DnA: Denoising Attention for Visual Tasks

Ron Campos[1], Subhajit Maity[1], Xin Li[1], Srijan Das[2], Aritra Dutta[1]

[1] University of Central Florida, USA
[2] University of North Carolina at Charlotte, USA
https://rjccv.github.io/DnA

Abstract. The softmax activation in multihead attention (MHA) is the de facto standard for attention-based models in visual perception tasks. However, standard softmax can produce noisy attention patterns that dilute relevant features and degrade its performance. In this paper, we propose **De**noising **A**ttention or **DnA**, in which, first, a positive query identifies which image features belong to the correct class, and a negative query identifies closely associated but irrelevant image features. DnA then projects these interactions into two distinct subspaces with larger principal angles, promoting subspace separation and improved discriminability. Using a ViT-B backbone, our proposed DnA achieves an absolute gain of 0.8% on ImageNet-1K compared to the baseline. We further show improvements across multiple visual understanding tasks, including video understanding with video transformers (1.8%) and video LLMs (0.5%). Our extensive empirical analyses justify the design choices involving two interacting subspaces and the denoising effect of DnA. The code is publicly available at https://github.com/rjccv/DnA.

1 Introduction

The softmax function is a fundamental component in machine learning and statistics that transforms arbitrary real-valued scores into a probability distribution. It plays an essential role in the Transformer architecture, particularly in the self-attention mechanism that forms the core of modern large language models [4, 6, 41, 42, 55] and other sequence-to-sequence models [12, 34, 46]. While efficiency-driven alternatives exist for specific use cases [21, 22, 39, 43, 47, 52, 63, 64, 68], softmax remains the gold standard for attention normalization due to its unique combination of theoretical properties and empirical performance [18, 47, 54, 58, 59]. The attention in Transformers computes contextualized representations by allowing each position in a sequence to attend to all positions, with the attention weights determined by the softmax function; see details in §2.1.

Although the softmax amplifies the dominant scores, it treats positive and negative values asymmetrically; the numerator amplifies the gaps between positive scores and compresses the gaps between negative scores. Indeed, we have the following observation: *For $\delta > 0$, we have* $e^{a+\delta} - e^a = e^a(e^\delta - 1) = e^a\delta + O(\delta^2)$.

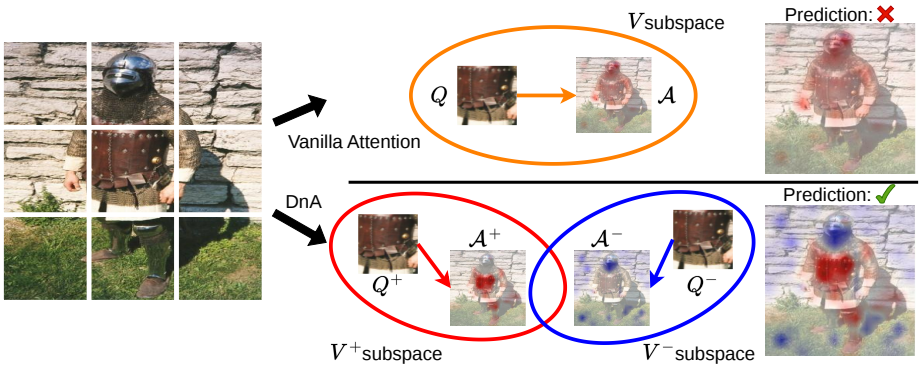

Fig. 1: The original image from ImageNet-1K [16] shows the primary object, **breastplate**, along with secondary adversarial objects, *person, helmet, etc.* The traditional attention [18, 56] and differential attention [62] project the interactions onto a single V subspace, resulting in misallocation of attention to secondary objects. Our proposed DnA explicitly models positive and negative interactions using two sets of queries, Q^+ and Q^-, and projects them onto two separate subspaces, V^+ and V^-, respectively, to effectively segregate closely associated adversarial features from the relevant ones. We define traditional attention as $\mathcal{A}$, and DnA branches as $\mathcal{A}^+$ and $\mathcal{A}^-$.

As a result, because softmax is used in transformers, larger key-query interactions receive much higher scores, while smaller interactions lose expressivity. This asymmetry may limit the representational capacity of standard attention.

Beyond score magnitudes, the geometry of feature subspaces also affects discriminability; larger principal angles between class subspaces reduce classification error [25]. This perspective motivates modeling positive and negative token interactions as a two-class problem, suggesting that subspace separation combined with both positive and negative score interactions can improve classifier performance. We present this in the context of the database query language [8,50], where the primary object might be associated with other secondary object classes. The key (token of interest), query (token-associated request), and value (token retrieval) interactions in traditional softmax attention; see Figure 1. In contrast to differential attention [62] for NLP tasks, which models the attention matrix using a single subspace V, we introduce two separate queries over the same keys. First, (i) Positive query: *Which features belong to the correct class?* and (ii) Negative query: *Which features do not belong to the correct class, but are closely associated?* Their interactions are projected into distinct subspaces, V^+ and V^-, with larger principal angles between them, promoting subspace separation and improved discriminability; see Figure 1.

Taken together, in this paper, we propose **De**noising **A**ttention or **DnA**, which replaces traditional attention in transformers; see §3. Since ViTs are integral to modern foundational models [28,44,45], we explore the efficacy of DnA in traditional transformers and Video LLMs. By replacing the traditional attention,

our proposed DnA outperforms ViT-B on ImageNet-1K [16] and achieves 0.8% higher test accuracy. DnA also excels on video understanding tasks, achieving up to an absolute 4.0% gain on Toyota Smarthome [15], 1.2% gain on NTU RGB+D 60 [51] with TimeSformer [3], and 0.5% gain on Ego-in-Exo PerceptionsMCQ [48] with the video LLM, VideoLLaMA3 [66]; see §4.

2 Background and Related Works

For completeness, we start with a brief description of multihead attention.

2.1 Multi-Head Attention (MHA) in Vision

Let $X \in \mathbb{R}^{C\times R\times W}$ be a C channel image with resolution $R \times W$. The image is flattened into N tokens in $\mathbb{R}^D$ by $p\times p$ patches such that, $N = \frac{RW}{p^2}$ and $D = p^2C$. The input, $X_p \in \mathbb{R}^{N\times D}$, is a sequence of N tokens of dimension D. Transformers process the input through a series of L encoder layers. For simplicity, we assume the input X_p is encoder-agnostic. We note that the following applies equally to decoder blocks; for brevity, we only refer to encoders.

In each encoder, *attention* projects an input sequence on the row-space of three learnable matrices $W_Q, W_K, W_V \in \mathbb{R}^{D\times D}$ to produce query, $Q = X_Q W_Q$, key, $K = X_{KV} W_K$, and value, $V = X_{KV} W_V$. In *self-attention*, the queries, keys, and values are derived from the same input, such that $X_Q = X_{KV} = X_p \in \mathbb{R}^{N\times D}$. In *cross-attention*, the queries are derived from one input sequence $X_Q \in \mathbb{R}^{N\times D}$, while the keys and values come from a different input sequence $X_{KV} \in \mathbb{R}^{M\times D}$, M is the number of tokens. Next, the query, key, and values are partitioned among H heads; for each head, h, we have $Q_h \in \mathbb{R}^{N\times d}$ and $K_h, V_h \in \mathbb{R}^{M\times d}$, where $d = \frac{D}{H}$. The notion of *multi-head* attention facilitates parallel attention calculation in all H heads. For any head, the *scaled dot-product*, $\frac{Q_h K_h^\top}{\sqrt{d}} \in \mathbb{R}^{N\times M}$, is further processed through nonlinear softmax activation function acting row-wise, followed by projection on the row space of the value matrix V, such that $\mathcal{A}_h = \boldsymbol{\sigma}\left(\frac{Q_h K_h^\top}{\sqrt{d}}\right) V_h \in \mathbb{R}^{N\times d}$. Self-attention is a special case where $M = N$. Once attention is calculated for all the heads, they are concatenated back to $\mathbb{R}^{N\times D}$.

The matrix, $\frac{Q_h K_h^\top}{\sqrt{d}}$, for a head inside an encoder, refers to how different image tokens *attend* to each other. Self-attention models interactions among tokens of the same sequence, while cross-attention models how tokens from the query sequence relate to tokens in the source sequence. The softmax activation, $\boldsymbol{\sigma}$, acts on each row, normalizes the key-query interactions by projecting them onto a probability simplex, and models the token-to-token relationships; see Figure 3(i).

Studies That Improve Upon Attention. Numerous works improve attention mechanisms in NLP and vision to enhance feature representation. DeiT [56] uses token-based distillation, CaiT [57] separates class-attention in deeper layers, DaViT [17] adds channel-wise attention, and Swin Transformer [37] employs

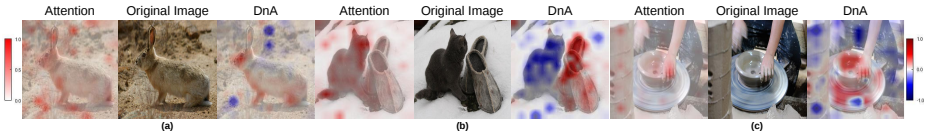

Fig. 2: Original images **(center)** from ImageNet-1K, their attention activation maps for traditional softmax *only* showing positive interactions **(left)**, and our denoising attention, DnA ($\mathcal{A}_h^{Q\pm V\pm}$) **(right)** showing *both* positive and negative interactions. **(a)** The softmax attention is dispersed and does not focus on the concerned object (hare), while our DnA focuses on the concerned object. **(b)** The softmax attention interacts heavily with a secondary adversarial class (cat), while our DnA focuses on the concerned object (clog) and considers the adversarial object as misleading. **(c)** The softmax attention does not attend to the primary object (potter's wheel); only a few sporadic interactions are picked up from the background. In contrast, DnA focuses on the primary object.

hierarchical shifted-window attention for scalability. ConViT [19] introduces convolutional inductive bias through gated positional self-attention, while SimA [29] replaces softmax with ℓ_1 normalization while maintaining competitive ViT performance. See broader discussion in Supplementary §A.

2.2 Attention Mechanisms in Video Understanding

Attention mechanisms are crucial to current video understanding. We group the related work into two categories: traditional video transformers and vision-language adapters for video LLMs.

Transformers for video understanding. ViViT [1] adapts a ViT to video by extracting the spatiotemporal tokens from videos and proposes several different transformer architectures that decouple the spatial and temporal dimensions. Video Swin Transformer [38] implements a 3D shifted window-based multi-head self-attention to the spatiotemporal space, reducing the computation by applying attention within local 3D windows. TimeSformer [3] proposes a divided space-time attention mechanism that decouples the attention operations.

Attention-based adapters in video LLMs. Q-Former [35] and Perceiver Resampler [26] use learnable queries that cross-attend to frozen visual encoder features, distilling visual context into a fixed number of tokens. Q-Former is used in Video-LLaMA [66], and Perceiver Resampler [26] has been successfully integrated as a bridge between visual and text contexts for LLMs [33, 60, 69]. mPLUG-Owl3 [61] introduces hyper-attention, which parallelizes cross- and self-attention by reusing language queries to attend to visual features and adaptively fuse them into the text stream. VisCoP [49] enables domain adaptation without fine-tuning the frozen vision encoder by using visual probes that cross-attend to intermediate features and extract domain-specific information.

3 Model Architecture

This section introduces **De**noising **A**ttention or **DnA**. We first analyze the softmax function, the core normalization mechanism in self-attention, and examine its properties from an information-theoretic perspective, highlighting potential limitations in representing negative token interactions that may be informative for visual understanding. Our focus is restricted to *self-attention*; other components of the ViT architecture, including positional embedding, class token representation, pre- and post-normalization, MLP, *etc*., remain untouched.

3.1 Why Softmax for *Attention*?

For a vector, $z \in \mathbb{R}^n$ with components $(z_1, \cdots, z_n)$, the softmax function is defined component-wise, for $i \in [n]$, as $\boldsymbol{\sigma}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}$, which is a unique solution to several constrained optimization problems. One of the most fundamental characterizations of the softmax function is the *maximum entropy* property given in the Theorem below [27].

Theorem 1 (*Maximum Entropy*). *Let $z = (z_1, \cdots, z_n)$ be a given vector of scores and let Z be a random variable taking values $z_1, ..., z_n$. The probability distribution $p^* = (p_1^*, \ldots, p_n^*)$ that maximizes the Shannon entropy, $H(p) = -\sum_{i=1}^n p_i \log p_i$, subject to the constraints that p is on the probability simplex, $\mathcal{P} = \{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\}$, and gives a fixed expected value $\mathbb{E}_{\mathbf{p}}[Z] = \mu$, for some $\mu \in \mathbb{R}$, is given by $p_i^* = \frac{e^{\lambda z_i}}{\sum_{j=1}^n e^{\lambda z_j}}$, where the parameter $\lambda \in \mathbb{R}$ is chosen such that $\sum_{i=1}^n p_i^* z_i = \mu$.*

The standard softmax is recovered when μ is the empirical average of the features, that is, $\mu = \sum_{i=1}^n z_i e^{z_i} / \sum_{j=1}^n e^{z_j}$.

The softmax distribution maximizes entropy subject to expected-value constraints, yielding the most noncommittal probability distribution consistent with the available information. This principle underlies its central role in information theory and statistical inference [14, 27]. In attention mechanisms, softmax is therefore naturally justified as a principled method for transforming raw attention scores into a valid probability distribution [2,58]; see Theorem 4 in §B. This mechanism computes a weighted average, where weights encode the relative importance of each position. The exponential form of softmax induces nonlinear amplification of dominant scores, producing two effects: (i) **Selective focus,** where large scores receive near-unit weight while small or negative scores are suppressed, emphasizing the most relevant information; and (ii) **Competitive normalization,** where increasing attention to one position necessarily decreases attention to others, enforcing a fixed attention budget.

Why do we need a *softmin* function? Because softmax pushes dominant scores toward 1 while compressing smaller scores toward 0, potentially informative low interactions may be suppressed. To capture such signals, we consider the softmin function, defined component-wise for $\mathbf{z} \in \mathbb{R}^n$ as $\hat{\boldsymbol{\sigma}}(\mathbf{z})_i :=$

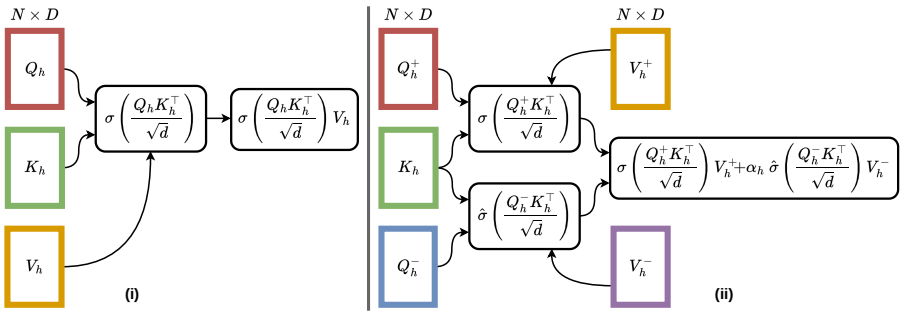

Fig. 3: A schematic comparison between traditional attention and the proposed denoising attention. *(i)* The multi-head attention uses a softmax activation on the rows of the scaled dot-product $\frac{Q_h K_h^\top}{\sqrt{d}}$ and combines it with V_h. *(ii)* The proposed denoising attention $\mathcal{A}_h^{Q\pm V\pm}$, uses two sets of values, V_h^+ and V_h^-, and two sets of queries Q_h^+ and Q_h^-. We use *softmax* and *softmin* two scaled dot-products, $\frac{Q_h^+ K_h^\top}{\sqrt{d}}$ and $\frac{Q_h^- K_h^\top}{\sqrt{d}}$, and are projected on the row-spaces of V_h^+ and V_h^-, respectively.

$\boldsymbol{\sigma}(-\mathbf{z})_i = \frac{e^{-z_i}}{\sum_{j=1}^n e^{-z_j}}$. By Theorem 1, $\hat{\boldsymbol{\sigma}}$ is the unique distribution that maximizes Shannon entropy subject to the constraint $\sum_{j=1}^n \hat{p}_j(-z_j) = \mu$, equivalently $\sum_{j=1}^n \hat{p}_j z_j = -\mu$. Thus, softmin maximizes entropy under a mean constraint with the opposite sign, providing a principled counterpart to softmax that emphasizes strongly lower interactions. Thus, it makes sense to promote both the largest and the smallest components, since using only one provides incomplete information.

To conclude, we obtain a new principle: *When we do not have any information about which extreme value (largest or smallest) of the scores, we should use both softmax and softmin.*

3.2 Designing Denoising Attention (DnA)

In ViTs [18,56], self-attention models token interactions by aggregating features via normalized weights derived from the attention matrix $\mathcal{A}_h$ (see Figure 2). High interaction scores reflect strong mutual relevance and typically correspond to object-related features; see Figure 2(a)-(b). However, as illustrated in Figure 2(c), softmax attention may assign weight to irrelevant context, introducing noise or distraction, which we interpret as negative token interactions. Formally, softmax projects each row of the scaled dot-product $\frac{Q_h K_h^\top}{\sqrt{d}}$ onto the probability simplex, exponentially amplifying large positive scores while mapping low, negative, and highly negative scores near zero. Although weak interactions may be uninformative, strongly negative interactions can encode informative contrastive structure. Standard softmax suppresses such signals; we instead aim to explicitly model and leverage negative token interactions to enhance visual understanding.

Recently, differential attention [62] for language modeling is motivated by differential amplifiers, where subtracting two signals suppresses shared noise. The

model projects the input X, to two sets of queries, $Q_1, Q_2 \in \mathbb{R}^{N\times d/2}$ and keys, $K_1, K_2 \in \mathbb{R}^{N\times d/2}$ and one value matrix $V \in \mathbb{R}^{N\times d}$, and the attention with a learnable parameter λ, is computed as $\mathcal{A}_D(X) = (\boldsymbol{\sigma}(\frac{Q_1K_1^T}{\sqrt{d}}) - \lambda\boldsymbol{\sigma}(\frac{Q_2K_2^T}{\sqrt{d}}))V$. By construction, $\mathcal{A}_D(X)$ lies in the row space, $C(V^\top)$ of V, meaning both interactions are projected onto the same subspace. This raises the question of whether improved modeling can be achieved by separating these projections.

To motivate this, we connect the performance of a subspace classifier and the principal angles between subspaces. Consider a two-class problem with labels, C_1 and C_2, concentrated near two subspaces with orthonormal bases, $U_1, U_2 \in \mathbb{R}^{n\times d}$. Let the class-conditional densities be modeled as:

$$p(x|C_1) = \int p(x|\alpha, C_1)p(\alpha)d\alpha, \text{ and } p(x|C_2) = \int p(x|\alpha, C_2)q(\alpha)d\alpha,$$

where the vector α does not essentially follow a multivariate normal distribution. Let $P(C_2|C_1)$ be the probability of mistaking C_2 for C_1, and $P(C_1|C_2)$ be the opposite. Then the *classification error*, $P_e = \frac{1}{2}\left(P(C_2|C_1) + P(C_1|C_2)\right)$. Following Bayes rule, one can rewrite $P(C_2|C_1)$ as

$$P(C_2|C_1) = \int P(C_2|C_1, \alpha)p(\alpha)d\alpha,$$

where $P(C_2|C_1, \alpha)$ can be bounded by writing $x = U_1\alpha + \mathbf{n}$, and considering $\mathbf{n} \in \mathcal{N}(0, \sigma^2\mathbf{I})$. Similarly, $P(C_1|C_2)$ can be bounded. With the formalization above, the theorem below bounds the classification error:

Theorem 2 (Classification error bound). *[25] As $\sigma^2 \to 0$ the classification error, P_e is upper bounded as $P_e \leq \int \frac{1}{2}\exp\left(-\frac{(\sum_{i=1}^d \sin^2\theta_i\alpha_i^2)^2}{8\sigma^2\sum_{i=1}^d \sin^2\theta_i(\alpha_i^2+\sigma^2)}\right)\frac{p(\alpha)+q(\alpha)}{2}d\alpha$, where $\{\cos(\theta_i)\}_{i=1}^d$ are the singular values of $U_1^\top U_2$ in ascending order, and $\{\theta_i\}_{i=1}^d$ are the principal angles between U_1 and U_2.*

Theorem 2 indicates that when the principal angles (or signal energy) are bigger, the classification error upper bound is smaller. When $\sigma^2 \to 0$, we have $\exp\left(-\frac{(\sum_{i=1}^d \sin^2\theta_i\alpha_i^2)^2}{8\sigma^2\sum_{i=1}^d \sin^2\theta_i(\alpha_i^2+\sigma^2)}\right) \approx \exp\left(-\frac{\sum_{i=1}^d \sin^2\theta_i\alpha_i^2}{8\sigma^2}\right)$, by ignoring the higher-order term of σ^2. This indicates that classification performance is a function of principal angles and *larger principal angles lead to a smaller classification error.*

Consequently, we can formulate the positive and negative interactions between the tokens as a two-class classification problem. We introduce two queries over the same keys: (i) a *positive* query capturing features belonging to the correct class, and (ii) a *negative* query capturing closely associated but irrelevant features. Their outputs are projected into distinct subspaces with large principal angles between them. Combining these projections suppresses non-essential patches (denoising) while maintaining small classification error through subspace separation.

Formally, the solution to the constrained minimization problem with a balancing parameter $\beta > 0$:

$$\min_{\theta\in\Theta} \mathcal{L}(\theta) + \beta\sum_{i=1}^{L}\sum_{j=1,j\neq l}^{H} \langle V_{ij}(\theta), V_{il}(\theta)\rangle_F, \tag{1}$$

promotes orthogonality between value subspaces and yields a solution set contained in that of the unconstrained problem:

$$\min_{\theta \in \Theta} \mathcal{L}(\theta), \tag{2}$$

where $\mathcal{L}(\theta)$ is the training loss. Nevertheless, to obtain solutions to (1) one needs to guarantee $\sum_{i=1}^{L} \sum_{j=1}^{H} \langle V_{ij}, V_{ik} \rangle_F$ is small at every iterate. Enforcing orthogonality at every iterate is difficult. If we can argue that $\sum_{i=1}^{L} \sum_{j=1, j \neq l}^{H} \langle V_{ij}^{(k)}, V_{il}^{(k)} \rangle_F$ is *quasi-orthogonal* for all k then it suffices to only solve (2).

In practice, $(W_{V_{ij}})_{pq} \sim \mathcal{N}(0, \sigma^2)$, chosen from a zero mean Gaussian distribution with standard deviation $\sigma = 0.02$. Since $Support(\mathcal{N}(0, \sigma^2)) = \mathbb{R}$, for all $\sigma \in \mathbb{R}^+$, we consider a compact support. Additionally, we follow [67], which views convergence in neural network training from an invariant measure perspective. Especially, Theorem 4.3 in [67] shows that as the training loss $\mathcal{L}(\theta)$ stabilizes, the iterates (or the weights) lie in a compact region almost surely. We modify the result of Theorem 4.3 below.

Proposition 1. *Let $W_{V_{ij}(0)}$ be initialized within a compact set $\mathcal{C}_V := \{W_{V_{ij}(0)} : \|W_{V_{ij}(0)}\| \leq \gamma\}$. Then for every k, the iterate $W_{V_{ij}(k)} \in \mathcal{C}_V$ with high probability.*

Proposition 1 implies that all iterates lie in a compact region almost surely. Based on this, we proposed the following Theorem, which states that, under certain standard assumptions, the solutions to (2) are orthogonal with high probability.

Theorem 3 (Quasi-orthogonality of the subspaces). *Let the entries of W_{V^+} and W_{V^-} be initialized as i.i.d. $\mathcal{N}(0, \sigma^2)$ over a finite support $[-\gamma, \gamma]$. Let the input $X_p \in \mathbb{R}^{N \times d}$ be such that $\|X_p\|_\infty \leq M$. Assume for every k, the entries of the iterates $\{W_{V^+(k)}, W_{V^-(k)}\}$ are i.i.d. sub Gaussian with parameter $\sigma > 0$. (*i*) Then $|(W_{V^+(k)})_{pq}| \leq \gamma$ and $|(W_{V^-(k)})_{pq}| \leq \gamma$, with high probability. (*ii*) If $\gamma = O(M^{-1}\epsilon^{-\frac{1}{2}}(N \log(\delta^{-1}))^{-\frac{1}{4}})$, then $|\langle V^+, V^- \rangle_F| \leq d\epsilon$ with probability $1 - \delta$.*

DnA: Modeling Discriminative Patches. While differential attention may remove *common-mode noise, it could also cancel useful signals if $Q_1 K_1^\top$ and $Q_2 K_2^\top$ both assign high attention weights to the same tokens.* Since both branches share a single V, subtracting them directly removes overlapping contributions. To avoid this, we introduce two distinct value projections, $V^+ = X_p W_{V^+} \in \mathbb{R}^{N \times D}$ and $V^- = X_p W_{V^-} \in \mathbb{R}^{N \times D}$, where $W_{V^+}, W_{V^-} \in \mathbb{R}^{D \times D}$ are learnable weight matrices. Consequently, we partition V^+ and V^- into H heads such that, for any head h, we head-specific value subspaces, $V_h^+, V_h^- \in \mathbb{R}^{N \times d}$. With this formalization, our denoising attention takes the form:

$$\mathcal{A}_h^{Q^\pm V^\pm} = \boldsymbol{\sigma}\left(\frac{Q_h^+ K_h^\top}{\sqrt{d}}\right) V_h^+ + \alpha_h \hat{\boldsymbol{\sigma}}\left(\frac{Q_h^- K_h^\top}{\sqrt{d}}\right) V_h^-,$$

where $\alpha_h \in \mathbb{R}$ is a learnable balancing parameter; see Figure 3(ii). It contains all previously discussed designing choices: (*i*) Two *q*uasi-orthogonal value spaces; (*ii*) that are scaled by using softmax for positive query and softmin for negative query, to capture complementary signal components, analogous to denoising or band-pass filtering [13, 32]; see Algorithm 1 in §C.1. For ablations on different design choices of DnA and their performances on ImageNet-1K [16], see §C.5.

3.3 Adapting DnA for Video Understanding

In this section, we explore how DnA can be integrated into video understanding frameworks, including conventional transformer-based architectures (e.g., TimeSformer [3]) and video foundation models [65]. This adaptation enables us to evaluate DnA on fine-grained video analysis tasks, where such an attention mechanism can be particularly beneficial by suppressing detrimental token interactions and improving representation quality.

Video Transformer [3] uses a dissociated space-time attention within each transformer block to capture discriminative action patterns. We integrate DnA into this framework by replacing both the spatial and temporal self-attention mechanisms with our DnA variant. This allows the mechanism to operate independently along both dimensions, filtering irrelevant spatial cues within each frame and temporal noise across frames.

Video LLMs involve a vision-language adapter, often consisting of simple MLPs or linear layers. Recently, visual probing [49] emerged as an additional visual reinforcement for the LLM. The role of visual probes in VisCoP [49] is to provide a compact set of learnable tokens $\mathbf{P}$ that extract domain-specific spatio-temporal cues $\mathbf{X}$ from the hidden states of the vision encoder. These probes augment the frozen vision encoder embeddings before they are passed to the LLM. At each layer l, the visual probes are updated through a cross-attention operation between learnable queries and the visual embeddings. We adapt DnA to video LLMs by replacing the cross-attention modules in VisCoP with our proposed attention mechanism. To the best of our knowledge, this is the first integration of explicit negative token interaction modeling within a foundation model through an adapter-based design, allowing the model to retain the benefits of large-scale pretraining while leveraging the denoising capability of our attention mechanism. Specifically, the denoising interaction module at layer ℓ introduces projection matrices $W_{Q^+}, W_{Q^-}, W_K, W_{V^+}, W_{V^-}$, such that the probe update becomes:

$$\mathbf{P}^{\ell+1} = \boldsymbol{\sigma}\left(\frac{\mathbf{P}^{\ell} W_{Q^+}^{\ell} (\mathbf{X}^{\ell} W_K^{\ell})^{\top}}{\sqrt{d_v}}\right)(\mathbf{X}^{\ell} W_{V^+}^{\ell}) + \alpha\, \hat{\boldsymbol{\sigma}}\left(\frac{\mathbf{P}^{\ell} W_{Q^-}^{\ell} (\mathbf{X}^{\ell} W_K^{\ell})^{\top}}{\sqrt{d_v}}\right)(\mathbf{X}^{\ell} W_{V^-}^{\ell}).$$

This *cross-DnA* formulation explicitly models interactions between positive and negative tokens, enabling the probes to capture fine-grained visual cues within the LLM embedding space.

4 Benchmarking and Evaluation

In this section, we benchmark DnA in traditional transformers and a foundation model for diverse visual understanding tasks, and analyze the effectiveness of DnA quantitatively and qualitatively.

4.1 Denoising Attention for Image Classification

Setup. For Image Classification, we train our models and baselines from scratch on ImageNet-1K [16] with images resized to 224×224. We implement DnA on

Model	Validation Set		Test Set		Params	GFLOPs	Throughput
	Acc.@1	Acc.@5	Acc.@1	Acc.@5			
ViT-B	81.1	95.6	81.1	95.6	86.6M	17.6	909.1 img/s
+Diff. Attn.	81.4(↑0.3)	95.7(↑0.1)	81.5(↑0.4)	95.6	86.6M	17.6	909.1 img/s
+Cog Attn.	81.4(↑0.3)	95.7(↑0.1)	81.5(↑0.4)	95.7(↑0.1)	86.6M	17.6	909.1 img/s
+DnA	**81.9**(↑0.8)	**95.9**(↑0.3)	**81.9**(↑0.8)	**95.8**(↑0.2)	100.7M	21.1	909.1 img/s

Table 1: Accuracy on the ImageNet-1K [16] validation and test set; the ↑ arrows indicate the absolute gain compared to the baseline ViT-B. In the supplementary material, we present: variance across multiple runs (Table 6); implementation details and hyperparameters (§C.1), with computational requirements (Table 7); and training loss and test accuracy curves (§C.2).

Model	ImageNet-A [24]		
	Acc.↑	RMSE↓	AURRA↑
ViT-B	24.1	26.6	33.5
+Diff. Attn.	**28.1**	25.8	39.7
+Cog Attn.	27.5	25.3	38.8
+DnA	27.7	**24.2**	**40.7**

Table 2: Robustness evaluation on ImageNet-A [24]. Improvements in ranking and calibration metrics reflect decision quality on adversarial samples.

the ViT-B backbone, using hyperparameters from DeiT [56]. Our models contain 12 encoder layers, with 12 heads in each layer, a patch size of $p = 16$, and a head dimension of 64. We train them on an NVIDIA H100 GPU for 300 epochs; see implementation details and training hyperparameters in Table 4 in §C.1.

Baselines. Our primary baseline is ViT-B, trained with a DeiT recipe [56]. For the image classification on ImageNet-1K, we employ cog attention [40] and differential attention [62] in ViT-B by replacing traditional MHSA. The rest of the ViT architecture remains unchanged. Cog attention [40] employs negative attention weights, thereby providing a direct point of comparison with our approach. Differential attention [62] employs a signal denoising mechanism for language modeling tasks by splitting keys and queries. See implementation details in §C.1.

Evaluation. Table 1 shows the evaluation results of our models on the ImageNet-1K [16] validation and test set compared to the baselines. The proposed DnA surpasses the baseline ViT-B by 0.8% and outperforms models using cog attention and differential attention by 0.5%, *while maintaining the same throughput during inference.*

Robustness Evaluation on ImageNet-A. ImageNet-A [24] consists of images that are full of adversarial examples and that confuse the image classifiers. We use this dataset to evaluate the robustness of our model to such adversarial examples. For evaluation, we use the calibration error metric RMSCE, the reliability metric AURRA, and standard accuracy. We show these results in Table 2. In terms of reliability and calibration, our proposed DnA substantially outperforms the baselines and competitors.

Additional Benchmarking. We use the weights of our ImageNet-1K trained models for downstream visual understanding tasks: (*i*) *image classification* on CIFAR-10, CIFAR-100 [31], and Stanford Cars [30], (*ii*) *object detection and instance segmentation* on MS COCO [36]. We further conduct (*iii*) *robustness evaluations* on ImageNet-O, -R [23,24]; see §C.3 and Tables 8–10. We incorporate DnA for a few additional backbones and show their performances in Table 11.

Model	Toyota Smarthome [15] CS	CV2	NTU60 [51] CS	CV
TimeSformer	67.5	59.5	81.2	88.6
+Diff. Attn.	66.6	59.4	81.5	89.1
+DnA	**68.8**	**63.5**	**82.4**	**89.2**

Model	Act.	Task	HOI	Hand	Avg.
VideoLLaMA3	74.9	75.8	75.2	65.3	72.8
VisCoP	81.8	86.1	**79.3**	65.1	78.1
+Diff. Attn.	78.8	83.0	77.5	64.9	76.1
+DnA	**83.1**	**87.1**	79.2	**65.1**	**78.6**

Table 3: (a)Left: Baseline comparison on Toyota Smarthome and NTU60. We report mean-class accuracy (mCA) for Toyota Smarthome and Top-1 accuracy for NTU60. Cross-subject (CS) and cross-view (CV) denote the data split protocols. **(b) Right:** Quantitative results on Ego-in-Exo PerceptionMCQ [48]. We evaluate across four egocentric video understanding tasks: action understanding (Act.), task-relevant region understanding (Task), human-object interactions (HOI), and hand identification (Hand).

4.2 Denoising Attention for Video Understanding

We evaluate DnA on fine-grained video understanding tasks to demonstrate the importance of specialized attention mechanisms for detailed video analysis. Specifically, we integrate DnA in two settings. First, we incorporate DnA into a conventional *Video Transformer* [3] for action classification on exocentric videos. Second, we integrate DnA into a *video LLM* [65] to address egocentric video understanding tasks.

Video Transformer [3]. We initialize the weights from the trained models in §4.1, and pretrain the models on Kinetics400 [7], then finetune and evaluate them on Toyota Smarthome [15] and NTU RGB+D 60 [51]. Toyota Smarthome consists of 16K videos spanning 31 action categories. NTU RGB+D 60 contains 57K video samples across 60 action classes. Models are trained for 15 epochs using divided space-time attention, with an input of 8 frames at a resolution of 224×224. For Toyota Smarthome, we evaluate using the cross-subject (CS) and cross-view (CV2) protocols, reporting mean class accuracy (mCA). For NTU60, we report Top-1 accuracy.

Results. Table 3(a) shows that replacing both the space and temporal attention in TimeSformer with our denoising mechanism consistently improves performance across both datasets. On Toyota Smarthome [15], we observe gains of $\uparrow$1.3% on CS and $\uparrow$4.0% on CV2. On NTU60 [51], denoising attention achieves improvements of $\uparrow$1.2% on CS and $\uparrow$0.6% on CV.

Overall, these consistent gains across two datasets show the effectiveness of incorporating denoising attention in video transformers.

Video LLM. Following the training strategy of VisCoP [49] and by using a subset of ego videos from EgoExo4D [20] ($\sim$46K QA pairs), we train VideoLLaMA3 with DnA as the attention mechanism and evaluate on the Ego-in-Exo PerceptionMCQ [48] benchmark. This benchmark comprises four evaluation categories: *Action Understanding*, which measures recognition of performed activities; *Task Regions*, which assesses identification of spatially salient areas in a scene; *Human–Object Interaction* (HOI), which evaluates reasoning about interactions between individuals and objects; and *Hand Identification*, which tests first-person perspective understanding through hand tracking and recognition.

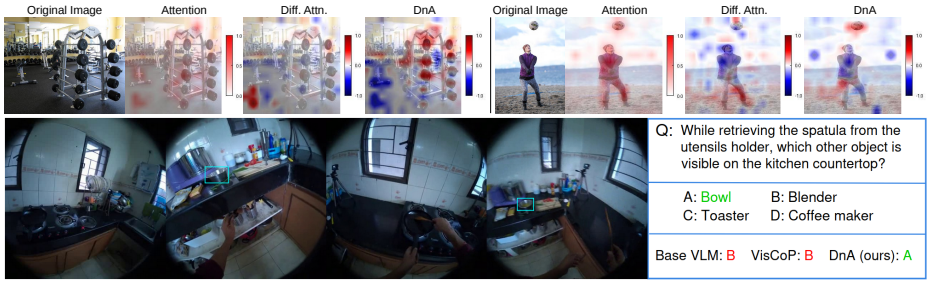

Fig. 4: (a) Top: Attention visualization for ViT-B [18] with softmax, differential attention [62], and our DnA. The objects are dumbbells and a volleyball. **(b) Bottom:** Example of DnA outperforming baseline methods on egocentric video QA. Four frames are uniformly sampled; the object of interest is highlighted by a cyan box. The correct answer is in **green**, the incorrect in **red**.

We initialize the positive projection weights from the pretrained vision encoder's cross-attention weights, and initialize $W^{\ell}_{Q^-}$ and $W^{\ell}_{V^-}$ as scaled-down copies of their positive counterparts (by 10^{-4}). The scaling factor avoids branch symmetry, which would arise from two equivalent initialization points [10, 11]; see §C.5 and Table 13. During finetuning, we train the interaction modules, visual probe projector, vision embedding projector, and apply LoRA to the LLM. We train for 3 epochs with an initial learning rate of 10^{-5}, using a cosine schedule for the projectors and the LLM. Evaluation uses a temperature of 0.

Results. Table 3(b) displays our results on Ego-in-Exo PerceptionMCQ [48]. Our denoising variant improves over the baseline VisCoP on average, achieving gains of $\uparrow$0.5%, showing that the denoising mechanism helps the visual probes extract more semantically meaningful domain-specific cues.

4.3 Qualitative Analysis: Attention Visualization

We use the gradient-based attention interpretability method from [9] for visualization on ViT-B. For DnA and differential attention [62], we extend this method to propagate and visualize signed relevance independently. Figure 4(a) demonstrates that softmax attention is densely spread around the image, but with limited focus on the classes of interest. Differential attention, on the other hand, shows improvement but retains substantial background noise. Finally, the positive activations (red) in DnA are focused on the classes of interest (dumbbell and volleyball). Crucially, the $\mathcal{A}_h^{Q^-V^-}$ branch of DnA suppresses irrelevant background areas, creating a contrast between class features and noisy distractors. Figure 4(b) displays how DnA thrives despite the presence of many visually prominent objects, by suppressing query-irrelevant tokens before passing them to the LLM, whereas VisCoP and the Base VLM are mislead by irrelevant visual cues. We provide additional visualizations in §C.5; see Figures 12 and 13.

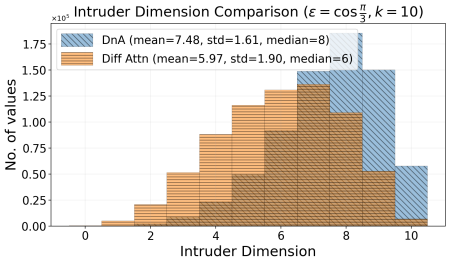

Fig. 5: Comparison of intruder dimension counts between DnA and differential attention $\epsilon = \cos\frac{\pi}{3}$, k=10.

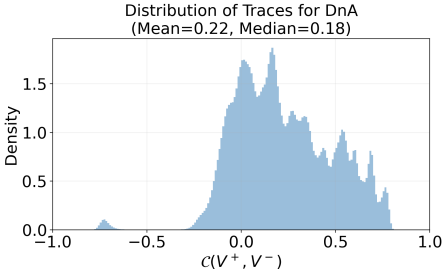

Fig. 6: Distribution of $\mathcal{C}(V^+, V^-)$ for each sample, head, and layer in the ImageNet-1K validation set (7.2M trace values).

4.4 Quantitative Analyses: Understanding the Attention Subspaces

This section empirically justifies our postulates in §3.2 that lead to the design choices of DnA. First, by calculating the intruder dimensions between two attention branches of the differential transformer and DnA, we show the necessity of considering two value subspaces. Next, by empirically measuring the similarities between the contrasting value subspaces, V^+ and V^- of DnA, we verify the claim of our main theoretical result in Theorem 3.

(a) **Intruder Dimension Analysis.** To observe how two branches in DnA relate to one another, we perform a subspace analysis that measures the divergence between $\mathcal{A}_h^{Q^+V^+}$ and $\mathcal{A}_h^{Q^-V^-}$ in DnA and $\mathcal{A}_1 V$ and $\mathcal{A}_2 V$ of the differential transformer. We compute the SVD of each branch and count how many of the Top-k left singular vectors are *near-orthogonal*. The near orthogonality of these vectors indicates the separability of the subspaces and hence the efficacy of denoising attention in reducing the classification error; see Theorem 2. Out of the Top-k singular vectors, we called the vectors with larger angles between them *intruders*, following [5, 53]; the higher the intruders, the better.

We calculated the intruders independently per sample, head, and layer across 5 images per class from the ImageNet-1K validation set (5K images total), resulting in 720K values for ViT-B. Out of the Top-k left singular vectors, a singular vector is considered an intruder if its cosine similarity with the rest of the singular vectors falls below a user-defined threshold, ϵ.

Figure 5 shows DnA consistently displays higher intruder dimensions, with its distribution peaking around $8-9$, compared to differential attention, which peaks around $6-7$ (we set $\epsilon = \cos\frac{\pi}{3}, k = 10$). This shows that in differential attention, two attention branches operate within the same subspace, performing a local refinement that may cancel common-mode noise, where DnA's branches interact more orthogonally. Since the subspaces are near-orthogonal, DnA can suppress noise without canceling any useful signal and has lower classification error, as seen in the visual understanding tasks. On the other hand, branches

$\mathcal{A}_1$ and $\mathcal{A}_2$ in differential attention may both assign high attention weights to the same tokens, losing a useful signal in the subtraction.

(b) **Similarity between V^+ and V^-.** We consider the normalized Frobenius inner product, $\mathcal{C}(V^+, V^-)$ to calculate the similarities between two value subspaces, given by $\mathcal{C}(V^+, V^-) := \langle \frac{V^+}{\|V^+\|_F}, \frac{V^-}{\|V^+\|_F} \rangle_F = \text{Trace}\left(\frac{{V^+}^\top V^-}{\|V^+\|_F \|V^-\|_F} \right)$. We normalize the matrices so that the range is $[-1, 1]$, where 1 indicates that the two matrices are aligned, and 0 indicates orthogonality. We plot $\mathcal{C}(V^+, V^-)$ for each sample, head, and layer using the ImageNet-1K validation set, resulting in 7.2M trace values; see Figure 6. The mean of the distribution is 0.22, while the median is 0.18. Figure 6 shows that the subspaces V^+ and V^- are quasi-orthogonal for most heads, with most of the mass near 0. However, the distribution shows a heavy right tail, suggesting that some attention heads in V^+ and V^- may not contribute equally to the denoising mechanism. *This empirical observation justifies our theoretical result in Theorem* 3. It also supports the claim that, instead of enforcing orthogonality at every iteration (as in the case of (1)), it is sufficient to solve the unconstrained problem (2), as the resulting subspaces are quasi-orthogonal during training.

Additionally, we compute the average cosine similarity between $\mathcal{A}_h^{Q^+V^+}$ and $\mathcal{A}_h^{Q^-V^-}$ for DnA and $\mathcal{A}_1 V$ and $\mathcal{A}_2 V$ for differential attention, over the entire ImageNet-1K validation set. Differential attention achieves a high average cosine similarity of 0.96, while DnA measures at only 0.32, further solidifying our claim.

4.5 Additional Quantitative Analyses & Ablations

We further examine DnA from multiple perspectives through extensive ablations.

Representational Robustness. We measure representational robustness by computing $\Delta_{ij} := 1 - \cos(\angle p_i, p_j)$ between two tokens, $p_i, p_j \in \mathcal{A}$, at each layer. Before calculating Δ_{ij}, token embeddings (excluding the CLS token) are mean-centered across feature space, then averaged per image and across the dataset. The higher this measurement, the better, as it indicates greater inter-token separation. To further substantiate our claim that our attention mechanism denoises, we infuse the input with Gaussian noise. Figure 7 shows how DnA stabilizes noisy inputs through layer-wise features. Overall, DnA improves robustness by suppressing perturbations and transforming them into features that downstream layers can process effectively. In contrast, softmax attention propagates perturbations, amplifying distortions in feature space and degrading accuracy. This denoising behavior lets DnA recover from early noise more reliably.

Additional Analysis. In §C.5, we trained two parameter-matched ViT-B baselines, and their performance confirms that DnA's gains stem from the architectural choices rather than extra parameters. Additionally, we discuss different design components of DnA and their behavior in §C.5. We further analyze attention head diversity, the evolution of the learnable parameter (α_h), and the entropy of attention distributions across encoder layers in Figure 14, §C.5.

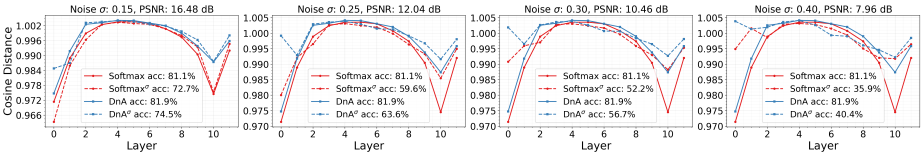

Fig. 7: Average pairwise cosine distances under varying noise levels; see noise variance (σ) and PSNR in titles. Solid lines represent models evaluated under clean data (DnA and softmax), while dashed lines reflect artifacts of noisy data (DnA$^\sigma$ and softmax$^\sigma$) for different noise, σ. DnA$^\sigma$ exhibits smaller accuracy drops than softmax$^\sigma$ across all noise levels, indicating more diverse feature representations and robustness to noise.

5 Limitation & Future Work

DnA introduces additional parameters, ranging from approximately 3% in video LLMs to 23% in video transformers. While the performance improvements stem from the discriminative capability of the proposed attention mechanism rather than simply increased parameter capacity (see Table 12), we acknowledge the added parameters and associated training cost. In practice, all the additional parameters reside in parallelizable attention heads, which mitigate the computational overhead. Consequently, we observe throughput comparable to a baseline ViT when compared with ViT augmented with DnA (see Table 7). Future work will focus on developing more parameter-efficient implementations while preserving the performance gains of the proposed mechanism.

6 Conclusion

We propose DnA, denoising attention that uses softmin and softmax and explicitly models negative token interactions. Our design principle for DnA is motivated by theoretical observations that indicate that subspace separation for the positive and negative token interactions can alleviate both attention misallocation and noise. By replacing traditional attention, the proposed DnA outperforms standard ViT-B on ImageNet-1K classification, improves on several video understanding tasks with both video transformers and video LLMs, and generalizes to multiple downstream tasks. To justify the design choices, we verify that the two branches in DnA, operating over distinct interacting subspaces, learn different token interactions that jointly enable the denoising effect.

Acknowledgments. Aritra Dutta is partially supported by the Florida Department of Health Grant, AWD00007072, and the National Science Foundation Grant, 2321986. Subhajit Maity is supported by the Florida Department of Health Grant AWD00007072. Srijan Das is partially supported by the National Science Foundation Grant, IIS-2245652. We thank Dominick Reilly for insightful discussions and technical assistance.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., Schmid, C.: ViViT: A Video Vision Transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: International Conference on Learning Representations (2015)
3. Bertasius, G., Wang, H., Torresani, L.: Is Space-Time Attention All You Need for Video Understanding? In: International Conference on Machine Learning. pp. 833–842 (2021)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems. pp. 1877–1901 (2020)
5. Cadenhead, E., McGee, C., Li, X., Bergou, E.H., Dutta, A.: Beyond LoRA: Is Sparsity-Induced Adaptation Better? arXiv preprint arXiv:2606.13767 (2026)
6. Campos, R., Vayani, A., Kulkarni, P.P., Gupta, R., Zafar, A., Dutta, A., Shah, M.: GAEA: A Geolocation Aware Conversational Model. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5236–5246 (2026)
7. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
8. Chamberlin, D.D., Boyce, R.F.: SEQUEL: A Structured English Query Language. In: Proceedings of the ACM SIGFIDET (Now SIGMOD) Workshop on Data Description, Access and Control. pp. 249–264 (1974)
9. Chefer, H., Gur, S., Wolf, L.: Transformer Interpretability Beyond Attention Visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 782–791 (2021)
10. Chen, C., Yin, Y., Shang, L., Jiang, X., Qin, Y., Wang, F., Wang, Z., Chen, X., Liu, Z., Liu, Q.: bert2BERT: Towards Reusable Pretrained Language Models. In: Annual Meeting of the Association for Computational Linguistics. pp. 2134–2148 (2022)
11. Chen, T., Goodfellow, I., Shlens, J.: Net2net: Accelerating Learning via Knowledge Transfer. In: International Conference on Learning Representations (2016)
12. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A.V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Ruiz, C.R., Steiner, A.P., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: PaLI: A Jointly-Scaled Multilingual Language-Image Model. In: International Conference on Learning Representations (2023)
13. Christiano, L.J., Fitzgerald, T.J.: The Band Pass Filter. International Economic Review **44**(2), 435–465 (2003)
14. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, 2 edn. (2006)
15. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota Smarthome: Real-World Activities of Daily Living. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 833–842 (2019)

16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2009)
17. Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., Yuan, L.: DaViT: Dual Attention Vision Transformers. In: Proceedings of the European Conference on Computer Vision. pp. 74–92 (2022)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2021)
19. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In: International Conference on Machine Learning. pp. 2286–2296 (2021)
20. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
21. Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., Zhang, Z.: Star-Transformer. In: Conference of the North American Chapter of the Association for Computational Linguistics. pp. 1315–1325 (2019)
22. Han, D., Pu, Y., Xia, Z., Han, Y., Pan, X., Li, X., Lu, J., Song, S., Huang, G.: Bridging the Divide: Reconsidering Softmax and Linear Attention. In: Advances in Neural Information Processing Systems. pp. 79221–79245 (2024)
23. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8349 (2021)
24. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural Adversarial Examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
25. Huang, J., Qiu, Q., Calderbank, R.: The Role of Principal Angles in Subspace Classification. IEEE Transactions on Signal Processing **64**(8), 1933–1945 (2016)
26. Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver: General Perception with Iterative Attention. In: International Conference on Machine Learning. pp. 4651–4664 (2021)
27. Jaynes, E.T.: Information Theory and Statistical Mechanics. Physical Review **106**(4), 620–630 (1957)

28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
29. Koohpayegani, S.A., Pirsiavash, H.: SimA: Simple Softmax-free Attention for Vision Transformers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2607–2617 (2024)
30. Krause, J., Deng, J., Stark, M., Fei-Fei, L.: Collecting a Large-Scale Dataset of Fine-Grained Cars. In: IEEE Workshop on 3D Representation and Recognition. Sydney, Australia (2013)
31. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Tech. rep., University of Toronto (2009)
32. Laplante, P.A. (ed.): Comprehensive Dictionary of Electrical Engineering. CRC Press, 2 edn. (2018)
33. Laurençon, H., Tronchon, L., Sanh, V.: What matters when building vision-language models? In: Advances in Neural Information Processing Systems. pp. 87874–87907 (2024)
34. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880 (2020)
35. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: International Conference on Machine Learning. pp. 19730–19742 (2023)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
37. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
38. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video Swin Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3202–3211 (2022)
39. Lu, J., Yao, J., Zhang, J., Zhu, X., Xu, H., Gao, W., Xu, C., Xiang, T., Zhang, L.: SOFT: Softmax-free Transformer with Linear Complexity. In: Advances in Neural Information Processing Systems. pp. 21297–21309 (2021)
40. Lv, A., Xie, R., Li, S., Liao, J., Sun, X., Kang, Z., Wang, D., Yan, R.: More Expressive Attention with Negative Weights. arXiv preprint arXiv:2411.07176 (2024)
41. Meta AI: Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/ (2024), Accessed: 2026-06-28
42. Meta AI: Llama 3.2: Vision and edge models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ (2024), Accessed: 2026-06-28
43. Nguyen, T.M., Nguyen, T., Bui, L., Do, H., Nguyen, D.K., Le, D.D., Tran-The, H., Ho, N., Osher, S.J., Baraniuk, R.G.: A Probabilistic Framework for Pruning Transformers Via a Finite Admixture of Keys. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2023)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V.,

Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
46. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
47. Rahimian, A.K., Govind, M.K., Maity, S., Reilly, D., Kümmerle, C., Das, S., Dutta, A.: Fibottention: Inceptive Visual Representation Learning with Diverse Attention Across Heads. arXiv preprint arXiv:2406.19391 (2024)
48. Reilly, D., Govind, M.K., Xue, L., Das, S.: From My View to Yours: Learning Egocentric Cues from Exocentric Video using Privileged Egocentric Supervision. In: Proceedings of the European Conference on Computer Vision (2026)
49. Reilly, D., Govind, M.K., Xue, L., Das, S.: VisCoP: Visual Probing for Video Domain Adaptation of Vision Language Models. In: Proceedings of the European Conference on Computer Vision (2026)
50. Selinger, P.G., Astrahan, M.M., Chamberlin, D.D., Lorie, R.A., Price, T.G.: Access Path Selection in a Relational Database Management System. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. pp. 23—-34 (1979)
51. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysist. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1010–1019 (2016)
52. Shi, H., Gao, J., Ren, X., Xu, H., Liang, X., Li, Z., Kwok, J.T.Y.: SparseBERT: Rethinking the Importance Analysis in Self-attention. In: International Conference on Machine Learning. pp. 9547–9557 (2021)
53. Shuttleworth, R., Andreas, J., Torralba, A., Sharma, P.: LoRA vs Full Fine-tuning: An Illusion of Equivalence. In: Advances in Neural Information Processing Systems. pp. 174627–174662 (2025)
54. Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient Transformers: A Survey. Association for Computing Machinery Computing Surveys **55**(6), 1–28 (2022)
55. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2023)
56. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357 (2021)
57. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going Deeper With Image Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 32–42 (2021)
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Łukasz Kaiser, Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems (2017)
59. Wang, S., Li, B., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020)
60. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Chen, C., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, R., Zou, Z., Zhang, H., Hu, S., Zheng, Z.,

Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M.: Efficient GPT-4V level multimodal large language model for deployment on edge devices. Nature Communications **16**(1), 5509 (2025)
61. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. In: International Conference on Learning Representations (2025)
62. Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., Wei, F.: Differential Transformer. In: International Conference on Learning Representations (2025)
63. Yun, C., Chang, Y.W., Bhojanapalli, S., Rawat, A.S., Reddi, S., Kumar, S.: $O(n)$ Connections are Expressive Enough: Universal Approximability of Sparse Transformers. In: Advances in Neural Information Processing Systems. pp. 13783–13794 (2020)
64. Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., Ahmed, A.: Big Bird: Transformers for Longer Sequences. In: Advances in Neural Information Processing Systems. pp. 17283–17297 (2020)
65. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., Zhao, D.: VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. arXiv preprint arXiv:2501.13106 (2025)
66. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553 (2023)
67. Zhang, J., Li, H., Sra, S., Jadbabaie, A.: Neural Network Weights Do Not Converge to Stationary Points: An Invariant Measure Perspective. In: International Conference on Machine Learning. pp. 26330–26346 (2022)
68. Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., Gao, J.: Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2998–3008 (2021)
69. Zohar, O., Wang, X., Dubois, Y., Mehta, N., Xiao, T., Hansen-Estruch, P., Yu, L., Wang, X., Juefei-Xu, F., Zhang, N., Yeung-Levy, S., Xia, X.: Apollo: An Exploration of Video Understanding in Large Multimodal Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18891–18901 (2025)