DIFFERENTIAL GATED SELF-ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers excel across a large variety of tasks but remain susceptible to corrupted inputs, since standard self-attention treats all query-key interactions uniformly. Inspired by lateral inhibition in biological neural circuits and building on the recent Differential Transformer's use of two parallel softmax subtraction for noise cancellation, we propose *Multihead Differential Gated Self-Attention (M-DGSA)* that learns per-head input-dependent gating to dynamically suppress attention noise. Each head splits into excitatory and inhibitory branches whose dual softmax maps are fused by a sigmoid gate predicted from the token embedding, yielding a context-aware contrast enhancement. M-DGSA integrates seamlessly into existing Transformer stacks with minimal computational overhead. We evaluate on both vision and language benchmarks, demonstrating consistent robustness gains over vanilla Transformer, Vision Transformer, and Differential Transformer baselines. Our contributions are (i) a novel input-dependent gating mechanism for self-attention grounded in lateral inhibition, (ii) a principled synthesis of biological contrast-enhancement and self-attention theory, and (iii) comprehensive experiments demonstrating noise resilience and cross-domain applicability.

1 INTRODUCTION

Attention-based Transformers (Vaswani et al., 2017) have revolutionized machine learning, driving breakthroughs in language, vision, and beyond. However, their uniform weighting of all query-key scores makes them vulnerable to input corruption - sensor noise in images or spurious tokens in text - which can be propagated and even amplified degrading performance in real-world datasets. Existing techniques for enhancing attention robustness or reducing its complexity, typically either apply global regularization (e.g. dropout, weight decay) or rework the attention computation itself via sparsity or low-rank factorization (Wang et al., 2020; Choromanski et al., 2020; Zaheer et al., 2020). More recently, the Differential Transformer (DIFF Transformer) (Ye et al., 2024) introduces a head-wise subtraction of two paired softmax maps to cancel common-mode noise, but its use of a single learned scalar limits its ability to adapt to low-level noise patterns that vary at the granularity of individual tokens.

In this work, we address the challenge of noise-cancellation in self-attention by introducing *Multihead Differential Gated Self-Attention (M-DGSA)*, a lightweight module that learns an input-dependent inhibitory gating per-head to dynamically adapt to and suppress attention noise.

Inspired by *lateral inhibition* in biological neural circuits, where neurons suppress neighboring activity to sharpen responses (Hartline, 1940; Kuffler, 1953), M-DGSA splits each head into excitatory and inhibitory branches. Dual softmax maps are fused via a sigmoid gate predicted from the token embedding, enabling fine-grained contrast enhancement tailored to each query-key pair. M-DGSA integrates seamlessly into off-the-shelf Transformer-based stacks with negligible overhead.

We evaluate M-DGSA on both vision and language benchmarks - namely CIFAR-10, CIFAR-100, FashionMNIST, SVHN, and ImageNet-1k for vision, Rotten Tomatoes, IMDB, AGNews, 20Newsgroups, and MNLI for language - under both clean and noisy settings. Empirically, M-DGSA consistently improves accuracy and noise resilience across these domains, outperforming vanilla Transformer (Vaswani et al., 2017), Vision Transformer (ViT) (Dosovitskiy et al., 2020), and Differential Transformer (Ye et al., 2024) baselines. These results underscore the practical benefit of adaptive, input-dependent inhibition for attention-based architectures for robust performance across diverse downstream tasks.

Our main contributions in this paper are as follows:

- Introducing a per-head, input-conditioned gating mechanism that dynamically fuses the two parallel attention streams, enabling fine-grained, token-wise noise suppression,
- Adopting the lateral-inhibition principles of biological systems' function into artificial attention mechanisms to provide an interpretable contrast-enhancement framework for self-attention,
- Validating M-DGSA across vision and language benchmarks, where it consistently boosts robustness and cross-domain adaptability under clean and corrupted conditions.

2 Related Work

Contrast-Enhancement and Inhibitory Motifs in Learning. In biological sensory neurons, lateral inhibition enhances spatial contrast by subtracting pooled neighboring activity, thereby sharpening feature edges and improving signal-to-noise ratio (Hartline, 1940; Kuffler, 1953).

Inspirations from lateral inhibition have informed modern deep networks. These include Local Response Normalization (LRN) which applies fixed local competition to sharpen feature maps (Krizhevsky et al., 2012); and spatial masking methods such as DropBlock, which attenuate contiguous activations to improve generalization (Ghiasi et al., 2018). More recently, LI-CNN learns per-channel subtractive inhibition filters to dynamically emphasize salient patterns (Zhuang et al., 2023), Selective Kernel Networks (SKNet) employ channel-wise gating to adaptively fuse multiple convolutional kernels and dampen irrelevant features, mirroring lateral inhibition (Li et al., 2019), while Gradient Mask applies inhibitory gating on gradients during backpropagation to filter gradient noise and stabilize training (Jiang et al., 2022).

Attention Variants Towards Robustness. Building on the Transformer (Vaswani et al., 2017) architecture, many attention variants have been proposed to improve efficiency and robustness. Low-rank methods such as Linformer (Wang et al., 2020) and Performer (Choromanski et al., 2020) reduce complexity via kernel approximations, while sparsity-driven models like BigBird (Zaheer et al., 2020) leverage block-sparse patterns. In vision, the Vision Transformer (Dosovitskiy et al., 2020) demonstrates that pure self-attention over image patches, with an extra class token, can rival convolutional networks. ConViT (d'Ascoli et al., 2021) then introduces locality-sensitive gating to blend convolutional inductive biases with self-attention, enabling dynamic control over spatial focus. Recently, the Differential Transformer (Ye et al., 2024) introduces head-wise subtraction of parallel softmax maps to cancel common-mode noise, improve spectral balance and reduce attention collapse. Multi-Token Attention (Golovneva et al., 2025) performs a key–query convolution on the attention scores and applies a head-wise convolution across groups of attention heads to retrieve richer, reliable information from the input data.

Despite this rich history, to the best of our knowledge, head-wise, input-dependent lateral-inhibition remains unexplored in multi-head self-attention.

3 Preliminaries

3.1 Scaled Dot-Product Attention

In the original Transformer formulation (Vaswani et al., 2017), given an input sequence of token embeddings

$$X = [x_1, \ldots, x_N] \in \mathbb{R}^{N \times d_{\text{model}}},$$

we compute query, key, and value matrices via learnable linear projections:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d_{\text{model}} \times d}$. The standard scaled dot-product attention then forms the weight matrix

$$A = \text{softmax}\big(\tfrac{QK^\top}{\sqrt{d}}\big) \in \mathbb{R}^{N \times N},$$

applying the softmax row-wise to normalize each query's affinities, and the final output is

$$\text{Attention}(Q, K, V) = A\,V \in \mathbb{R}^{N \times d}.$$

3.2 DIFFERENTIAL ATTENTION

(Ye et al., 2024) mitigate attention noise by contrasting two parallel attention maps. For an input sequence

$$X = [x_1, \ldots, x_N] \ \in \ \mathbb{R}^{N \times d_{\text{model}}},$$

they first split queries and keys into paired streams and project values jointly:

$$[Q_1; Q_2] = X\, W_Q, \quad [K_1; K_2] = X\, W_K, \quad V = X\, W_V,$$

with $W_Q, W_K \in \mathbb{R}^{d_{\text{model}} \times 2d'}$ and $W_V \in \mathbb{R}^{d_{\text{model}} \times 2d'}$. Each stream then forms a standard scaled dot-product attention,

$$A_i = \text{softmax}\big(\tfrac{Q_i K_i^\top}{\sqrt{d}}\big) \ \in \mathbb{R}^{N \times N}, , \quad i \in \{1, 2\},$$

and their outputs are combined via a learnable subtraction:

$$\text{DiffAttn}(X) = (A_1 - \lambda\, A_2)\, V, \in \mathbb{R}^{N \times 2d'}.$$

Here, λ is reparameterized to stabilize learning:

$$\lambda = \exp\big(\lambda_{q1}\lambda_{k1}\big) - \exp\big(\lambda_{q2}\lambda_{k2}\big) + \lambda_{\text{init}},$$

where $\lambda_{q1}, \lambda_{q2}, \lambda_{k1}, \lambda_{k2} \in \mathbb{R}^d$ are learnable vectors and $\lambda_{\text{init}} = 0.8 - 0.6\, e^{-0.3(l-1)}$ is a layer-dependent constant initialization for layer index $l \in [1, L]$.

3.3 LATERAL INHIBITION

Lateral inhibition is a fundamental neural mechanism that enhances contrast by suppressing nearby activity. In rate-based models, the neuron's response is characterized by

$$r_i = \phi\Big(e_i - \alpha \sum_{j \in \mathcal{N}(i)} e_j\Big),$$

where e_i is the direct excitatory input, $\sum_j e_j$ aggregates neighboring signals, $\alpha > 0$ controls the inhibition strength, and ϕ is a smooth nonlinearity. This local subtractive process accentuates salient features while damping background noise, providing the theoretical underpinning for input-dependent gating in self-attention.

4 METHOD

We propose *Differential Gated Self-Attention (M-DGSA)*, the first self-attention variant to embed an input-dependent lateral inhibition mechanism, combining the common-mode noise cancellation of differential attention with the adaptability of biological contrast-enhancement. Building on the Differential Transformer (Ye et al., 2024) implementation (dual-softmax streams, GroupNorm (Wu & He, 2018), and λ_{init} scheme), we replace its subtraction scalar with a lightweight gating network that learns input-dependent and head-specific inhibitory weights.

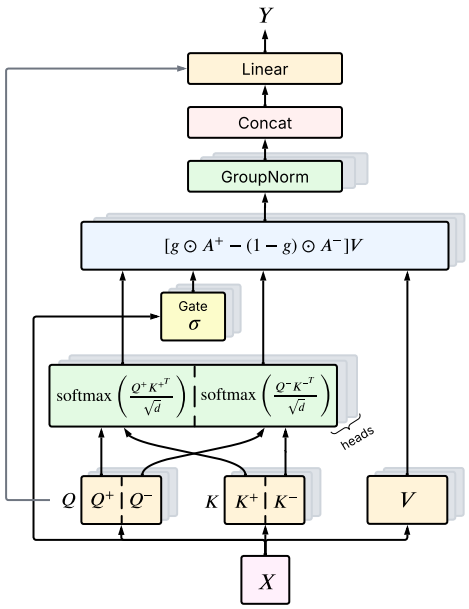
4.1 DIFFERENTIAL GATED SELF-ATTENTION

In our proposed *Differential Gated Self-Attention*, the input tensor $X \in \mathbb{R}^{N \times d_{model}}$ is linearly projected into two query-key substreams, $Q^+, K^+ \in \mathbb{R}^{N \times d'}$ (excitatory) and $Q^-, K^- \in \mathbb{R}^{N \times d'}$ (inhibitory), alongside a unified value stream $V \in \mathbb{R}^{N \times 2d'}$:

$$[\, Q^+; \, Q^- \,] = X\, W_Q, \quad [\, K^+; \, K^- \,] = X\, W_K, \quad V = X\, W_V,$$

where $W_Q, W_K, W_V \ \in \ \mathbb{R}^{d_{\text{model}} \times 2d'}$ are the learnable weight matrices. Then, each substream computes its own scaled dot-product softmax attention map, denoted A^+ and A^-:

$$A^+ = \text{softmax}\big(\tfrac{Q^+ {K^+}^\top}{\sqrt{d}}\big), \quad A^- = \text{softmax}\big(\tfrac{Q^- {K^-}^\top}{\sqrt{d}}\big).$$



Algorithm 1: Multihead Differential Gated Self-Attention (M-DGSA)

Input: $X \in \mathbb{R}^{N \times d_{model}}$, layer index l, weights W_Q, W_K, W_V, W_g, W_o, init λ_{init}

$\lambda_l \leftarrow 0.8 - 0.6\exp(-0.3(l-1))$

$Q \leftarrow XW_Q, \quad K \leftarrow XW_K, \quad V \leftarrow XW_V$

Split Q into $[Q^+; Q^-]$, split K into $[K^+; K^-]$

$A^+ \leftarrow \text{softmax}(Q^+ {K^+}^\top / \sqrt{d'})$

$A^- \leftarrow \text{softmax}(Q^- {K^-}^\top / \sqrt{d'})$

$g \leftarrow \sigma(XW_g + b_g)$

$A \leftarrow g \odot A^+ - (1-g) \odot A^-$

$H \leftarrow (1-\lambda_l)\text{GroupNorm}(AV)$

$O \leftarrow$ reshape and concat heads of H

$Y \leftarrow Q + OW_o$

Output: $Y \in \mathbb{R}^{N \times hd'}$

Figure 1: Design of Multihead Differential Gated Self-Attention. The input tensor is projected into excitatory (Q^+, K^+) and inhibitory (Q^-, K^-) query-key pairs and shared values V. Each branch generates its own softmax map A^+ and A^-, which are then "filtered" by a lightweight input-dependent headwise gating network that learns to attenuate noise - mimicking lateral inhibition in biological sensory systems. The fused attention is applied to V and each head's output undergoes head-wise GroupNorm (Wu & He, 2018) before all heads are concatenated and linearly projected. An optional residual connection around the attention block can be added to improve gradient flow and training stability.

Here we introduce a gating network, a per-input (token) t, per-head h sigmoid gate

$$g_{t,head_i} = \sigma\big(w_{g,head_i} x_t + b_{g,head_i}\big),$$

which adaptively fuses excitation and inhibition streams

$$A_{t,head_i} = g_{t,head_i}\, A^+_{t,head_i} \;-\; (1 - g_{t,head_i})\, A^-_{t,head_i}. \tag{1}$$

We emphasize subtracting the two gated-weighted terms, rather than blending them via weighted addition as in conventional gating mechanisms. By doing so, each position can dynamically choose how much of the inhibitory branch to apply, amplifying strong signals where needed and quelling noise elsewhere.

4.2 Multi-head Differential Gated Self-Attention

We extend the DGSA to h heads, as in the original Transformer multi-head attention mechanism (Vaswani et al., 2017). Thereby, we tile the projection matrices into

$$[\mathbf{Q}^+; \mathbf{Q}^-],\ [\mathbf{K}^+; \mathbf{K}^-] \;\in\; \mathbb{R}^{N \times 2h \times d'}, \quad \mathbf{V} \;\in\; \mathbb{R}^{N \times h \times 2d'},$$

and then reshape into $\{Q^{\pm}_{head_i}, K^{\pm}_{head_i}, V_{head_i}\}_1^h$, each of shape $N \times 2 \times d'$. Each head independently computes its fused map A_{head_i} as in equation 1 and used to attend over $\mathbf{V}$, in the form of

$$H_{head_i} = A_{head_i}\, V_{head_i}.$$

This is followed by a GroupNorm (Wu & He, 2018) where RMSNorm (Zhang & Sennrich, 2019) is applied in each head and scaled by $(1 - \lambda_{init})$.

Concatenating $[head_1, \ldots, head_i, \ldots, head_h]$ along the embedding dimension and projecting with $W_O \in \mathbb{R}^{hd' \times hd'}$ and adding an optional residual connection from $[\mathbf{Q}^+; \mathbf{Q}^-]$ to preserve information flow in our model. This completes the design of our multi-head attention block.

4.3 DGT AND DGVIT: TRANSFORMER & VISION TRANSFORMER INSTANTIATIONS

Our M-DGSA can be applied in any multi-head attention within Transformer-based models. We explore two instantiations:

Differential Gated Transformer (DGT). Starting from the standard Transformer encoder, we replace every multi-head self-attention block with M-DGSA. Each M-DGSA output is followed by a feed-forward network using the SwiGLU activation (Shazeer, 2020). Unlike in vision tasks, we found that an additional skip connection (shown in Fig. 1) around the M-DGSA block did not improve convergence so we omit it in DGT.

Differential Gated Vision Transformer (DGViT). Building on the ViT backbone, DGViT replaces each attention module with an M-DGSA layer, retaining the learnable class token in the attention computation and substitutes the GeLU activation (Hendrycks & Gimpel, 2016) in the feed-forward network with SwiGLU. It also integrates an optional residual connection inside the M-DGSA block to facilitate gradient flow, stabilize training, and boost model's performance.

5 EXPERIMENTS

We evaluate the proposed Multihead Differential Gated Self-Attention (M-DGSA) across two representative domains: vision and natural language classification. The complete experimental setup, hyperparameter schedules, and dataset details are provided in Appendix B.

All models were trained from scratch (i.e., without pretrained weights) to isolate the contribution of the attention mechanism. As a result, absolute accuracies may be lower than prior work that finetunes pretrained backbones, however, this setting provides a controlled basis for cross-architecture comparison. Experiments were implemented in PyTorch 2.x and executed on a single machine equipped with an NVIDIA A10G (24 GB) for smaller datasets, and on servers with NVIDIA A100 (40 GB and 80 GB) GPUs for ImageNet-1k.

5.1 VISION CLASSIFICATION

For vision classification, we leverage the built-in `torchvision.datasets` module, which provides standardized benchmarks including, CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), Fashion-MNIST (Xiao et al., 2017), and the Street View House Numbers (SVHN) dataset (Netzer et al., 2011). For large-scale evaluation, we additionally include ImageNet-1k (Deng et al., 2009). To assess robustness under challenging conditions, we further inject synthetic noise into the training images at multiple distortion levels.

Models. For the four smaller benchmarks, comparisons include the original Vision Transformer (ViT) and our Differential Gated ViT (DGViT). We also adapt the Differential Transformer (Ye et al., 2024) to the vision setting, denoted DViT. In DGViT and DViT, the feed-forward module expands the hidden dimension by a factor of two or four, applies a SwiGLU activation, and then projects back to the base size. To study the role of regularization, we interleave one or two dropout layers between these projection and activation steps, thereby assessing the effect of different levels of stochastic masking during training.

For ImageNet-1k, we compare the performance of ViT, DViT, and DGViT under two configurations: a $4\times$ expansion with two dropout layers, which serves as our default, and a $16/3\times$ expansion with two dropout layers, corresponding to the canonical setting of the Differential Transformer and providing a higher-capacity regime for comparison.

Results on small benchmarks. Table 1 reports accuracies averaged over five seeds on CIFAR-10, CIFAR-100, FashionMNIST, and SVHN. DGViT consistently outperforms the baseline ViT, with the largest improvement on CIFAR-10 (+2.2%), highlighting the value of gated inhibition in low-resolution, noise-sensitive settings. On the other datasets, improvements are more modest (typically within 0.2-0.6%), yet they remain steady across configurations and never come at the cost of degraded accuracy. When compared directly to DViT, DGViT yields competitive or superior performance in all settings. On CIFAR-100, where DViT underperforms relative to ViT, DGViT closes this gap and establishes a clear advantage. On FashionMNIST, DGViT slightly surpasses DViT, demonstrating that input-dependent gating remains beneficial even in simpler domains. On SVHN, DGViT matches or marginally improves over DViT, confirming that the gating mechanism does not erode the gains provided by differential attention.

Table 1: Classification accuracy (mean ± std) on small vision benchmarks. Averaged over 5 seeds.

		CIFAR-10 (%)	CIFAR-100 (%)	FashionMNIST (%)	SVHN (%)
#Classes		10	100	10	10
#Size		60,000	60,000	70,000	600,000
ViT		75.90 ± 0.09	47.01 ± 0.56	93.15 ± 0.11	93.22 ± 0.28
DViT					
proj.	drop.				
$\times 2$	$\times 2$	76.76 ± 0.28	45.49 ± 0.52	93.39 ± 0.10	93.79 ± 0.003
$\times 4$	$\times 1$	76.40 ± 0.40	44.82 ± 0.41	93.32 ± 0.001	93.67 ± 0.001
$\times 4$	$\times 2$	76.82 ± 0.42	45.55 ± 0.01	93.34 ± 0.001	93.80 ± 0.003
$\times 16/3$	$\times 2$	77.16 ± 0.01	45.66 ± 0.26	93.34 ± 0.002	93.99 ± 0.001
DGViT (ours)					
proj.	drop.				
$\times 2$	$\times 2$	77.51 ± 0.44	47.13 ± 0.26	93.42 ± 0.07	93.93 ± 0.003
$\times 4$	$\times 1$	77.77 ± 0.29	47.32 ± 0.38	93.32 ± 0.10	$\mathbf{94.16 \pm 0.004}$
$\times 4$	$\times 2$	78.06 ± 0.41	$\mathbf{47.72 \pm 0.35}$	$\mathbf{93.45 \pm 0.13}$	93.94 ± 0.002
$\times 16/3$	$\times 2$	$\mathbf{78.31 \pm 0.01}$	47.43 ± 0.03	$\mathbf{93.45 \pm 0.001}$	93.89 ± 0.002

Table 2: Classification accuracy (mean ± std) on ImageNet. Averaged over 5 seeds.

Model		Accuracy (%)
ViT		59.94 ± 0.04
DViT		
proj.	drop.	
$\times 4$	$\times 2$	65.75 ± 0.01
$\times 16/3$	$\times 2$	66.66 ± 0.006
DGViT (ours)		
proj.	drop.	
$\times 4$	$\times 2$	66.10 ± 0.004
$\times 16/3$	$\times 2$	$\mathbf{67.08 \pm 0.003}$

These narrow but steady margins across diverse datasets suggest that the contribution of the gating mechanism is not limited to specific domains but rather yields a systematic robustness benefit, complementing differential attention without introducing instability. In practice, this reliability is as important as large improvements on single datasets, as it demonstrates that the method scales consistently across different visual conditions.

Results on ImageNet-1k. Table 2 summarizes performance at scale. Both DViT and DGViT substantially improve over the baseline ViT, with gains exceeding +6% absolute accuracy. Between the two differential variants, DGViT achieves consistent, though moderate, improvements over DViT for both expansion factors. These findings indicate that input-dependent gating provides an additional layer of robustness on top of differential attention, yielding the strongest overall performance while preserving computational efficiency.

5.2 Language Classification

For language classification, we design a progression of tasks that increase in difficulty with respect to the number of target classes and the diversity of linguistic structure. The evaluation begins with binary sentiment classification (Rotten Tomatoes (Pang & Lee, 2005), IMDB (Maas et al., 2011)), progresses to four-way news categorization task (AG News (Del Corso et al., 2005)), and extends to the 20-class 20 Newsgroups benchmark (Mitchell, 1997), which demands fine-grained discrimination across diverse categories. To complement this controlled scaling, we additionally evaluate on MNLI (Williams et al., 2018), a large-scale natural language inference benchmark spanning multiple genres, which serves as a rigorous test of generalization under heterogeneous conditions. Together, this suite of datasets provides a principled framework for assessing both the performance and robustness of our approach across varying levels of classification complexity.

Table 3: Classification accuracy (%) (mean ± std) on language benchmarks. Averaged over 5 seeds.

	Rotten Tom.	IMDB	MNLI	AGNews	20 Newsg.
#Classes	2	2	3	4	20
#Size	10,000	50,000	433,000	127,600	18,000
Transformer	72.18 ± 0.51	85.34 ± 0.22	55.03 ± 0.01	91.61 ± 0.08	51.54 ± 0.49
DT					
proj.					
$\times 2$	73.99 ± 0.91	86.41 ± 0.14	57.00 ± 0.004	91.85 ± 0.23	56.49 ± 1.28
$\times 4$	73.06 ± 0.70	86.49 ± 0.29	56.11 ± 0.01	92.03 ± 0.20	58.63 ± 0.87
$\times 16/3$	73.19 ± 0.85	86.43 ± 0.16	56.90 ± 0.003	92.00 ± 0.19	58.75 ± 1.21
DGT (ours)					
proj.					
$\times 2$	74.26 ± 0.97	86.47 ± 0.17	57.60 ± 0.004	$\mathbf{92.25 \pm 0.14}$	60.32 ± 0.49
$\times 4$	73.83 ± 1.15	$\mathbf{86.97 \pm 0.04}$	$\mathbf{58.07 \pm 0.001}$	91.89 ± 0.14	59.07 ± 0.48
$\times 16/3$	$\mathbf{74.40 \pm 0.42}$	86.53 ± 0.33	57.26 ± 0.004	91.98 ± 0.27	$\mathbf{63.53 \pm 0.87}$

Models. We compare our DGT model's performance to both the vanilla Transformer and the Differential Transformer (DT) under default configurations. To ensure fairness and to isolate the contribution of the attention mechanism, we adopt identical experimental conditions for DT and DGT: the feed-forward module expands the hidden dimension by factors of 2, 4, and 16/3 relative to the output dimension Y, applying SwiGLU before projecting back to the base dimension. This alignment guarantees that performance differences can be attributed to the gating mechanism rather than to architectural or training discrepancies.

Results. Table 3 reports mean accuracy over five seeds. DGT improves over both baselines across all datasets. On the simpler sentiment tasks (Rotten Tomatoes, IMDB) and the four-class AG News, gains are modest (typically +1-2%), reflecting that these benchmarks are near-saturated with strong baselines. Importantly, DGT maintains or improves performance across all expansion settings, indicating that input-dependent gating introduces no regressions even in low-complexity regimes. The advantages are most evident on 20 Newsgroups, where DGT consistently surpasses both baselines, demonstrating its capacity to handle fine-grained, high-variance classification with greater robustness. On MNLI, DGT consistently outperforms both baselines across expansion factors, with improvements of up to +3% over the vanilla Transformer and +1-1.5% over DT, underscoring its effectiveness in large-scale natural language inference. Overall, the findings indicate that input-dependent gating yields consistent improvements on easier benchmarks and delivers pronounced advantages as task difficulty rises, underscoring its ability to scale reliably and adapt to diverse linguistic settings.

5.3 Attention Visualization.

We use the attention-rollout method (Abnar & Zuidema, 2020) to inspect how our models filter noise and sharpen structure on representative vision and language examples.

Vision (DGViT vs Baselines). Fig. 2 illustrates attention-rollout heatmaps for DGViT, DViT, and ViT on CIFAR-100 and SVHN. While ViT spreads attention broadly, often assigning high weights to background regions alongside the object of interest, DViT applies a differential contrast mechanism that partly sharpens focus but still leaves residual noise. DGViT, by contrast, produces the most discriminative patterns: excitatory weights are concentrated on semantically salient structures (e.g., mushroom contours or digit strokes), while inhibitory weights actively suppress background clutter, resulting in clearer boundaries and a more reliable focus on task-relevant regions.

Language (DGT vs Baselines). Fig. 3 presents attention-rollout for (a) Transformer, (b) Differential Transformer, and (c) our DGT on a 20 Newsgroups sample, input text: *"Most graphics systems I have seen have drawing routines that also specify a color for drawing, like Drawpoint(x,y,color) or Drawline(x1,y1,x2,y2,color) or Fillrectangle(x1,y1,x2,y2,color)... With X, I..."*, label: *comp.windows.x*. The vanilla Transformer disperses its weights broadly over both common function words and various candidate labels, failing to single out the true class while the Differential Transformer partly sup-

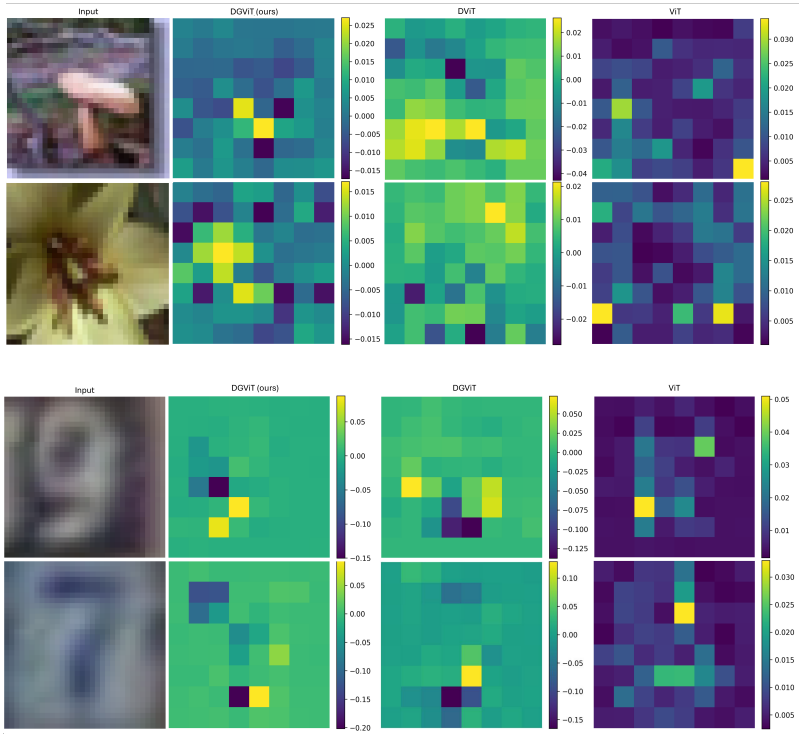


Figure 2: Attention-rollout heatmaps for DGViT vs DViT vs ViT on CIFAR-100 (top two rows) and SVHN (bottom two rows). Bright yellow indicates high attention weights, whereas deep blue denotes low or suppressed attention. ViT disperses attention across both objects and background, DViT partially cancels common-mode noise but retains some distraction, while DGViT exhibits sharper boundary delineation and stronger alignment with task-relevant features.

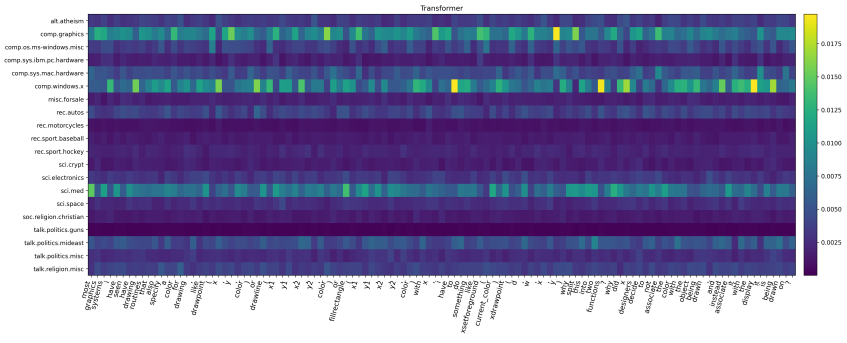
presses common-mode noise yet still spreads attention to competing classes, diluting its focus on the correct one. DGT's input-dependent gating, in contrast, concentrates focus on key class-indicative words, filtering out irrelevant text and yielding a more interpretable, discriminative attention pattern, increasing attention mass on ground-truth and aligning sharply with *comp.windows.x*.

6 Discussion

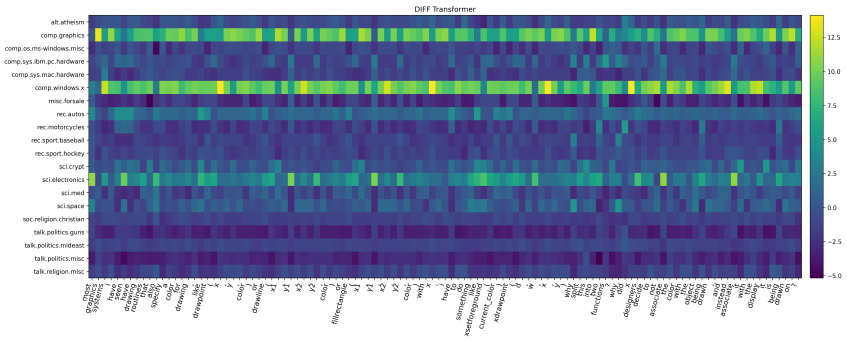
Our results demonstrate that embedding an input-conditioned lateral inhibition mechanism into self-attention yields substantial gains in noise resilience and generalization across both vision and language tasks. M-DGSA consistently sharpens attention maps, assigning strong excitatory weights to semantically relevant regions while suppressing background noise, without requiring any task-specific tuning. Its capacity to delineate fine boundaries suggests applications in safety-critical domains - such as lane-marking detection in autonomous driving or edge delineation in medical imaging - where robust, fine-grained attention is paramount. Despite these benefits, several limitations invite further exploration. We have so far evaluated primarily on mid- and large-scale benchmarks with synthetic noise, broadening the analysis to real-world distortions (e.g., motion blur, occlusions) may uncover additional strengths or weaknesses. Moreover, our study focused on self-attention, incorporating lateral inhibition into cross-attention - such as in vision-language models for VQA or image captioning - could enable dynamic, context-conditioned suppression of irrelevant inputs and further strengthen multimodal alignment. Looking forward, adapting lateral-inhibition gating to multimodal architectures, scaling up to natural corruption regimes, and developing hardware-efficient implementations that leverage the induced sparsity represent particularly promising directions.

7 Conclusion

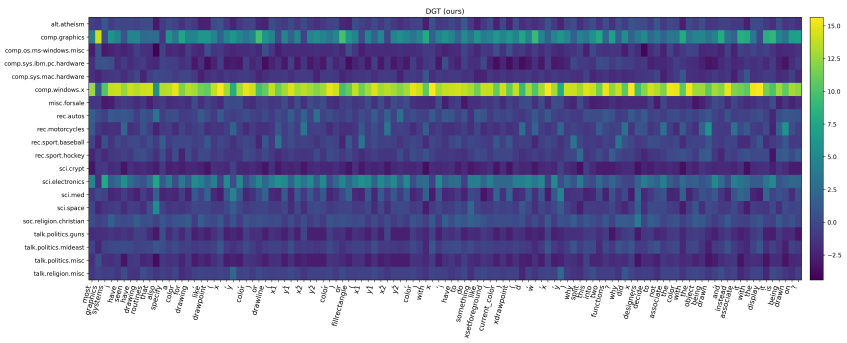
We have introduced Multihead Differential Gated Self-Attention (M-DGSA), a lightweight yet powerful extension of multihead self-attention that embeds per-head input-dependent lateral inhibition



(a) Transformer model attention



(b) Differential Transformer model attention



(c) DGT (ours) model attention

Figure 3: Attention-rollout for a 20 Newsgroups sample. Bright yellow indicates strong positive attention, and deep blue denotes inhibitory weights. (a) The standard Transformer baseline spreads its attention diffusely across both function words and content tokens, failing to isolate the class-defining terms. (b) The Differential Transformer partially suppresses common-mode noise, but still assigns non-negligible weight to other classes too. (c) Our DGT model sharply concentrates on the key parts of the input while attenuating filler words. The focused, input-dependent gating aligns attention with the correct *comp.windows.x* label and filters out background noise.

into the dual-softmax framework. Leveraging dynamic, context-sensitive gating, M-DGSA produces sharper and more selective attention maps, enhancing noise resilience and improving cross-domain performance on both vision and language benchmarks. Importantly, it integrates seamlessly into existing Transformer and ViT architectures with negligible overhead, making it directly applicable to a wide range of tasks.

Beyond empirical improvements, M-DGSA advances the interpretability of attention mechanisms: its gating patterns mirror biological contrast-enhancement, yielding clear visualizations of how models emphasize salient features and quench noise. We view this biologically inspired paradigm as a significant step toward more robust, trustworthy attention models for next-generation deep learning.

Reproducibility Statement

We have taken several measures to facilitate reproducibility of our results. All datasets are standard public benchmarks with canonical train/test splits, requiring no additional preprocessing beyond what is provided in their official releases B.1, B.2. Hyperparameter settings, model sizes, and training schedules for both vision and language tasks are detailed in Tables 4-8 of the Appendix. Compute requirements and runtime estimates are reported in Appendix B.4, and additional ablation studies on initialization and gate depth are provided in Appendix B.5. Our implementation is based on PyTorch 2.x, and an anonymous link to the source code is also included in the supplementary material B.5 to enable full replication of our experiments.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Gianna M Del Corso, Antonio Gulli, and Francesco Romani. Ranking a stream of news. In *Proceedings of the 14th international conference on World Wide Web*, pp. 97–106, 2005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, pp. 2286–2296. PMLR, 2021.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. Multi-token attention. *arXiv preprint arXiv:2504.00927*, 2025.

Haldan K Hartline. The receptive fields of optic nerve fibers. *American Journal of Physiology-Legacy Content*, 130(4):690–699, 1940.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Lei Jiang, Yongqing Liu, Shihai Xiao, and Yansong Chua. Gradient mask: Lateral inhibition mechanism improves performance in artificial neural networks. *arXiv preprint arXiv:2208.06918*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 510–519, 2019.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*, 2005.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

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

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

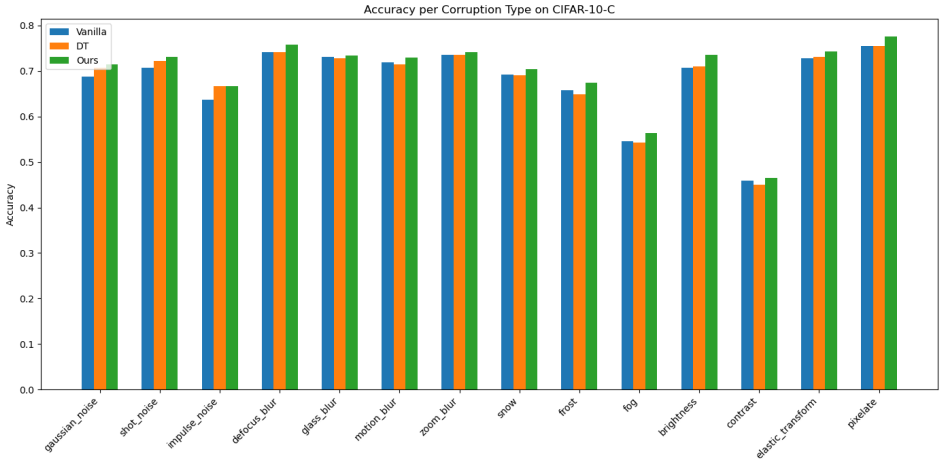
Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

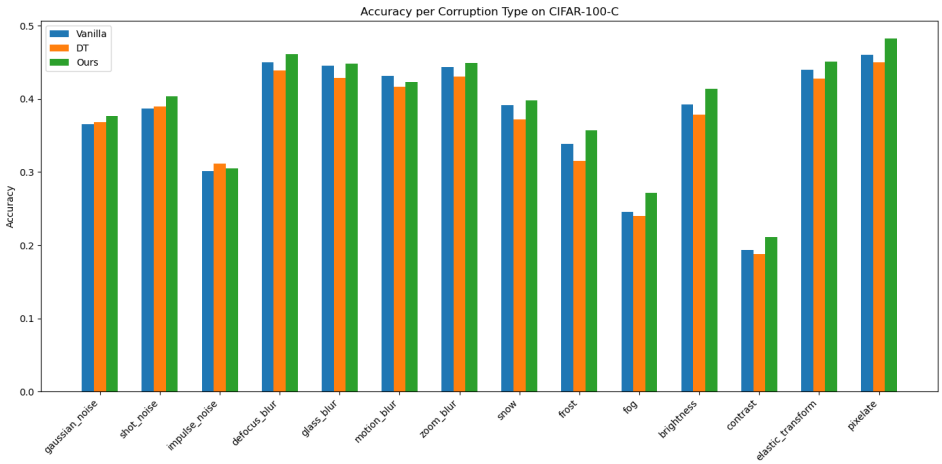
Chengyuan Zhuang, Xiaohui Yuan, and XUAN GUO. Suppression helps: Lateral inhibition-inspired convolutional neural network for image classification, 2023. URL https://openreview.net/forum?id=N3kGYG3ZcTi.

APPENDIX

A ADDITIONAL EXPERIMENTS



(a)



(b)

Figure 4: Experiments of the vanilla ViT, Differential Transformer (DT) and our model on the (a) CIFAR-10-C and (b) CIFAR-100-C datasets.

Across both CIFAR-10-C and CIFAR-100-C, our method provides consistent, corruption-agnostic robustness improvements over both the Vanilla and DT baselines. On CIFAR-10-C (Fig. 4a), our model achieves higher accuracy for nearly all corruption types, with particularly strong gains on noise (Gaussian, Shot), blur (Defocus, Motion, Zoom), weather effects (Snow, Frost), and compression artifacts (JPEG). The largest improvements occur under severe global distribution shifts, such as fog and contrast, where both baselines degrade sharply. These gains demonstrate that our approach preserves both high-frequency and global structural features under complex perturbations. We note that Vanilla performs marginally better on impulse_noise at low severities; however, this difference is extremely small and disappears at higher severities. Importantly, impulse noise is the only corruption where the baseline is competitive, underscoring the broad generality of our robustness enhancements.

On CIFAR-100-C (Fig. 4b), where robustness is significantly more challenging due to the increased label space and lower clean accuracy, the same pattern holds. Our method outperforms both baselines across almost all corruptions, including the hardest ones (fog, frost, contrast), while maintaining competitive performance even in the few cases - such as impulse noise - where the Vanilla model

shows a slight advantage. Overall, the results across both datasets highlight that our method provides uniform, corruption-agnostic robustness improvements, rather than overfitting to a specific corruption family, a property that is critical for real-world robustness.

B Additional Experimental Details

B.1 Detailed description of the used datasets

Rotten Tomatoes. (Pang & Lee, 2005) $10,662$ movie reviews equally split between positive and negative sentiment.

IMDB. (Maas et al., 2011) $50,000$ highly polarized reviews ($25,000$ train / $25,000$ test) for binary sentiment classification.

MNLI. (Williams et al., 2018) The Multi-Genre Natural Language Inference corpus is a dataset of 433,000 sentence pairs annotated with textual entailment, genre and label.

AG News. (Del Corso et al., 2005) Derived from AG's corpus of news articles, the AG News dataset comprises four categories - World, Sports, Business, and Science/Technology - by combining each article's title and description and selecting the four largest classes. It contains $120,000$ training samples ($30,000$ per category) and $7,600$ test samples ($1,900$ per category).

20 Newsgroups. (Mitchell, 1997) $18,000$ posts organized into 20 distinct newsgroups, each corresponding to a specific topic such as computer hardware, recreational activities, politics, science, or religion. We adopt the canonical split of $11,314$ training and $7,532$ test posts, yielding on average about 566 training and 377 test samples per class.

CIFAR-10 / CIFAR-100. (Krizhevsky et al., 2009) $60,000$ colored images of size 32x32 pixels, evenly distributed across 10 distinct classes such as airplanes, automobiles, birds, cats, deers, dogs, frogs, horses, ships, and trucks. Each class includes $6,000$ images, with the dataset split into $50,000$ training and $10,000$ test images. CIFAR-100 is based on the same concept but with 100 classes and contains 600 samples for each class.

CIFAR-10-C / CIFAR-100-C. (Hendrycks & Dietterich, 2019) CIFAR-10-C and CIFAR-100-C are corruption robustness benchmarks derived from CIFAR-10 and CIFAR-100, where test images are modified using 19 common corruption types (e.g., noise, blur, weather, and digital distortions) at five severity levels to systematically evaluate model robustness.

FashionMNIST. (Xiao et al., 2017) $70,000$ 28×28 pixel grayscale Zalando article images of 10 fashion item categories (e.g., t-shirt/top, trouser, bag, sandal). Maintaining the same structure and format as MNIST, FashionMNIST provides a standardized yet more complex alternative for evaluating machine learning models on image classification tasks.

Street View House Numbers (SVHN). (Netzer et al., 2011) A large-scale, real-world image dataset comprising over 600,000 32×32-pixel digit crops extracted from Google Street View house numbers. It includes 10 classes (digits 0-9): 73,257 training images, 26,032 test images, and 531,131 additional "extra" samples of comparatively easier digits. By capturing digits in varied natural scenes - with diverse backgrounds, lighting conditions, and occlusions - SVHN poses a substantially more challenging recognition task than MNIST.

ImageNet. (Deng et al., 2009) A large-scale, human-annotated image dataset organized according to the WordNet hierarchy, aiming to provide around 1000 images per concept, with tens of millions of labeled images overall. It was created to advance computer vision research by offering a high-quality benchmark for object categorization and supplying the large amounts of data needed for developing more generalizable machine learning methods.

Licenses for existing assets

- **Rotten Tomatoes** (Pang & Lee, 2005) Academic use only, see dataset terms at `https://nlp.stanford.edu/sentiment/index.html`

- **IMDB** (Maas et al., 2011) Non-commercial academic use (see Stanford ACL IMDB dataset page). https://ai.stanford.edu/~amaas/data/sentiment/
- **MLNI** OANC's license. https://huggingface.co/datasets/nyu-mll/multi_nli
- **AG News** (Del Corso et al., 2005) Distributed under the CC BY-SA 3.0 license. https://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
- **20 Newsgroups** (Mitchell, 1997) Originally Usenet posts—distributed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). http://qwone.com/~jason/20Newsgroups/
- **CIFAR-10 / CIFAR-100** (Krizhevsky et al., 2009) Released under the MIT License. https://www.cs.toronto.edu/~kriz/cifar.html
- **CIFAR-10-C / CIFAR-100-C** (Hendrycks & Dietterich, 2019) Released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. https://zenodo.org/record/2535967, https://zenodo.org/record/3555552
- **Fashion-MNIST** (Xiao et al., 2017) Public domain (MNIST) / MIT License (Fashion-MNIST). http://yann.lecun.com/exdb/mnist/, https://github.com/zalandoresearch/fashion-mnist
- **SVHN** (Netzer et al., 2011) Released under the MIT License. http://ufldl.stanford.edu/housenumbers/
- **ImageNet** (Deng et al., 2009) Non-commercial academic research and educational use. https://www.image-net.org/index.php

B.2 Dataset Splits

For each language dataset, AG News, IMDB, 20 Newsgroups, and Rotten Tomatoes, we further split the training fold into 80% train and 20% validation (stratified by class). All final results are reported on the official test set. For all vision benchmarks, CIFAR-10, CIFAR-100, Fashion-MNIST, and SVHN, we use the standard train/test partitions provided by each dataset and no further hold-out split was applied.

B.3 Hyperparameters

B.3.1 Vision Benchmarks

For all vision datasets, we use the hyperparameters reported in Table 4. For ImageNet-1k, we set the batch size to 64 and the patch size to 16.

Table 4: Hyperparameters for vision datasets.

Hyperparameter	Value
Epochs	100
Batch size	128
Learning rate	3×10^{-4}
Embedding dimension	256
# Layers	8
# Heads	8
Dropout probability	0.05
Weight decay	0.01
Optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.999,\ \epsilon = 10^{-8}$)
LR scheduler	CosineAnnealingLR
Patch size	4
Max sequence length	100

B.3.2 LANGUAGE BENCHMARKS

For AG News, IMDB, and 20 Newsgroups, all text-classification experiments use the hyperparameters listed in Table 5. For Rotten Tomatoes, we instead employ 4 layers, a learning rate of 5×10^{-4}, and 500 warmup steps. The hyperparameters for MNLI are reported separately in Table 6.

Table 5: Hyperparameters for small text datasets.

Hyperparameter	Value
Epochs	10
Batch size	32
Learning rate	3×10^{-4}
Warmup steps	1000
Max sequence length	256
Minimum token frequency	2
Maximum vocabulary size	60,000
Embedding dimension	256
# layers	6
# heads	8
Dropout probability	0.1
Optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.98,\ \epsilon = 10^{-8}$)
Weight decay	0.01 (0 for bias & LayerNorm/RMSNorm weights)
LR scheduler	Linear warmup (1000 steps) + linear decay
Gradient clipping	Max-norm = 1.0

Table 6: Hyperparameters for MNLI dataset.

Hyperparameter	Value
Epochs	5
Batch size	128
Learning rate	3×10^{-4}
Warmup steps	4000
Max sequence length	256
Minimum token frequency	2
Maximum vocabulary size	30,000
Embedding dimension	512
# layers	6
# heads	8
Dropout probability	0.1
Optimizer	AdamW ($\beta_1 = 0.9,\ \beta_2 = 0.98,\ \epsilon = 10^{-8}$)
Weight decay	0.01 (0 for bias & LayerNorm/RMSNorm weights)
LR scheduler	Linear warmup (4000 steps) + linear decay
Gradient clipping	Max-norm = 1.0

B.4 COMPUTE & RUNTIME

ImageNet-1k experiments ran on NVIDIA A100 (40 GB and 80 GB) GPUs. All other experiments ran in PyTorch on a single NVIDIA A10G (24 GB).

- Vision: $\sim$0.5min/epoch on CIFAR-10/100, $\sim$0.6 min on FashionMNIST, $\sim$0.76 min on SVHN, and $\sim$31min/epoch on ImageNet.
- Language: $\sim$0.3min/epoch on Rotten Tomatoes, $\sim$0.39 min on IMDB, $\sim$0.57 min on AG News, $\sim$0.5 min on 20NewsGroups, and $\sim$18min/epoch on MNLI.

With DT and DGT, our goal was to avoid increasing - and in some cases even to reduce - the number of learnable parameters relative to the standard Transformer. This reduction stems from the use of

SwiGLU, which splits projections in half, whereas the Transformer employs GeLU over the full hidden dimension.

The standard ViT has 5.9M parameters across CIFAR-10, CIFAR-100, FashionMNIST, and SVHN. For DViT and DGViT, the parameter count varies with the projection factor: a $\times 2$ projection yields 3.8M parameters, while a $\times 4$ projection increases this to 5.4M.

In the ImageNet experiments, all models maintain a comparable computational cost of 1.2 GFLOPs regardless of architecture. The baseline ViT has 5.7M learnable parameters and uses 2.7 GB of memory. DViT and DGViT with a $\times 4$ projection keep the parameter count at 5.7M, but memory usage rises to 4.3 GB. The $\times 16/3$ projection further increases parameters to 6.7M, while leaving the memory footprint unchanged. These results indicate that our architectural modifications primarily affect parameter count and memory requirements, while leaving overall FLOPs constant.

The number of trainable parameters in the language tasks of Rotten Tomatoes, IMDB, AGNews and 20 Newsgroups are reported in Table 7 and for the MNLI in Table 8.

Table 7: Number of learnable parameters for language benchmarks.

	Rotten Tomatoes	IMDB	AGNews	20 Newsgroups
Transformer	5.1M	27.6M	15.1M	13.3M
DT/DGT				
proj. $\times 2$	3.8M	25.7M	13.2M	12M
proj. $\times 4$	4.6M	26.9M	14.4M	12.8M
proj. $\times 16/3$	5.1M	27.7M	15.2M	13.4M

Table 8: Metrics from the MNLI experiments.

	No. of learnable params	Memory usage (in MB)	GFLOPs
Transformer	31.4M	1.3k	4.4
DT/DGT			
proj. $\times 2$	26.7M	2.1k	3.2
proj. $\times 4$	31.4M	2.1k	4.4
proj. $\times 16/3$	34.6M	2.1k	5.2

Dataset	Model	# Params	FLOPs / Img	Throughput	Latency	Peak Fwd Mem	Peak Train Mem
CIFAR-10	ViTx4	6,591,626	1.441 GFLOPs	40.17 it/s (2320.09 img/s)	24.89 ms	164.5 MB	2594.8 MB
CIFAR-10	DViTx4	5,543,818	1.234 GFLOPs	30.49 it/s (1951.11 img/s)	32.80 ms	236.8 MB	4070.3 MB
CIFAR-10	DGViTx4	5,559,754	1.238 GFLOPs	23.73 it/s (1518.91 img/s)	42.14 ms	274.8 MB	4082.9 MB
CIFAR-100	ViTx4	6,611,428	1.441 GFLOPs	39.95 it/s (2557.08 img/s)	25.03 ms	164.6 MB	2594.9 MB
CIFAR-100	DViTx4	5,555,428	1.234 GFLOPs	30.39 it/s (1944.94 img/s)	32.91 ms	236.8 MB	4070.5 MB
CIFAR-100	DGViTx4	5,571,364	1.238 GFLOPs	23.74 it/s (1519.66 img/s)	42.11 ms	274.8 MB	4083.2 MB
FashionMNIST	ViTx4	6,591,882	1.441 GFLOPs	39.94 it/s (2556.39 img/s)	24.45 ms	164.5 MB	2820.4 MB
FashionMNIST	DViTx4	5,543,818	1.234 GFLOPs	30.41 it/s (1946.48 img/s)	32.88 ms	236.8 MB	4070.3 MB
FashionMNIST	DGViTx4	5,559,754	1.238 GFLOPs	23.73 it/s (1518.74 img/s)	42.14 ms	274.8 MB	4082.9 MB
SVHN	ViTx4	6,591,626	1.441 GFLOPs	39.95 it/s (2556.93 img/s)	25.03 ms	164.5 MB	2594.7 MB
SVHN	DViTx4	5,543,818	1.234 GFLOPs	30.40 it/s (1945.49 img/s)	32.90 ms	236.8 MB	4070.3 MB
SVHN	DGViTx4	5,559,754	1.238 GFLOPs	23.74 it/s (1519.32 img/s)	42.12 ms	274.8 MB	4082.9 MB
ImageNet-1k	ViTx4	6,727,528	1.441 GFLOPs	36.25 it/s (2320.06 img/s)	25.05 ms	165.0 MB	2597.0 MB
ImageNet-1k	DViTx4	5,671,528	1.235 GFLOPs	30.40 it/s (1945.78 img/s)	32.89 ms	237.3 MB	4072.5 MB
ImageNet-1k	DGViTx4	5,687,464	1.238 GFLOPs	23.73 it/s (1518.64 img/s)	42.14 ms	275.3 MB	4085.2 MB

Table 9: Performance and resource metrics for ViTx4, DViTx4, and DGViTx4 across datasets.

B.5 Additional Analyses

Ablation on λ_{init} After experimenting with various λ_{init} settings, we found that holding it fixed at 0.8 outperforms both other constant values and treating it as a learnable parameter.

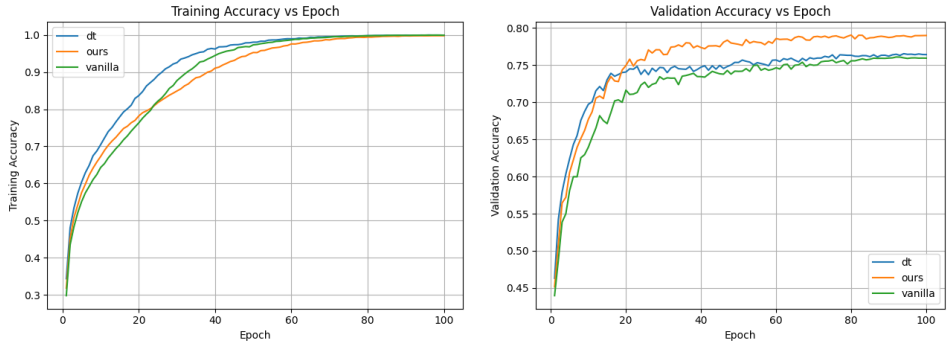


Figure 5: Training and validation convergence of the vanilla Transformer (vanilla), Differential Transformer (dt) and our models.

Gate depth We observed that increasing the depth of the MLP gating layers consistently degraded the performance, therefore, our method employs a single-layer gating mechanism.

CODE AVAILABILITY

Code is available at: `https://anonymous.4open.science/r/DGSA-1-BDE1/`