet-1k dataset images, which have been independently subjected to a diverse suite of 19 image degradation types, including noise, blur, weather, and digital distortions, applied at varying severity levels (1-5).

We adopt an intermediate severity level (3), which provides a balanced configuration that introduces meaningful degradations while preserving image recognizability. This setting enables a realistic assessment of robustness without saturating error rates or obscuring relative performance trends. All evaluations were conducted while employing the *SPOT* models fine-tuned on ImageNet-1K, as detailed in Section 4 and reported in Table 1, using the same computational budgets and without any additional adaptation or fine-tuning. Model accuracy is reported in Table 5 to quantify stability across perturbations relative to the baseline. Unless otherwise specified, all reported GFLOPS values denote total computational cost (equivalent to "Overall GFLOPS").

Across models, *SPOT* consistently maintains comparable accuracy to the baseline, and in most cases surpasses it, demonstrating strong robustness to input noise and distributional shifts despite operating at significantly lower computational budgets. This behavior can be attributed to *SPOT*'s estimation and utilization of most relevant tokens, which can better preserve the signal in the input despite perturbations, mitigating the amplification of corrupted or spurious tokens at a certain layer. Consequently, when exposed to input perturbations, the model's predictions are guided by a more robust and context-aware representation of token dynamics, which better preserves the integrity of salient feature representations and yields higher accuracy despite reduced computational capacity. These mechanisms underpin *SPOT*'s stable and broadly applicable computational benefits.

The results indicate that *SPOT* generalizes effectively beyond clean visual conditions, exhibiting resilience to both

Attention aggregation	Accuracy (%)	Predictor GFLOPS	Overall GFLOPS
Baseline [43]	79.8	–	4.6
Averaging across layers	73.6 (-6.2)	–	2.9 (-1.7)
SPOT (ours)	**79.2 (-0.6)**	<0.01	2.9 (-1.7)

Table 7. Comparison between learnable and non-learnable (heuristic) attention aggregation strategies on the DeiT-S backbone. '–' on the Predictor GFLOPS column denotes that no learnable predictor is used.

high-frequency corruptions (e.g., impulse noise) and low-frequency distortions (e.g., fog). This robustness underscores that *SPOT*'s attention dynamics regularization implicitly contributes to robust modeling of salient dependencies between tokens, reinforcing its suitability for deployment in real-world, noisy visual environments.

7.2. Information Aggregation and Granularity Analysis

We analyze the impact of different information aggregation strategies, comparing our learnable predictor against a non-learnable heuristic and evaluating the effect of information granularity within the learnable model.

First, to isolate the contribution of *SPOT*'s learnable attention-derived information-based importance prediction from simpler aggregation, we implemented a deterministic, non-learnable variant of *SPOT*, comparing both under identical conditions on the DeiT-S backbone and without token-derived information.

In this heuristic configuration, the learnable predictor was replaced by a fixed, attention-averaging mechanism, thereby eliminating contributions from learned weighting of attention-derived features. At each sparsification stage, token importance was computed as the mean class-token attention score ($A_{l,h}^{cls,out}$), averaged across all attention heads and all preceding layers ($l = 1, \ldots, k$). This scalar value captures the cumulative amount of attention a token has received from the class token throughout the network up to stage k, integrating attention information over multiple layers and heads. The ρ_k tokens exhibiting the highest mean scores were deterministically retained, while the remaining tokens were pruned, in accordance with attention-based token redundancy reduction principles, as detailed in Section 2. To ensure a fair comparison, the DeiT-S backbone was fine-tuned following the same optimization setup described in Section 4.1 and Appendix 6.2.

As summarized in Table 7, the described heuristic-based approach achieved an accuracy of 73.6% with a computational cost of 2.9 GFLOPS when evaluated on the ImageNet-1K dataset. This reflects a notable drop in accuracy relative to the baseline and to *SPOT* ablation variant that retained the learnable predictor while similarly excluding token-derived information, as detailed in the reduced token-derived information ablation in Section 4.3 for $d_{\text{remap}} = 0$. The mentioned ablation attained an accuracy of 79.2%, as illustrated in Figure 5. Importantly, this degradation in accuracy does not correspond to a noticeable computational benefit, since both configurations exhibit an equivalent computational cost of 2.9 GFLOPS (as reported in Section 4.3), given that the processing of attention-derived features within the proposed predictor incurs only negligible computational overhead.

This finding demonstrates that *SPOT*'s efficacy cannot be attributed solely to multi-layer attention aggregation, but rather to its learnable modeling of attention dynamics. This ablation confirms that the *SPOT* predictor captures complementary, context-aware information not fully perceptible to fixed, non-learnable aggregation heuristics, thereby reinforcing the validity of our proposed learnable sparsification framework.

We note that this variant is inspired by earlier work in token pruning for vision transformers that uses class-token attention to score and prune tokens in a single-layer fashion, e.g., EViT [32].

Second, we explore the efficacy of incorporating more granular information and modules versus aggregated ones. The results for granularity experiments are summarized in Table 8.

The default *SPOT* configuration feeds per-head attention values to its predictor. In a multi-head attention layer with H heads, each head learns different types of token relationships. By providing all H sets of attention-derived features, the predictor has access to this rich, diverse information. To this end, aiming to check how providing averaged attention values across the multi-head attention H heads, $D_l = \frac{1}{H} \cdot \sum_{h=1}^{H} D_{l,h}$, performs compared to per-head attention values.

Finally, we investigate the importance of an iteration-specific learnable token relevance prediction module. The default *SPOT* configuration includes separate predictor modules for each stage. For instance, the predictor at the first sparsification stage is trained specifically to find unimportant tokens based on information up to its corresponding layer, and does not share weights with other learnable modules. This experiment employs one common token detection module. This single, shared predictor module with one set of weights is being trained and applied repeatedly at all K sparsification stages.

7.3. Qualitative Results

Visualizations of the token identification process applied to samples from the ImageNet-1K dataset for the DeiT-T model are presented in Figure 6. The integration of *SPOT* with the DeiT-T architecture demonstrably preserves class-discriminative tokens despite its smaller capacity compared to the DeiT-S model, whose results are detailed in Figure 3.

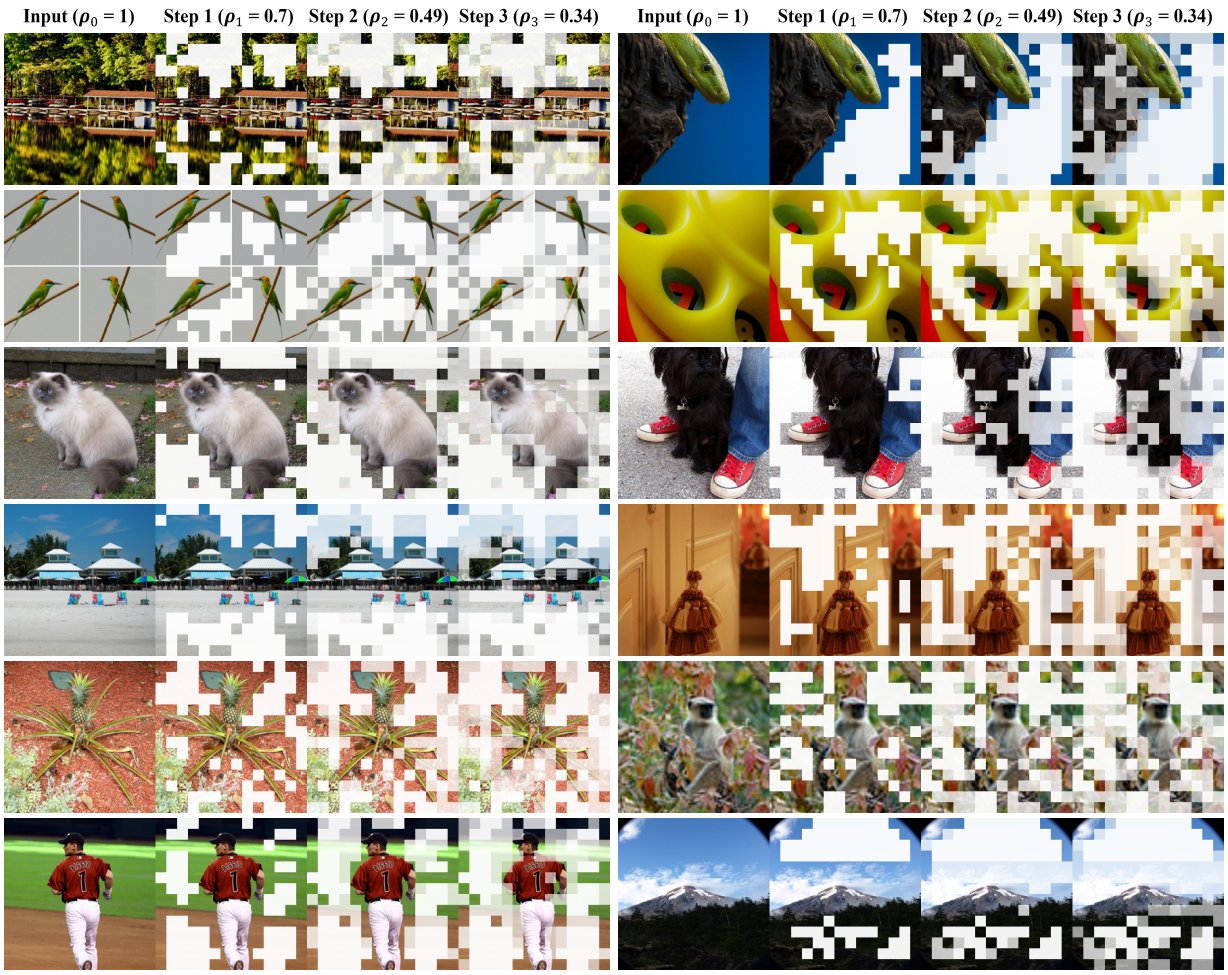

Figure 6. Visualizations of the gradual redundant token detection performed by our proposed approach on DeiT-T on samples from ImageNet-1K validation set. Increasingly transparent masking shades indicate later detection. Tokens identified as more informative, and thereby retained, are well aligned with semantic image objects and visual features, pointing to *SPOT*'s interpretability.

Predictor modification	Accuracy (%)	Predictor GFLOPS	Overall GFLOPS
Head-averaged attention	79.5 (-0.3)	0.106	3.0 (-1.6)
Iterations-shared module	79.4 (-0.4)	0.129	3.0 (-1.6)

Table 8. Granularity analysis results of *SPOT* on the DeiT-S backbone contrasting two module modifications: using head-averaged attention values versus per-head inputs, and using a single, shared predictor module versus stage-specific modules.

This observation further underscores the interpretability and generalization capabilities of the proposed methodology.

7.4. Throughput Analysis

Furthermore, we present an empirical throughput analysis for the various *SPOT* variants, which are characterized by the quantity and nature of the information received as input. Consistent with the variants examined in Section 4.3, Figure 7 provides a comprehensive overview of the empirical throughput performance of *SPOT* when integrated with the DeiT-S architecture with different token retention rates, denoted as ρ. All measurements were obtained under the experimental configuration outlined in Section 4.1. It is important to note that the throughput of *SPOT* is largely dictated by the underlying base architecture or method with which it is integrated, since *SPOT* does not modify the architecture or the computational pipeline. Accordingly, we focus our throughput analysis on comparing different *SPOT* variants within the same architecture (DeiT-S), thereby isolating the impact of its redundant-token prediction mechanisms.

As illustrated in Figure 7, *SPOT*'s versatility allows for a range of throughput levels that can be calibrated to meet different computational constraints and performance objectives. Such objectives are characterized by the token retention rate, ρ, and the predictor's input composition, as empirically validated in Section 4. By selectively incorporat-

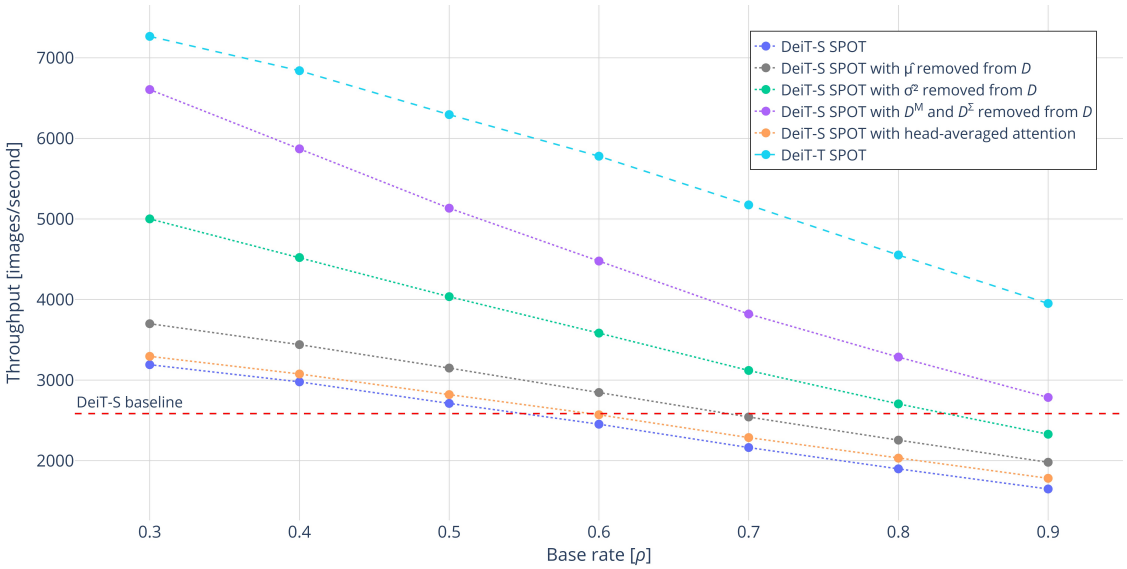

Figure 7. Comparative throughput analysis of *SPOT*. We evaluate the DeiT-S architecture integrated with all variants of the proposed *SPOT* redundant token prediction module, alongside the DeiT-T architecture utilizing the standard *SPOT* configuration. The evaluation was performed on a single NVIDIA GeForce RTX 4090 GPU, employing a batch size of 512. Each variant is characterized by the distinct combination of token and attention-derived information it incorporates for prediction. This design ensures that *SPOT* is adaptable, allowing users to calibrate its parameters to meet specific computational constraints and performance objectives.

ing or removing token and attention-derived information, *SPOT* enables a trade-off between throughput and accuracy across varying retention rates. For instance, one *SPOT* variant achieves a throughput of 3,880 images per second, over x1.5 higher compared to the baseline, while operating at 3.0 GFLOPS and maintaining an accuracy of 79.5%. This demonstrates the flexibility of *SPOT* in adapting to diverse operational settings while preserving strong performance.

8. Theoretical Perspective

The efficacy of *SPOT* can be interpreted through the lens of statistical reduction of sample variance and structured aggregation of attention signals.

Guo et al. [17] show that in Vision Transformers (ViTs), a single attention layer provides an inherently noisy estimation of a token's underlying importance, denoted as $R[i]$, and that ViTs often suffer from unstable distributions of high-attention tokens, where even mild perturbations can significantly alter which tokens appear salient. The attention scores $A_l[i]$ assigned to token i at layer l within a specific model head, and by other tokens, can be modeled as

$$A_l[i] = R[i] + \epsilon_l[i],$$

where $\epsilon_l[i]$ represents fluctuations arising from spurious spikes or localized instabilities.

When token relevance is inferred from one layer, as common in prior literature, the variance of the estimator is $\mathrm{Var}(A_l[i]) = \sigma_\epsilon^2$. Conversely, *SPOT* aggregates across L layers to formulate a composite estimator:

$$\hat{R}[i] = \frac{1}{L}\sum_{l=1}^{L} A_l[i].$$

This aggregation strategy substantially reduces the estimator's variance, which becomes:

$$\mathrm{Var}(\hat{R}[i]) = \mathrm{Var}\left(\frac{1}{L}\sum_{l=1}^{L} A_l[i]\right) \approx \frac{1}{L^2} L\sigma_\epsilon^2 = \frac{\sigma_\epsilon^2}{L},$$

under the assumption of weakly correlated noise across different layers. Thus, inter-layer aggregation functions as a form of ensemble averaging [18], yielding a more robust and lower-variance estimate of token saliency. Importantly, this procedure is performed independently for each attention head, after which the resulting descriptors are further aggregated across all heads, ensuring that complementary interaction patterns captured by distinct heads are synergistically leveraged.

As detailed in Section 3, *SPOT* compresses attention maps into compact descriptors that capture both outgoing and incoming dynamics. These descriptors are constructed from several components, including the first row and column of the attention map, which quantify the distribution of outgoing influence from the global aggregator, and the contribution of each patch token to the class token, respectively. To further summarize interaction patterns, the framework computes row-wise and column-wise statistics,

namely the mean and variance (formulated in equations 4, 5, 6, 7). The row-wise moments encode how a given token distributes its attention across all other tokens, while the column-wise moments describe how a token is attended to by others, thereby capturing its global significance.

The utilization of these moment-based descriptors yields two principal advantages. First, by summarizing attention distributions with statistical moments, the influence of spurious spikes and outliers within the raw attention maps is suppressed. Second, the explicit differentiation between row-wise and column-wise patterns provides crucial directional context, enabling *SPOT* to distinguish between tokens that broadly influence others and those that consistently attract attention.

Beyond single-layer descriptors, *SPOT* also models the trajectories of these statistical moments across layers. The mean of these statistics across layers, $D^{M}_{l,h}$, serves as an indicator of a token's influence and cumulative saliency throughout the model, highlighting which tokens consistently exert meaningful impact on information flow. Simultaneously, the variance of these values across layers, $D^{\Sigma}_{l,h}$, measures the stability of a token's role, indicating whether its importance is consistent or fluctuates with local context. This characterization of inter-layer attention evolution helps prevent the premature pruning of tokens that may become relevant in deeper layers of the network.

By combining the raw first row and column vectors, summary statistics from row and column-wise moments, aggregated cross-layer moments, and token embeddings from the current layer, *SPOT* constructs a compact yet information-rich representation of tokens and their attention dynamics. This composite representation comprehensively encodes both local and global interactions, tracking short-term fluctuations and long-term stability, while remaining resilient to transient attention noise. Thus, *SPOT* offers a holistic and theoretically-supported framework for more interpretable token sparsification.