IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 June 2025, accepted 4 July 2025, date of publication 7 July 2025, date of current version 15 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3586739

RESEARCH ARTICLE

Geometry Meets Attention: Interpretable Transformers via SVD Inspiration

VASILEIOS ARAMPATZAKIS[1,2], GEORGE PAVLIDIS[2], (Senior Member, IEEE), NIKOLAOS MITIANOUDIS[1], (Senior Member, IEEE), AND NIKOS PAPAMARKOS[1], (Senior Member, IEEE)

[1]Democritus University of Thrace, University Campus at Kimmeria, 67100 Xanthi, Greece
[2]Athena Research Center, University Campus at Kimmeria, 67100 Xanthi, Greece

Corresponding author: Vasileios Arampatzakis (vaarampa@ee.duth.gr)

This work has been partially supported by the ARGUS EU project (Grant Agreement No. 101132308), funded by the European Union.

ABSTRACT Self-attention is a cornerstone of modern deep learning, yet its dense dot-product formulation offers limited interpretability and lacks explicit structural constraints. We propose *SVD-inspired Attention* (SVDA), a novel self-attention mechanism that introduces normalized query/key projections and a learnable diagonal spectral modulation, drawing direct motivation from the structure of Singular Value Decomposition (SVD). This formulation separates directional alignment from spectral emphasis, offering a geometrically grounded and interpretable variant of attention. We formalize SVDA within a standard multi-head Transformer architecture and introduce a suite of structure-aware indicators—such as spectral entropy, effective rank, and selectivity—that quantify interpretability and sparsity in attention dynamics. Our analysis highlights SVDA's capacity for structured, energy-aware attention without compromising architectural compatibility or expressiveness. This work provides a theoretical foundation and diagnostic framework for structured attention models aimed at interpretability, compression, and semantic transparency.

INDEX TERMS Transformers, self-attention, singular value decomposition (SVD), interpretable AI, spectral modulation, attention mechanism, deep learning theory, structured representations, orthogonality regularization, explainable machine learning.

I. INTRODUCTION

Self-attention has become a foundational operation in modern deep learning architectures, particularly in Transformer models. Despite their remarkable performance, standard attention mechanisms face two core challenges: high computational cost and limited interpretability. The dense dot-product attention formulation distributes relevance broadly and opaquely, making it difficult to understand how decisions emerge from model internals.

Motivation. The lack of interpretability is especially problematic in sensitive domains such as healthcare, scientific discovery, and cultural heritage, where understanding the rationale behind model predictions is critical. Interpretability tools such as attention visualizations and probing often fall short in capturing structured patterns or explaining model behavior. To address this, recent works have explored constraints and decompositions that impose meaningful structure on the attention mechanism.

Our Approach. In this paper, we propose a novel modification, *SVD-inspired Attention (SVDA)*, that introduces soft-orthonormal query/key projections and a learned diagonal spectrum inspired by Singular Value Decomposition (SVD). SVD is a powerful matrix decomposition that expresses a matrix as $X = U\Sigma V^\top$, where U and V are orthonormal and Σ is diagonal. This factorization separates directional information (geometry) from magnitude (importance), which aligns naturally with the semantics of attention.

Spectral Modulation and Interpretability. SVDA adapts the SVD principle to the attention setting by applying normalization and spectral weighting in the query-key interaction. The diagonal matrix Σ becomes a learned parameter,

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

modulating the influence of different latent dimensions. This separation provides geometric constraints and spectral priors that yield better focus, regularization, and interpretability. Moreover, SVDA enables new structural indicators such as spectral entropy, sparsity, and effective rank to analyze the attention mechanism in a principled way.

Scope Clarification. Unlike many contemporary works that prioritize empirical performance, this paper is primarily intended as a principled theoretical foundation. We aim to formalize SVDA's core design rationale, mathematical structure, and interpretability potential, without focusing on extensive benchmark comparisons or application-specific fine-tuning. Theoretical clarity and structural diagnostics are emphasized over performance gains, which are to be addressed separately in future empirical work.

Contributions. The main contributions of this work include:

- A new SVD-inspired attention mechanism combining normalized directional similarity and learned spectral modulation.
- A fully compatible multi-head implementation that integrates with standard Transformer and ViT pipelines.
- A set of novel interpretability-oriented indicators (e.g., spectral entropy, effective rank, spectral sparsity, selectivity index, perturbation sensitivity) that quantitatively assess structure and robustness in attention.
- Extensive analytical and demonstrative empirical insights into the behavior of spectral weights and their connection to attention structure and focus.

II. RELATED WORK

The Transformer architecture introduced by Vaswani et al. [1] established self-attention as the backbone of modern deep models in NLP and beyond. The attention mechanism computes pairwise interactions using scaled dot products between projected queries and keys. This is the **Standard Attention in Transformers (baseline)** and despite its empirical power, this formulation offers little insight into latent interactions and incurs quadratic complexity in sequence length.

Several studies have proposed alternatives that reduce this burden, toward **Efficient and Low-Rank Attention**. Linformer [2] approximates attention matrices via low-rank projections on keys and values. Performer [3] uses kernel-based random feature maps to approximate the softmax kernel. The Nyströmformer [4] accelerates self-attention using Nyström matrix approximation. Linear Transformers [5] reformulate attention using kernel-based linearization, enabling autoregressive decoding with reduced complexity.

Structured attention with spectral or geometric biases defines a growing line of work under **Spectral and SVD-Inspired Attention**. Singularformer [6] models attention as a decomposed low-rank factorization, while Primal-Attention [7] applies kernel SVD in primal space. Cosformer [8] replaces the softmax kernel with cosine similarity for sharper gradients and alignment. SpecFormer [9] incorporates spectral decomposition to stabilize and diversify attention heads. These methods are included for completeness as they represent structurally relevant efforts that also explore spectral or factorized attention, albeit from different methodological standpoints, helping position the proposed method within the broader landscape.

In the line of research towards **Angular and Normalized Attention** models, NormFormer [10] improves training stability by inserting LayerNorm in attention and feedforward layers. Unlike these approaches, our model enforces normalized query/key directions and introduces a learned diagonal spectrum, drawing a formal analogy to singular value decomposition.

Beyond computational goals, there is increasing interest in **Interpretability-Oriented Modifications**. Chefer et al. [11] go beyond raw attention scores, proposing saliency-based attribution for vision Transformers. Bastings and Filippova [12] challenge attention as explanation, pointing instead to sparse and disentangled internal representations.

Our work is most closely related to *Angular and Normalized Attention* models. Our method contributes a structured spectral lens on interpretability, enabling compression-awareness and directional probing through indicators like entropy and spectral sparsity.

TABLE 1[1] summarizes key distinctions between SVDA and selected prior methods in structure, spectrum, and complexity.

In parallel to spectral or interpretability-driven mechanisms, **sparse attention architectures** have been proposed to improve scalability. Notable examples include fixed attention patterns in Sparse Transformers [13], locality-sensitive hashing in Reformer [14], and randomized sparse connectivity in BigBird [15]. Routing Transformers [16] group tokens dynamically, while entmax-based methods [17] impose soft sparsity via generalized softmax functions. Unlike these approaches, SVDA does not rely on sparsity or token masking to regulate attention. Instead, it introduces orthonormal query-key projections and diagonal spectral scaling, drawing from matrix factorization theory. While sparse models target computational efficiency, SVDA prioritizes interpretability, structural decomposability, and controlled expressiveness through spectral modulation.

III. SVD-INSPIRED ATTENTION – SVDA

A. DEFINITION OF SVDA

The Singular Value Decomposition (SVD) of a matrix $M \in \mathbb{R}^{n\times d}$ expresses it as the product:

$$M = U\Sigma V^{\top} \tag{1}$$

where $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{d\times r}$ are orthonormal matrices spanning the left and right singular subspaces, respectively, and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix of singular values,

[1] It should be noted that in SVDA, $d = d_k$ denotes projection dimension. Additional costs $\mathcal{O}(d_k^2)$ (training only) and $\mathcal{O}(nd_k)$ (modulation) are asymptotically minor when $n \gg d_k$. as explained in subsequent sections.

TABLE 1. Comparison of SVDA with verified related attention mechanisms.

Method	Attention Approximation	Low-Rank	Spectral	Complexity	Notes
SVDA (ours)	$Q\Sigma K^\top$, norm(Q), norm(K)	Optional	✓	$\mathcal{O}(n^2 d)$	Orthonormal projections and learned Σ enable structured and interpretable attention.
Linformer [2]	$Q(EK)^\top$	✓	✗	$\mathcal{O}(nd)$	Uses fixed-rank projections on keys/values for efficiency.
Performer [3]	$\phi(Q)\phi(K)^\top$	✗	✗	$\mathcal{O}(nd)$	Kernel-based attention using random features.
Nyströmformer [4]	Nyström sampling	✓	✗	$\mathcal{O}(nmd)$	Approximates full attention using landmark-based estimation.
Linear Transf. [5]	Kernel reformulation	✗	✗	$\mathcal{O}(nd)$	Avoids softmax via kernel-based linearization.
Singularformer [6]	MLP-learned $QK^\top$ factorization	✓	implicit	$\mathcal{O}(n^2 d)$	Learns low-rank factors without spectral separation.
Primal-Attention [7]	Asymmetric kernel SVD	✓	✗	$\mathcal{O}(nd)$	Applies SVD in feature space; asymmetry preserved.
Cosformer [8]	Cosine kernel	✗	implicit	$\mathcal{O}(nd)$	Replaces dot-product with cosine similarity for sharper focus.
NormFormer [10]	LayerNorm-enhanced attention	✗	implicit	$\mathcal{O}(n^2 d)$	Stabilizes gradients with extra normalization.
SpecFormer [9]	Eigen-decomp. of A	✓	✓	$\mathcal{O}(n^2 d)$	Applies spectral decomposition to full attention matrix.
Chefer et al. [11]	Gradient-based saliency	✗	✗	-	Enhances interpretability beyond raw attention scores.
Bastings et al. [12]	Structural probing	✗	✗	-	Challenges attention weights as explanations.
Vaswani et al. [1]	$QK^\top/\sqrt{d_k}$	✗	✗	$\mathcal{O}(n^2 d)$	Canonical baseline for all Transformer variants.

reflecting the importance of each component. This decomposition separates geometric directionality (given by U and V) from magnitude (given by Σ) and underpins many applications in compression, signal processing, and learning.

Inspired by this structure, our proposed SVD-Attention (SVDA) mechanism integrates a learnable spectral modulation Σ into the attention operation. Specifically, we model attention weights through a decomposition-like form:

$$A \sim Q\Sigma K^\top \tag{2}$$

where Q and K are normalized query and key projections (analogous to orthonormal matrices in SVD), and Σ is a learned diagonal matrix that scales dimensions in a head-specific manner.[2] As will be discussed, tested and detailed in subsequent sections, *this design encourages structured attention that disentangles directional alignment from spectral emphasis, improving interpretability and control*.

Standard attention uses dot products of dense projections without constraint. We hypothesize that imposing orthogonality on queries/keys and modulating their interaction with learned singular values offers better semantic separation and control. Supposing the general case of multi-head attention, the following key dimensionalities are defined:

- d_k: the dimensionality of query and key vectors (per head)
- d_v: the dimensionality of value vectors (per head)
- h: the number of attention heads
- d_{model}: the model embedding dimension (i.e., the dimensionality of token representations after embedding and throughout the Transformer)

[2]Per-head, diagonal, and independently learned.

- we adopt $d_k = d_v = d_{\text{model}}/h$ as an architectural choice, motivated by: (a) simplicity, so that concatenated outputs preserve the model dimension; (b) parameter efficiency; and (c) implementation convenience.

Let $X \in \mathbb{R}^{n\times d_{\text{model}}}$ be the input sequence of n tokens, each represented with d_{model}-dimensional embeddings. In the proposed SVDA approach, we divide the model into h attention heads. The attention for each head $i = 1, \ldots, h$ is computed as:

$$\begin{aligned}
Q_i &= \text{normalize}(XW_i^Q) \in \mathbb{R}^{n\times d_k}\\
K_i &= \text{normalize}(XW_i^K) \in \mathbb{R}^{n\times d_k}\\
\Sigma_i &= \text{diag}(\sigma_1^{(i)}, \ldots, \sigma_{d_k}^{(i)}) \in \mathbb{R}^{d_k\times d_k}\\
V_i &= XW_i^V \in \mathbb{R}^{n\times d_v}\\
A_i &= \text{softmax}\left(\frac{(Q_i\Sigma_i)K_i^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{n\times n}\\
Z_i &= A_iV_i \in \mathbb{R}^{n\times d_v}
\end{aligned} \tag{3}$$

where $W_i^Q, W_i^K \in \mathbb{R}^{d_{\text{model}}\times d_k}$, and $W_i^V \in \mathbb{R}^{d_{\text{model}}\times d_v}$ are the projection matrices. The normalize operator applies row-wise ℓ_2-normalization

$$\text{normalize}(Y) = \left[\frac{y_i}{\|y_i\|_2}\right]_{i=1}^{n}, \quad \text{for any } Y \in \mathbb{R}^{n\times d_k} \tag{4}$$

ensuring that each token vector has unit norm. The final attention output is obtained by concatenating all head outputs

$$Z = \text{Concat}(Z_1, \ldots, Z_h)W^O \in \mathbb{R}^{n\times d_{\text{model}}} \tag{5}$$

where $\text{Concat}(Z_1, \ldots, Z_h) \in \mathbb{R}^{n\times(h\cdot d_v)}$, and $W^O \in \mathbb{R}^{(h\cdot d_v)\times d_{\text{model}}}$ is a shared output projection matrix that maps

the concatenated multi-head output back to the model dimension.[3]

To reinforce the SVD structure, we introduce a *soft orthogonality regularization* term on the query and key projection weights, indirectly imposing an **orthonormality constraint**. This is defined as:

$$\begin{aligned}\mathcal{L}^{Q}_{\text{ortho}} &= \|W^{Q^\top} W^{Q} - I\|_F^2 \\ \mathcal{L}^{K}_{\text{ortho}} &= \|W^{K^\top} W^{K} - I\|_F^2\end{aligned} \tag{6}$$

where $\|\cdot\|_F$ denotes the *Frobenius norm*, which for any matrix $Y \in \mathbb{R}^{d_k \times d_k}$ is defined as $\|Y\|_F = \sqrt{\sum_{i,j} |Y_{ij}|^2}$. In our case, Y corresponds to the deviation of a *Gram matrix*[4] from the identity, i.e., $W^{Q^\top} W^{Q} - I$ or $W^{K^\top} W^{K} - I$, each in $\mathbb{R}^{d_k \times d_k}$. The Frobenius norm penalizes all off-diagonal correlations and diagonal scale mismatches uniformly, encouraging the columns of W^Q and W^K to form nearly orthonormal bases. This choice supports the intended SVD-like structure by promoting global orthonormality across all projection directions. Additionally, the Frobenius norm is computationally efficient and fully differentiable, making it a practical regularizer during training. These regularization terms are added to the overall training objective with a weighting factor λ_{ortho}, encouraging the learned projections to span nearly orthonormal subspaces, in line with the SVD decomposition. The total training loss is thus given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{ortho}} \left(\mathcal{L}^{Q}_{\text{ortho}} + \mathcal{L}^{K}_{\text{ortho}} \right) \tag{7}$$

where $\mathcal{L}_{\text{task}}$ denotes the primary task-specific loss function, such as cross-entropy for classification or mean squared error for regression. This regularization encourages W^Q and W^K to have approximately orthonormal columns, by minimizing the deviation of their corresponding Gram matrices $W^{Q^\top} W^{Q}$ and $W^{K^\top} W^{K}$ from the identity.

B. LEARNING BEHAVIOR AND DIFFERENCES

SVDA introduces key differences to standard attention architectures as it (a) imposes geometric structure via normalized projections and soft orthogonality, (b) introduces learnable spectral modulation via diagonal Σ and (c) mimics SVD, $A \sim Q\Sigma K^\top$, separating direction (geometry) from magnitude (importance). Traditional attention mechanisms rely solely on dot products of learned projections, which may entangle content and scale. In contrast, SVDA disentangles geometric alignment (via normalized Q, K) from energy shaping (via Σ), and soft orthogonality reinforces interpretable attention subspaces.

[3]Given our assumption $d_v = d_{\text{model}}/h$, the concatenated dimension $h \cdot d_v$ equals d_{model}, so W^O becomes square.

[4]The *Gram matrix* captures all pairwise inner products between the columns of a matrix, encoding the geometry of their span, including lengths, angles, and linear dependencies. In SVDA, it is used to measure and softly enforce orthonormality of the learned query/key projection directions. When the Gram matrix equals the identity, the vectors are perfectly orthonormal, which is the structural prior you're regularizing toward.

In SVDA, the projection matrices W^Q, W^K, and W^V are still fully learned, as in standard attention. The key distinction lies in the treatment of the projected query and key vectors, and in the introduction of an additional diagonal parameter matrix Σ.

The introduction of the learnable diagonal matrix Σ, which modulates the energy spectrum of each query vector before computing attention defined in (3), introduces directional selectivity, allowing the model to emphasize or suppress specific latent dimensions in a structured and interpretable way, achieving a *learned spectral scaling*. All projection matrices and the spectral coefficients in Σ are learned jointly via backpropagation. Σ introduces only d_k new parameters per head and can optionally be regularized or sparsified. At the end, the learned Σ enables fine-grained, direction-specific scaling of attention energy, allowing the model to learn a spectrum of focus that can emphasize or suppress certain semantic directions, thus *offering spectral modulation*.

Normalization of Q and K ensures cosine-like similarity structure in attention computation, removing the influence of vector norms and isolating angular relationships. The normalization step fixes the length of vectors and effectively constrains the geometry of the attention computation to the hypersphere. Since the use of normalized query and key vectors ensures attention scores reflect angular (cosine-like) similarity, providing geometric meaning to the attention distribution, *SVDA offers geometric interpretability*.

Soft orthogonality of Q and K is imposed as described in (6), leading to a new loss function as shown in (7). The orthogonality regularization provides an additional structural bias toward semantic disentanglement.

C. INTERPRETABILITY AND TRADE-OFFS

While SVDA introduces only *a small number of additional parameters* via the diagonal matrix Σ, it provides substantial gains in structure, transparency, and semantic introspection.

For each attention head, SVDA adds only d_k parameters, as shown in (3), with $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d_k})$, where $\sigma_j \in \mathbb{R}$ being learned, which is minimal compared to the standard W^Q, W^K, especially in multi-head settings. By reusing projection dimensions and using a single learned spectrum, *SVDA offers parameter efficiency* as it introduces minimal additional overhead, while capturing rich interactions across tokens.

Unlike standard attention, which combines vector magnitudes and directions into a raw dot product, SVDA explicitly decouples these effects by applying the new workflow (3), where now Q, K reflect pure angular (cosine) similarity, whereas Σ modulates the spectral importance of each semantic direction. This separation enables interpretation of attention in terms of both direction and importance of latent features, achieving a *geometry and energy separation*.

In addition, the explicit separation of directions (via Q/K) and their weighting (via Σ) mirrors the structure of SVD, leading to a more disentangled representation of attention

dynamics, rendering *SVDA capable of offering structured attention*.

The diagonal matrix Σ acts as a *tunable spectrum* that modulates the similarity computation between queries and keys. This corresponds to a dimension-wise reweighting of inner products, in which for any query-key pair (Q_i, K_j), their similarity is computed as $\text{sim}(Q_i, K_j) = \sum_{l=1}^{d_k} \sigma_l \cdot q_{il} \cdot k_{jl}$. Each σ_l thus controls the relative importance of the l-th spectral component, *achieving learnable spectral control over the attention mechanism*.

As *SVDA offers the potential for spectral regularization*, the singular value spectrum can be regularized or constrained (e.g., via soft-thresholding or sparsity), making the approach compatible with model pruning, interpretability, or compression strategies. While the current formulation learns the spectral matrix Σ without explicit regularization, the framework allows the inclusion of L_1, L_2, or sparsity-inducing penalties to promote dimensional compression, low-rank adaptation, or structured interpretability. Exploring such constraints on Σ remains a promising direction for future work.

Unlike low-rank approximations, *SVDA offers full-rank expressiveness* (as Σ is diagonal and unconstrained, full rank can be preserved so long as no σ_i is zero), as it maintains full rank in attention computation, avoiding information loss while still enabling structured compression when desired (e.g., by thresholding Σ).

SVDA offers compatibility with standard architectures can be seamlessly integrated into Transformer encoders, multi-head blocks, or ViT-style models with no change to positional encoding or downstream components.

SVDA's structure brings some important *interpretability benefits*, as it allows inspection of Σ at each head and layer, enabling answers to questions like

- Which dimensions are emphasized in attention?
- Does the model use a few dominant directions (low-rank behavior)?
- How does the spectrum evolve during training?

While the spectrum Σ enables per-head interpretability via direct inspection of learned directional importance, SVDA's structural design also supports broader introspection. By enforcing row-wise normalization and applying soft orthogonality regularization on Q and K, the model ensures that attention interactions arise from well-conditioned, geometrically interpretable subspaces. This structured formulation allows subsequent derivation of interpretable indicators, grounded in spectral concentration, directional coherence, and robustness, further explored in the next sections. Thus, SVDA transforms attention from an opaque mechanism into an analyzable and diagnostic object.

TABLE 2 provides a conceptual comparison between standard attention and SVDA, underscoring SVDA's interpretability benefits. This structure makes SVDA well-suited for tasks demanding transparency, control, and semantic interpretability, including natural language, structured vision, and biomedical domains. Notably, SVDA delivers these interpretability gains without compromising full-rank expressivity.

D. ARCHITECTURAL INTEGRATION

SVDA integrates seemlessly into existing Transformer architectures. It preserves layer normalization, positional encoding, and residual pathways, and requires no modifications to data preprocessing or surrounding model components.

TABLE 2. Conceptual Standard vs SVDA attention.

Aspect	Standard	SVDA
Query/Key Vectors	Raw projections	Normalized unit vectors
Similarity Basis	Dot product	Cosine + spectral weighting
Spectral Control	None	Learned $\Sigma \in \mathbb{R}^{d_k \times d_k}$
Interpretability	Low	High (dimension-wise importance)
Rank Structure	Unstructured	Structured spectrum
Additional Params	None	d_k per head

IV. QUANTIFYING STRUCTURE AND INTERPRETABILITY

As SVDA's added value lies in its structured geometry and potential for interpretability, we introduce a set of *indicators* that help reveal the internal behavior and latent structure of attention. These indicators are designed to make attention dynamics transparent and analyzable, offering structured insights into the model's internal mechanisms. Importantly, the indicators are *descriptive*, *not prescriptive*; they are not meant to enforce specific behaviors or serve as optimization targets; their role is diagnostic, aiming to highlight emergent properties of attention in transformer models without imposing normative design goals.

A. SPECTRAL ENTROPY

Given the learned diagonal spectrum $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d_k})$, we define the normalized energy distribution as

$$p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2} \tag{8}$$

The *spectral entropy* is then defined as

$$\mathcal{H}_\Sigma = -\sum_{i=1}^{d_k} p_i \log(p_i) \tag{9}$$

Lower entropy indicates concentration on fewer dimensions (selective modes), while higher entropy implies diffuse attention. This definition stems straight from the formal definition of Entropy, well-established in information theory.

B. EFFECTIVE RANK

Using the same normalized energy distribution p_i, we define a soft measure of rank as

$$\text{rank}_{\text{eff}}(\Sigma) = e^{\mathcal{H}_\Sigma} \tag{10}$$

This reflects how many latent directions contribute meaningfully to the attention interaction. It is bounded between 1

(when all energy is concentrated in one component) and d_k (for a uniform spectrum), and acts as a continuous relaxation of matrix rank, well-suited for gradient-based optimization and introspection. This definition is based on [18], [19].

C. ANGULAR ALIGNMENT

Because Q, K are normalized in SVDA, attention weights reflect cosine similarity. We may analyze

$$\cos\theta_{ij} = Q_i \cdot K_j \tag{11}$$

to quantify how closely queries align with keys. Aggregate statistics (mean and variance of $\cos\theta$) reveal how focused or diffuse the attention is geometrically. This measure builds on the classical notion of cosine-based alignment, formalized in information-retrieval literature, and rooted in directional statistics and linear algebra, where the cosine of the angle between normalized vectors is used to quantify geometric alignment [20].

D. ATTENTION SELECTIVITY INDEX

To quantify how concentrated the attention distribution is across tokens, we adapt the *Herfindahl-Hirschman Index* (HHI) [21], a metric traditionally used in economics to measure market concentration. In the context of attention, we define the *Attention Selectivity Index* as

$$\mathcal{S}(A_i) = 1 - \frac{\sum_j A_{ij}^2}{\left(\sum_j A_{ij}\right)^2} \tag{12}$$

where $A_i \in \mathbb{R}^n$ is the attention distribution over n tokens for a given head and query. A lower selectivity value implies higher concentration –that is, attention is sharply focused on fewer tokens –which may indicate more interpretable and localized information routing. To our knowledge, this is the first use of an HHI-derived index to quantify concentration in Transformer attention distributions.

E. SPECTRAL SPARSITY

To measure whether SVDA prunes unimportant directions, we define:

$$\mathcal{P}(\Sigma) = \frac{|\{\sigma_i : |\sigma_i| < \varepsilon\}|}{d_k} \tag{13}$$

for a small threshold ε. This indicates how many dimensions are effectively ignored. Unlike prior approaches that analyze sparsity in attention weights or perform post-hoc pruning of projection matrices [22], our formulation defines *Spectral Sparsity* directly over the learned diagonal matrix Σ, introduced as part of the SVDA attention mechanism. This allows us to assess how many latent semantic directions are *actively used* during attention computation. If many σ_i values are near zero, the model implicitly suppresses uninformative dimensions, revealing its internal spectral focus. This metric thus provides a built-in, interpretable signal of dimensional pruning–complementary to entropy or rank–and may support future compression, pruning, or attention regularization strategies.

F. ROBUSTNESS TO INPUT PERTURBATIONS

Let $x' = x + \delta$ be a perturbed version of the input with small additive noise. We assume $\|\delta\| \ll 1,\ \delta \sim \mathcal{N}(0, \epsilon^2)$, meaning that the perturbation has small magnitude relative to the scale of x, typically normalized pixel values in $[0, 1]$ or embeddings with unit norm. To measure robustness, we define the change in the attention distribution as:

$$\Delta A = \|A(x) - A(x + \delta)\|_F \tag{14}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This indicator quantifies how much the attention matrix A changes in response to small input perturbations. Prior work has studied theoretical local sensitivity bounds [23], formal verification of attention under semantic shifts [24], and adversarial stability of attention mechanisms [25]. In contrast, our formulation offers a practical, continuous, and interpretable diagnostic, by enabling direct head-wise and layer-wise robustness monitoring during training and evaluation, reinforcing SVDA's suitability for transparent, structured attention modeling.

TABLE 3 summarizes and categorizes these interpretability and structure-aware indicators. While the indicators offer structured insight into the internal behavior of SVDA, it is important to distinguish between their roles in practice:

- Some indicators help quantify geometric or spectral properties (e.g., entropy, angular alignment).
- Others describe sparsity, robustness, or selectivity patterns that are relevant for human understanding of model behavior.
- Only one satisfies the formal axioms of a metric in the mathematical sense.

Apparently, according to the classical formalism for a *metric*, in which a function $d : X \times X \to \mathbb{R}$ is a formal *metric* if it satisfies non-negativity, identity of indiscernibles, symmetry, and the triangle inequality, the following can be stated about our indicators:

- Spectral Entropy and Effective Rank are *not metrics*, but scalar "summaries" of structure.
- Angular Alignment, if expressed via angle, or $1-\cos(\theta)$, is a valid similarity-based *pseudo-metric*, as it may yield zero for non-identical but colinear vectors.[5]
- Selectivity and Sparsity are scalar indicators of structure and sparsity, *not distances*.
- Frobenius Norm used in perturbation robustness is a valid matrix norm and *satisfies all metric axioms*.

Furthermore, while the defined indicators do not provide direct feature-level attributions or counterfactual explanations, they enhance interpretability in the sense of:

- Making attention behavior *inspectable* (via Σ, A).
- Enabling reasoning about *which latent directions matter*.

[5] A *pseudo-metric* is a distance-like function $d : X \times X \to \mathbb{R}_{\geq 0}$ satisfying non-negativity, symmetry, and the triangle inequality, but not necessarily the identity of indiscernibles–meaning $d(x, y) = 0$ does not imply $x = y$. For example, the cosine distance $d_{\cos}(x, y) = 1 - (x \cdot y)/(\|x\|\|y\|)$ yields zero when $x = \lambda y$ with $\lambda > 0$, even though $x \neq y$. It is distinct from a *measure*, which quantifies the size of subsets of a space, rather than distances between elements.

TABLE 3. Interpretability-oriented indicators categorized by formal properties.

Indicator	Definition	Type	Interpretability Contribution
Spectral Entropy	$\mathcal{H}_\Sigma = -\sum p_i \log p_i$	Information indicator	Captures focus/disorder in latent dimension usage
Effective Rank	$\text{rank}_{\text{eff}}(\Sigma) = e^{\mathcal{H}_\Sigma}$	Scalar Indicator	Measures compressibility or dimension reduction capacity
Angular Alignment	$\cos\theta_{ij} = Q_i \cdot K_j$	Pseudo-metric	Reflects semantic closeness in latent space
Selectivity Index	$\mathcal{S}(A_i) = 1 - \frac{\sum_j A_{ij}^2}{\left(\sum_j A_{ij}\right)^2}$	Statistical indicator	Token-level sparsity and focus in attention matrix
Spectral Sparsity	$\mathcal{P}(\Sigma) = \frac{\lvert\{\sigma_i : \lvert\sigma_i\rvert < \varepsilon\}\rvert}{d_k}$	Structural indicator	Indicates inactive latent directions and pruning potential
Perturbation Robustness	$\Delta A = \lVert A(x) - A(x+\delta)\rVert_F$	Valid Metric	Quantifies behavioral consistency under small perturbations

- Allowing diagnosis of *spectrum sparsity*, *focus*, and *robustness*.
- Providing quantitative tools for model designers to *control or regularize* attention behavior.

We therefore propose these indicators as part of a structural diagnostic toolkit for understanding attention dynamics in models that prioritize both performance and transparency.

G. ABSENCE OF STRUCTURED INDICATORS IN BASELINE ATTENTION

The baseline model computes attention via raw scaled dot products

$$A_{\text{base}} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \tag{15}$$

This formulation does not support an explicit decomposition of similarity or importance across latent directions. As a result, structural indicators such as spectral entropy, effective rank, or sparsity cannot be meaningfully defined or measured.

V. DISCUSSION

SVDA introduces a principled structure to attention by blending the geometry of cosine similarity with the spectral scaling properties of SVD. Despite quadratic time complexity, SVDA enables interpretability, making it a candidate for research-focused or hybrid-efficiency models. Examples presented here serve an illustrative role. They aim to reveal the internal dynamics of SVDA, including entropy decay, selectivity, and angular alignment, using small-scale settings. These insights help visualize the structural differences introduced by SVDA, rather than provide competitive benchmarks or application-specific optimization.

While our current implementation learns Σ without constraints, future extensions could include L_1 or soft-thresholding regularization to promote sparsity and interpretability in the spectral profile. This opens avenues for rank-adaptive attention and spectral pruning.

Demonstrative Results: While this paper prioritizes theoretical formulation over exhaustive empirical tuning, we run a number of experiments to illustrate the interpretability and structural behavior of SVDA across diverse input modalities. Specifically, we conducted three representative sets of experiments on standard benchmarks from time series, text, and image domains. In all cases, the experimental design isolates the role of SVDA between baseline Transformer attention and our proposed formulation, keeping all other hyperparameters constant. Each configuration is trained for 100 epochs using the Adam optimizer with learning rate 3×10^{-4} and appropriate orthogonality regularization weight $\lambda_{\text{ortho}} \in \{10^{-1}, 10^{-2}, 10^{-3}\}$ respectively.

- **ECG5000 (time series)**: We use univariate sequences of length 140 with 5-class classification. The model comprises 4 layers, 2 attention heads, and an embedding dimension of 256. This setup tests the model's ability to capture fine-grained temporal patterns in physiological signals.
- **SST-2 (text)**: For sentiment classification over sentence-level inputs, we apply a lightweight Transformer with 2 layers and 4 heads, using a vocabulary (size 30522), and a maximum input length of 64 tokens. This evaluates SVDA in natural language settings with dense semantic structure.
- **CIFAR-10 (image)**: We train a Vision Transformer using 4×4 patches over 32×32 RGB images, with 4 layers and 4 heads. This configuration enables analysis of spatial attention dynamics and visual interpretability over 10 classes.

These datasets were chosen not for benchmarking, but to illustrate how interpretability metrics behave across diverse input modalities: (i) time series (ECG5000), (ii) text with discrete token structure (SST-2), and (iii) structured vision tasks (CIFAR-10). Each domain highlights different facets of attention behavior, offering a representative testbed for the proposed indicators.

All three models share the same embedding dimension ($d_{\text{model}} = 256$) and batch size (64), allowing for fair comparison across tasks. For each run, we log all six proposed interpretability metrics per head, per layer, and per epoch, providing a consistent diagnostic view of the internal attention behavior induced by SVDA. For Spectral Entropy, Effective Rank, and Spectral Sparsity, reported values are averaged across all heads and all layers to yield a single scalar value for each training epoch. This global aggregation captures the overall structural behavior of attention across the network, reflecting emergent patterns of selectivity, compression, and spectral concentration throughout the model depth. In contrast, the Selectivity Index, Perturbation Robustness,

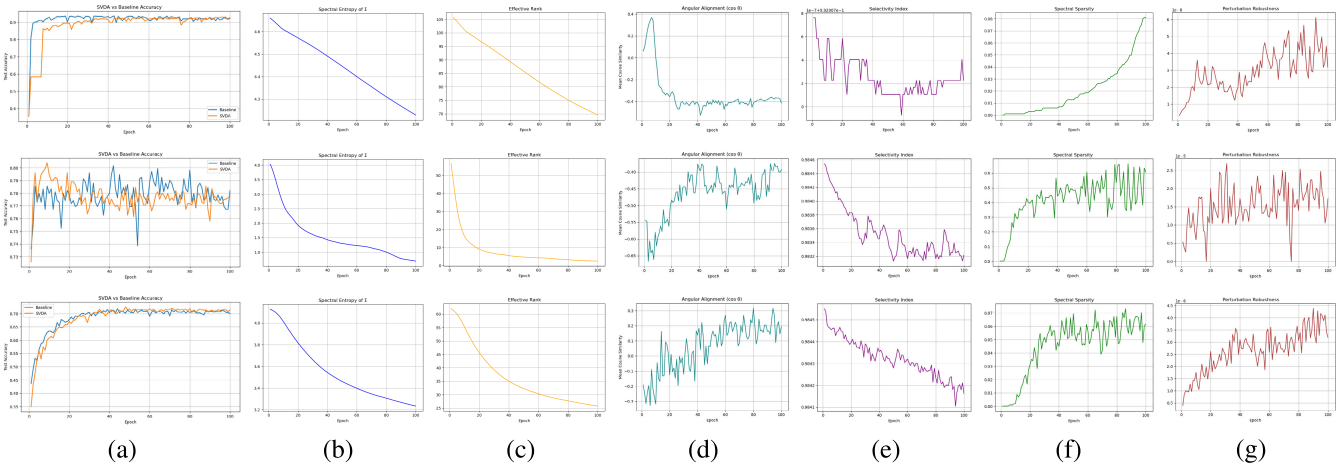

FIGURE 1. Evolution of attention statistics during training with SVDA across three datasets: ECG5000 (top row), SST-2 (middle row), and CIFAR-10 (bottom row). For each dataset, we plot (from left to right): (a) test accuracy, (b) spectral entropy, (c) effective rank, (d) angular alignment, (e) selectivity index, (f) spectral sparsity, and (g) perturbation robustness.

and Angular Alignment are reported using only the final attention layer. The rationale is that these indicators are most informative when measured at the layer directly influencing the model's output. This differentiated treatment ensures that each metric reflects the phenomenon it is best suited to capture, thus reporting them this way avoids dilution of their interpretive value and aligns the measurement with the point of semantic decision, balancing architectural insights with decision-focused analysis. All metrics are computed at the end of each training epoch and plotted to visualize their trajectories, providing insight on the inner workings of the tested architectures. In addition, the same architectures and tests were run for the corresponding *baseline attention* architecture just to capture and compare the accuracy in both cases, to showcase that there is no practical degradation in that aspect.

FIGURE 1 presents all the attained results, including the accuracy comparison, and the indicators for each of the experiment evolving by epoch. The top row shows the results for ECG5000, the middle row for SST-2, and the bottom row for CIFAR-10. It is apparent that the accuracy attained by SVDA is similar to the baseline, while the indicators convey significant information on the evolution of the attention 'statistics', as discussed in the following. The dropping *Spectral Entropy indicator* reveals that fewer singular values dominate, signaling that attention is more focused or selective. Accordingly, from the decreasing *Effective Rank indicator*, we may deduce that the model is learning to rely on fewer informative directions, potentially discarding noise or irrelevant features. Interestingly, the evolution of *Angular Alignment indicator* exhibits modality-specific behavior: in the time-series task, angular alignment consistently decreases and remains negative, suggesting that the model learns to emphasize directional dissimilarity between queries and keys, potentially to highlight temporal transitions; in the NLP experiment, alignment values start negative and gradually increase, reflecting a shift toward more coherent latent structures without reaching full alignment; in the image classification setting, angular alignment increases and becomes positive, indicating that spatial attention benefits from aligned query-key directions, consistent with the locality and redundancy of image patches. The *Selectivity Index indicator* exhibits subtle variations across all three experimental settings: in the time-series task, the index initially decreases over training epochs, suggesting a diffusion of attention focus, but later exhibits a minor upward correction, though the overall change remains within a negligible magnitude; for both the NLP and image classification tasks, the Selectivity Index shows a consistent but marginal decline, indicating a stable yet slightly decreasing trend in the sharpness of attention distribution. The *Spectral Sparsity indicator* increases across training epochs in all experimental settings, reflecting a growing number of near-zero components in the learned spectral weights: in the time-series experiment, the subtle increase indicates modest emergence of inactive spectral dimensions; in the NLP case, sparsity rises more markedly, although with visible fluctuations, suggesting intermittent shifts in spectral pruning behavior; for the image classification task, the trend is similar to that of NLP but with a tighter range. The *Perturbation Robustness indicator* exhibits an increasing trend across all experimental settings, indicating a growing sensitivity of the attention maps in the final layer to small input perturbations: in the time-series experiment, the Frobenius norm difference remains within a very narrow range, reflecting extremely stable attention behavior; for the NLP experiment, robustness values reveal significantly stable behavior; in the image classification case, the corresponding values reveal very stable behavior. It should be noted here that by using the heuristics of matrix norm scaling, the attention matrix size, the precision limits of floats in practical implementations, and the empirical behavior in real models, rough limits in the characterization of robustness can be set, for example as shown in TABLE 4.

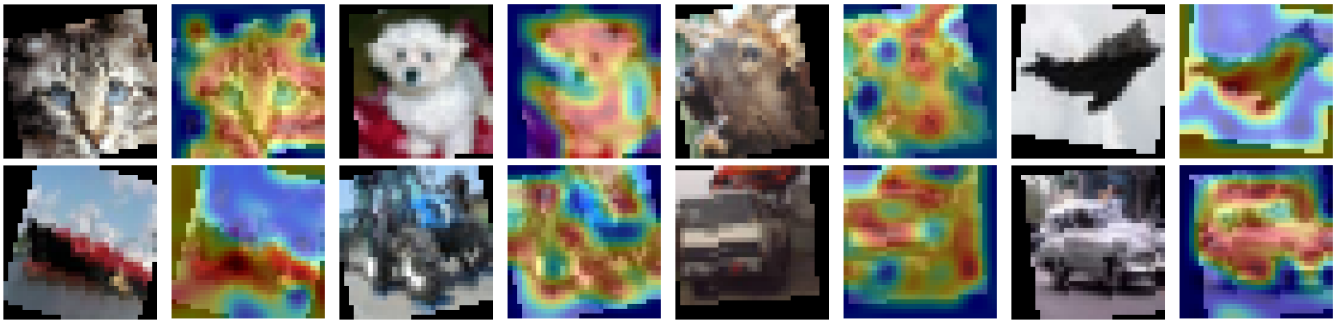

FIGURE 2. Visualization of random samples and attention map pairs.

TABLE 4. Interpretation thresholds for attention Perturbation magnitudes.

Perturbation Magnitude	Interpretation
$< 10^{-7}$	Below numerical precision (float32); negligible change
$10^{-7} - 10^{-6}$	Extremely stable attention; changes within noise floor
$10^{-6} - 10^{-5}$	Minor structural variation; generally stable
$10^{-5} - 10^{-4}$	Noticeable attention shift; potentially meaningful
$10^{-4} - 10^{-3}$	Substantial alteration; likely to affect model behavior
$> 10^{-3}$	Significant disruption; often indicative of instability or adversarial sensitivity

It should be stressed that those limits are indicative and can be adjusted per domain of application. These thresholds were chosen based on empirical ranges observed during training, informed by the numerical precision of floating-point operations and practical experience across the three domains. They serve as domain-agnostic heuristics but can be further calibrated per application.

In addition, FIGURE 2 presents random samples and attention map pairs from the image classification experiment. Qualitative inspection of attention maps confirms that SVDA preserves the ability of the model to focus attention meaningfully. The visual patterns remain coherent and semantically aligned with salient image regions, indicating that the structural constraints introduced by Σ-based modulation do not hinder the effective operation of the attention mechanism. This supports the claim that SVDA enhances interpretability without sacrificing the model's capacity for token-wise or region-wise focus. While detailed visualizations are only shown for the image domain, analogous attention map inspections for SST-2 and ECG5000 confirm domain-aligned patterns, such as token-cluster focus in NLP and transition-points in time series, validating the generality of SVDA's interpretability behavior across modalities.

Implications for Interpretability and Design: The additional structure in SVDA not only improves our ability to inspect what the model is attending to, but also opens pathways for architectural refinement. For instance, spectral sparsity could be exploited for model pruning, and high entropy could be regularized for robustness or fairness.

Computational Considerations: While SVDA preserves the quadratic complexity of standard attention ($\mathcal{O}(n^2 d_k)$), it introduces two additional costs: (i) *orthogonality regularization* for W_Q, W_K, which adds a term of cost $\mathcal{O}(d_k^2)$ per head during training, and (ii) *spectral modulation* via a learned diagonal matrix $\Sigma \in \mathbb{R}^{d_k \times d_k}$, applied once to each query or key vector, with cost $\mathcal{O}(nd_k)$ per head. These costs are negligible compared to full dot-product attention when $n \gg d_k$, but may modestly increase training time due to the regularization gradients. Importantly, SVDA incurs no significant inference-time penalty since Σ is diagonal and shared across all positions (the matrix-vector product $Q\Sigma$ reduces to element-wise scaling across feature dimensions). Further optimization is possible via low-rank or sparse spectral approximations in future work. In our experiments, training time increased by approximately 25% (on average) relative to baseline attention, while inference speed remained virtually unchanged, confirming that SVDA remains practically viable.

Limitations: While SVDA enhances interpretability, it incurs additional computational overhead due to spectral modulation and orthogonality constraints. Scaling to very large models may require low-rank approximations or hybridization with efficient attention mechanisms.

VI. CONCLUSION

In this paper, we introduced SVD-Inspired Attention (SVDA), a novel attention mechanism that incorporates learned spectral modulation over orthonormal query/key projections, enabling structured, interpretable attention without altering Transformer architecture. This work provides a theoretical and structural foundation for SVDA, offering a spectrum-based perspective on attention that supports diagnostics, interpretability, and potential compression. Unlike methods driven primarily by empirical performance, SVDA is designed from first principles, prioritizing semantic transparency and geometric decomposition, and presented here as a theoretical foundation work. Despite its added spectral structure, SVDA incurs minimal inference-time overhead, with only modest training cost introduced via orthogonality regularization and head-specific spectral modulation. SVDA thus lays the groundwork for a family of attention

mechanisms guided by interpretable design constraints, rather than solely empirical heuristics. As SVDA is designed as a holistic structural modification, where query normalization, orthogonality, and spectral modulation co-act to enforce interpretability, component-wise ablation would misrepresent its intended function and obscure the emergent properties that arise from their interaction, thus it is omitted from this paper. Future research will focus on adapting SVDA to different applications, attaining interpretability, and even improving performance through prescriptive use of its diagnostic power, in diverse scenarios.

APPENDIX A
THEORETICAL FOUNDATIONS OF SVDA

This Appendix provides the theoretical foundations of SVDA, formalizing its structural properties and offering analytical justifications for the interpretability metrics (indicators) introduced in the main text. Through rigorous treatment of orthogonality, spectral modulation, and normalized projections, we derive key properties such as row-stochasticity, bounded attention dynamics, and spectrum-based controllability. These results support the claim that SVDA enables interpretable and expressive attention without increasing architectural complexity. For analytical clarity without loss of generality, we consider the case of a single attention layer with a single head.

A. THEORETICAL PROPERTIES OF SVDA

Theorem 1 (SVDA Attention is Row-Stochastic and Spectrally Modulated): Let $Q, K \in \mathbb{R}^{n\times d_k}$ be row-wise ℓ_2-normalized projections (i.e., each row Q_i, K_j satisfies $\|Q_i\|_2 = \|K_j\|_2 = 1$), and let $\Sigma \in \mathbb{R}^{d_k\times d_k}$ be a learned diagonal matrix. Define the attention matrix as:

$$A = \text{softmax}\left(\frac{Q\Sigma K^\top}{\sqrt{d_k}}\right)$$

Then:

1) $A \in \mathbb{R}^{n\times n}$ is row-stochastic: $\sum_j A_{ij} = 1, A_{ij} \geq 0$
2) The matrix Σ enables spectral reweighting of attention components
3) When $\Sigma = I$, SVDA reduces to standard cosine (angular) attention

Proof: Let $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d_k})$. Then each entry of the un-normalized score matrix $S \in \mathbb{R}^{n\times n}$ is:

$S_{ij} = \frac{1}{\sqrt{d_k}} \sum_{\ell=1}^{d_k} \sigma_\ell Q_{i\ell} K_{j\ell} = \frac{1}{\sqrt{d_k}} \langle Q_i, \Sigma K_j\rangle$

Because $Q_i, K_j \in \mathbb{R}^{d_k}$ are unit-norm vectors, and Σ is diagonal, S_{ij} is a *weighted inner product* over dimensions, where σ_ℓ modulates the contribution of each dimension ℓ. This shows that Σ acts as a spectral modulator, selectively scaling dimensions in the attention dot product, thereby establishing point (2).

By applying the softmax function row-wise over S, we define:

$A_{ij} = \frac{e^{S_{ij}}}{\sum_{j'} e^{S_{ij'}}}$ with $A_{ij} \geq 0, \quad \sum_j A_{ij} = 1$

These properties follow from the standard definition of the softmax function and imply that A is *row-stochastic*, satisfying point (1) (cf. [26], [27] for standard softmax properties).

Finally, when $\Sigma = I$, we have:

$S_{ij} = \frac{1}{\sqrt{d_k}} Q_i^\top K_j$

Since both vectors are normalized, this is their *cosine similarity*, and so SVDA becomes cosine-based attention, satisfying (3). □

1) COROLLARIES AND INTERPRETABILITY IMPLICATIONS

(C1) **Spectral Reweighting (Controllability in Feature Space).**
The diagonal matrix $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d_k})$ scales each feature direction independently during attention scoring:

$$(Q\Sigma K^\top)_{ij} = \sum_{\ell=1}^{d_k} \sigma_\ell Q_{i\ell} K_{j\ell}$$

Each σ_ℓ acts as a learned modulation over the contribution of the corresponding latent dimension. This spectral decomposition provides insight into which feature subspaces are emphasized during attention –enabling compression, pruning, or emphasis of semantically relevant dimensions.

(C2) **Reduction to Angular Attention.**
In the special case where $\Sigma = I$, the formulation simplifies to:

$$A_{ij} = \text{softmax}_j\left(\frac{Q_i \cdot K_j}{\sqrt{d_k}}\right)$$

Since Q_i, K_j are normalized, this is equivalent to computing cosine similarity:

$$Q_i \cdot K_j = \cos\theta_{ij}$$

This shows that SVDA generalizes angular attention by introducing a spectrum-aware scaling term that enriches directional selectivity.

(C3) **Row-Stochasticity Enables Probabilistic Semantics.**
The attention matrix $A \in \mathbb{R}^{n\times n}$ is row-stochastic by construction:

$$\forall i : \sum_j A_{ij} = 1 \quad \text{and} \quad A_{ij} \geq 0$$

This implies that each output token is computed as a convex combination of value vectors V:

$$Z_i = \sum_{j=1}^{n} A_{ij} V_j$$

As a result:

- It permits the interpretation of A_i as a probability distribution over tokens.
- It guarantees the boundedness of outputs ($Z_i \in \text{conv}(\{V_j\})$), avoiding uncontrolled activation growth.

- It facilitates the use of well-defined interpretability indicators such as entropy:

$$\mathcal{H}(A_i) = -\sum_j A_{ij} \log A_{ij}$$

and selectivity:

$$\mathcal{S}(A_i) = 1 - \frac{\sum_j A_{ij}^2}{\left(\sum_j A_{ij}\right)^2}$$

which depend on valid distributions over the token set.
- It aligns the attention mechanism with theoretical tools from Markov chains, diffusion processes, and information theory, which all assume stochastic or doubly-stochastic operators.

Theorem 2 (Orthogonality Regularization Leads to Bounded Projections): Let $W_Q, W_K \in \mathbb{R}^{d\times d_k}$ be the projection matrices for queries and keys, respectively. Define the orthogonality regularization loss as:

$\mathcal{L}_{\text{ortho}} = \|W_Q^\top W_Q - I\|_F^2 + \|W_K^\top W_K - I\|_F^2$

Then, minimizing $\mathcal{L}_{\text{ortho}}$ leads to:

1) The columns of W_Q and W_K becoming approximately orthonormal.
2) The spectral norms of W_Q and W_K approaching 1, i.e., $\|W_Q\|_2 \to 1, \quad \|W_K\|_2 \to 1$

Proof: Consider the Frobenius penalty term for a generic projection matrix $W \in \mathbb{R}^{d\times d_k}$:

$\|W^\top W - I\|_F^2 = \sum_{i=1}^{d_k}\sum_{j=1}^{d_k}\left(\langle w_i, w_j\rangle - \delta_{ij}\right)^2$

where w_i denotes the i-th column of W, and δ_{ij} is the Kronecker delta. This term penalizes deviations from orthonormality in the columns of W.

By the singular value decomposition (SVD), $W = U\Sigma V^\top$, where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d_k})$. Then:

$W^\top W = V\Sigma^\top\Sigma V^\top, \quad \text{so} \quad \|W^\top W - I\|_F^2 = \|\Sigma^2 - I\|_F^2$

Therefore:

$\|W^\top W - I\|_F^2 = \sum_{i=1}^{d_k}(\sigma_i^2 - 1)^2$

This is minimized when each $\sigma_i = 1$, i.e., when all singular values of W equal 1. This implies that W has orthonormal columns (i.e., $W^\top W = I$) and hence:

$\|W\|_2 = \max_i \sigma_i = 1$

This ensures that W_Q and W_K are *norm-preserving*, bounded in operator norm, and yield stable geometric transformations. The result follows from standard matrix analysis [28]. □

2) COROLLARIES AND INTERPRETABILITY IMPLICATIONS

(C1) **Norm-Preserving Projections (Geometric Stability).**
When the orthogonality loss $\mathcal{L}_{\text{ortho}} = \|W_Q^\top W_Q - I\|_F^2 + \|W_K^\top W_K - I\|_F^2$ is minimized, the learned projection matrices approximate orthonormality. Thus, for any input $x \in \mathbb{R}^d$, the projected vectors satisfy:

$$\|W_Q^\top x\|_2 \le \|x\|_2, \quad \|W_K^\top x\|_2 \le \|x\|_2$$

This bounds the energy of query/key vectors and prevents unbounded activation growth in the attention dot products.

(C2) **Spectral Regularity and Bounded Operator Norm.**
As the singular values of W_Q, W_K converge to 1, the operator norm of each projection satisfies:

$$\|W_Q\|_2 \to 1, \quad \|W_K\|_2 \to 1$$

This guarantees numerical stability and ensures that no direction is disproportionately amplified, which improves conditioning for both the attention matrix and gradients.

(C3) **Interpretability via Angular Similarity.**
Orthonormal projections enable the separation of magnitude and direction. Given row-wise ℓ_2-normalized outputs:

$$\hat{Q}_i = \frac{W_Q^\top x_i}{\|W_Q^\top x_i\|}, \quad \hat{K}_j = \frac{W_K^\top x_j}{\|W_K^\top x_j\|}$$

the dot product $\hat{Q}_i \cdot \hat{K}_j = \cos\theta_{ij}$ encodes pure directional alignment, independent of scale. This supports meaningful geometric analysis of attention patterns.

(C4) **Improved Gradient Flow.**
The Frobenius regularization stabilizes the gradient landscape by reducing anisotropy in weight matrices. Since

$$\frac{\partial \mathcal{L}}{\partial W} \propto (W^\top W - I)W$$

the update dynamics push the weight matrix toward a flat, orthonormal manifold, reducing curvature in the optimization trajectory.

Theorem 3 (Spectral Modulation Enables Dimension-Wise Control): Let $\hat{Q}, \hat{K} \in \mathbb{R}^{n\times d_k}$ be row-wise normalized query and key matrices such that $\|\hat{Q}_i\|_2 = \|\hat{K}_j\|_2 = 1$ for all i, j, and let $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d_k}) \in \mathbb{R}^{d_k\times d_k}$ be a learned diagonal matrix of spectral weights. Then the attention score between tokens i and j is:

$\hat{Q}_i^\top \Sigma \hat{K}_j = \sum_{k=1}^{d_k} \sigma_k \cdot q_{ik} \cdot k_{jk}$

This defines a spectrally weighted inner product, enabling selective amplification or suppression of latent components.

Proof: Let $q_i = \hat{Q}_i \in \mathbb{R}^{d_k}$, $k_j = \hat{K}_j \in \mathbb{R}^{d_k}$. Then:

$q_i^\top \Sigma k_j = \sum_{k=1}^{d_k} q_{ik} \cdot \sigma_k \cdot k_{jk}$

This follows directly from the definition of matrix-vector multiplication with a diagonal operator. Because Σ is diagonal, there is no cross-feature mixing; each latent dimension k contributes independently, scaled by σ_k. This enables token-level attention control in the spectral domain. See [28] for properties of bilinear forms with diagonal matrices. □

3) COROLLARIES AND INTERPRETABILITY IMPLICATIONS

(C1) **Selective Feature Emphasis (Directional Reweighting).**
Each entry σ_k of Σ scales the attention interaction along the k-th latent direction. Large values of σ_k

enhance importance of corresponding basis vectors $q_{\cdot k}, k_{\cdot k}$, while small (or near-zero) values suppress them. This supports:

- Pruning of irrelevant directions,
- Emphasis of task-relevant semantic axes,
- Compression of latent space.

(C2) **Spectral Interpretability via Σ.**
The set $\{\sigma_k\}$ can be interpreted as a spectral profile of the attention mechanism, enabling the definition of indicators like:

$$\text{Spectral Entropy: } \mathcal{H}_\Sigma = -\sum_k p_k \log p_k,$$

$$\text{with } p_k = \frac{\sigma_k^2}{\sum_\ell \sigma_\ell^2}$$

and effective rank:

$$\text{rank}_{\text{eff}}(\Sigma) = \exp(\mathcal{H}_\Sigma)$$

These summarize diversity, focus, and compressibility of spectral usage.

(C3) **Structure-Preserving Sparsification.**
Since modulation is diagonal, sparse configurations of Σ induce structured sparsity without altering projection geometry. We define:

$$\mathcal{P}(\Sigma) = \frac{|\{k : |\sigma_k| < \varepsilon\}|}{d_k}$$

as the spectral sparsity index – counting the proportion of suppressed dimensions.

(C4) **Smooth Differentiable Control of Subspace Attention.**
Unlike hard masking or low-rank truncation, SVDA provides a differentiable means of reweighting attention contributions per latent direction. This enables:

- Learning interpretable spectra end-to-end,
- Avoiding discrete discontinuities (e.g., thresholding),
- Smooth interpolation between full and pruned attention modes.

B. MATHEMATICAL JUSTIFICATION OF INTERPRETABILITY INDICATORS

We define a set of interpretable indicators that quantify the structure and robustness of attention, especially in the context of spectral modulation via SVDA.

1) SPECTRAL ENTROPY

Given $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d_k})$, define the normalized spectral energy:

$$p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2}$$

Then the spectral entropy is:

$$\mathcal{H}_\Sigma = -\sum_{i=1}^{d_k} p_i \log p_i$$

This measures the disorder of energy distribution across latent directions. Lower entropy reflects sharper and more selective use of dimensions.

2) EFFECTIVE RANK

The soft estimate of active dimensions is given by:

$$\text{rank}_{\text{eff}}(\Sigma) = \exp(\mathcal{H}_\Sigma)$$

It reflects the number of effectively used directions and complements the entropy.

3) SPECTRAL SPARSITY

Given a small threshold $\varepsilon > 0$, define:

$$\mathcal{P}(\Sigma) = \frac{|\{i : |\sigma_i| < \varepsilon\}|}{d_k}$$

This fraction quantifies how many spectral directions are effectively pruned or inactive.

4) SELECTIVITY INDEX (BASED ON HERFINDAHL-HIRSCHMAN INDEX)

For attention row vector $A_i \in \mathbb{R}^n$, define:

$$\mathcal{S}(A_i) = 1 - \frac{\sum_j A_{ij}^2}{\left(\sum_j A_{ij}\right)^2}$$

It quantifies the spread of attention. Lower values indicate token-level sharpness (fewer tokens attended), higher values imply uniformity.

5) ANGULAR ALIGNMENT

Using normalized projections $\hat{Q}_i = Q_i/\|Q_i\|$, $\hat{K}_j = K_j/\|K_j\|$, define:

$$\cos(\theta_{ij}) = \hat{Q}_i \cdot \hat{K}_j$$

This reflects directional similarity in latent space. The pseudo-distance $1 - \cos(\theta_{ij})$ encodes angular discrepancy.

6) PERTURBATION ROBUSTNESS

Let $x' = x + \delta$ be a perturbed input with $\|\delta\| \ll 1$. Then:

$$\Delta A = \|A(x) - A(x')\|_F$$

measures the Frobenius norm of change in attention under input noise. Low values imply more stable, geometry-aware attention.

ACKNOWLEDGMENT

This work has been partially supported by the ARGUS EU project (Grant Agreement No. 101132308), funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or of the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[2] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," 2020, *arXiv:2006.04768*.

[3] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Davis, A. Mohiuddin, Ł. Kaiser, D. Belanger, L. J. Colwell, and A. Weller, "Rethinking attention with performers," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2020, pp. 1–16.

[4] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh, "Nystrmformer: A nystrm-based algorithm for approximating self-attention," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 14138–14148.

[5] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. 37th Int. Conf. Mach. Learn.*, Jan. 2020, pp. 5156–5165.

[6] Y. Wu, S. Kan, M. Zeng, and M. Li, "Singularformer: Learning to decompose self-attention to linearize the complexity of transformer," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 4433–4441.

[7] Y. Chen, Q. Tao, F. Tonin, and J. A. K. Suykens, "Primal-attention: Self-attention through asymmetric kernel SVD in primal representation," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 65088–65101.

[8] Z. Qin, W. Sun, H. Deng, D. Li, Y. Wei, B. Lv, J. Yan, L. Kong, and Y. Zhong, "CosFormer: Rethinking softmax in attention," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–14.

[9] D. Bo, C. Shi, L. Wang, and R. Liao, "Specformer: Spectral graph neural networks meet transformers," in *Proc. 11th Int. Conf. Learn. Represent.*, Jan. 2023, pp. 1–18.

[10] S. Shleifer, J. Weston, and M. Ott, "NormFormer: Improved transformer pretraining with extra normalization," in *Proc. 10th Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–18.

[11] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.

[12] J. Bastings and K. Filippova, "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?" in *Proc. 3rd BlackboxNLP Workshop Analyzing Interpreting Neural Netw.*, 2020, pp. 149–155.

[13] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.

[14] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2020, pp. 1–18.

[15] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 17283–17297.

[16] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, "Efficient content-based sparse attention with routing transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 53–68, Feb. 2021.

[17] G. M. Correia, V. Niculae, and A. F. T. Martins, "Adaptively sparse transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, 2019, pp. 2174–2184.

[18] L. L. Campbell, "Minimum coefficient rate for stationary random processes," *Inf. Control*, vol. 3, no. 4, pp. 360–371, Dec. 1960.

[19] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimensionality," in *Proc. 15th Eur. Signal Process. Conf.*, Sep. 2007, pp. 606–610.

[20] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, 1983.

[21] S. A. Rhoades, "The herfindahl-hirschman index," *Federal Reserve Bull.*, vol. 79, no. 3, pp. 188–189, Jan. 1993.

[22] H. Li, M. Wang, S. Zhang, S. Liu, and P.-Y. Chen, "Learning on transformers is provable low-rank and sparse: A one-layer analysis," in *Proc. IEEE 13rd Sensor Array Multichannel Signal Process. Workshop (SAM)*, Jul. 2024, pp. 1–5.

[23] A. J. Havens, A. Araujo, H. Zhang, and B. Hu, "Fine-grained local sensitivity analysis of standard dot-product self-attention," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 1–16.

[24] S. Munakata, C. Urban, H. Yokoyama, K. Yamamoto, and K. Munakata, "Verifying attention robustness of deep neural networks against semantic perturbations," in *Proc. 29th Asia–Pacific Softw. Eng. Conf. (APSEC)*, Dec. 2022, pp. 560–561.

[25] S. Kitada and H. Iyatomi, "Attention meets perturbations: Robust and interpretable attention with adversarial training," *IEEE Access*, vol. 9, pp. 92974–92985, 2021.

[26] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

VASILEIOS ARAMPATZAKIS is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece. He is currently a Scientific Associate with the Athena Research Center. His research interests include digital image processing, computer vision, machine learning, 3D digitization and reconstruction, and document processing.

GEORGE PAVLIDIS (Senior Member, IEEE) received the Diploma and Ph.D. degrees in electrical and computer engineering. He is currently the Research Director of the Athena Research Center and also heading the Media Department. His research interests include digital image and multimedia technologies, content analysis and retrieval, machine learning and artificial intelligence, 3D digitization, and cross-reality applications.

NIKOLAOS MITIANOUDIS (Senior Member, IEEE) received the Diploma and Ph.D. degrees in electronic and computer engineering. In 2010, he joined the Electrical and Computer Engineering Department, Democritus University of Thrace, where he is currently an Associate Professor. His research interests include applications of statistical signal processing, machine learning, audition, and vision.

NIKOS PAPAMARKOS (Senior Member, IEEE) received the Diploma degree in electrical and mechanical engineering and the Ph.D. degree in electrical engineering. He has been a Professor with the Democritus University of Thrace, since 2003. His research interests include digital image processing, computer vision, document processing, analysis and recognition, pattern recognition, neural networks, signal processing, filter design, and optimization algorithms.