Density Adaptive Attention Mechanism: Robust & Explainable Representations Across Modalities

Georgios Ioannides[1,2], Aman Chadha[3,2], Aaron Elkins[4]
[1]Carnegie Mellon University [2]Amazon GenAI*
[3]Stanford University [4]James Silberrad Brown Center for AI
*Work does not relate to position at Amazon.

Why a new attention mechanism?

Scaled dot-product attention is the workhorse of Transformers, but it has three key issues on real-world, non-stationary data:

- **Low-entropy softmax**: attention often collapses to a few positions.
- **Fixed context**: limited adaptability to changing time scales and local/non-local patterns.
- **Weak explainability**: weights encode correlations, not task-aligned feature importance.

- **DAAM**: learns Gaussian means & variances to modulate features instead of dot products.
- **Multi-head**: mixtures of Gaussians approximate arbitrary densities.
- **Parameter-efficient**: 0.002–0.082M parameters per module.
- **Explainable**: produces *Importance Factors* (IF) for each feature and layer.

Core idea: Density Adaptive Attention

For input tensor x (batch $\times$ features), DAAM computes per-head Gaussian gates instead of pairwise dot products.

Per-head transformation:

Compute mean and variance along a chosen axis (e.g. time or channels).

Add learnable mean offset δ and scale variance with learnable ξ.

Normalize: $y_{\text{norm}} = \dfrac{x - (\mu + \delta)}{\sqrt{\sigma^2 + \varepsilon}}$.

Gaussian gate: $g = \exp\left(-\dfrac{y_{\text{norm}}^2}{2c^2}\right)$ with learnable c.

Modulate: $x' = x \odot g$.

DAAM replaces "*where are things similar?*" with *"where is the density unusually informative?"*

Algorithm sketch

Algorithm 1: Density Adaptive Attention

Input: x, normAxis, H heads
for each head $h = 1..H$:
Compute μ_h, σ_h^2 along normAxis
$\mu_h^{\text{adj}} = \mu_h + \delta_h$ (learnable)
$y_h = \dfrac{x - \mu_h^{\text{adj}}}{\sqrt{\sigma_h^2 + \varepsilon}}$
$g_h = \exp\left(-\dfrac{y_h^2}{2c_h^2}\right)$
$x'_h = x \odot g_h$
return concat$(x'_1, \ldots, x'_H)$

Universal approximation

Multi-head DAAM with K Gaussians per head approximates mixtures of Gaussians; with enough heads/components, it can approximate any continuous density. This lets the model match the *true* feature distribution of each modality instead of relying on fixed softmax shape.

Architecture & integration

Plug-in module for frozen encoders

Setup. We keep large pre-trained models *frozen* and add a light DAAM-based decoder:

- **Speech**: WavLM-Large (24 layers, 1024-d).
- **Text**: Llama2-13B (40 layers, 5120-d).
- **Vision**: BEiT-Large (24 layers, 1024-d).

Pipeline.

Mean-pool hidden states from each transformer layer.
Stack pooled vectors into $X \in \mathbb{R}^{N \times d}$.
Apply Multi-Head DAAM (or GQDAAM) across layer dimension.
Feed contextualized representation into a small conv / MLP head.

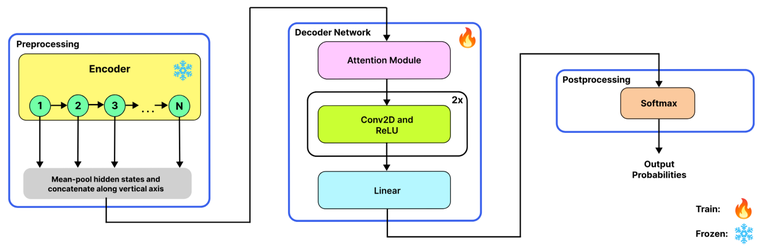

Fig. 1. Frozen encoder, DAAM-based attention module over layer-wise features, and task-specific head.

Parameter cost

For d-dimensional features and H heads:

$$P_{\text{DAAM}} = 2 \times H \times d$$

versus

$$P_{\text{SA}} \approx 4d^2$$

for standard self-attention projection matrices. In practice, DAAM adds **0.016%–0.08%** parameters to GQA and uses **$\sim$80% fewer** parameters than LoRA.

Experimental results

Cross-modal, parameter-efficient gains

Modality	Dataset	Baseline	Best DAAM	Δ
Speech	IEMOCAP	62.3	67.4	+5.1
Vision	CIFAR-100	61.7	80.6	+18.9
Text	AG News	94.6	94.9	+0.3

Explainability via Importance Factors

DAAM yields **Importance Factors (IF)** from its Gaussian gates:

$$\text{IF} = \frac{\text{DA} - \min(\text{DA})}{\max(\text{DA}) - \min(\text{DA})} \in [0, 1]$$

where DA are the density-based attention values.

Importance Factor Heatmaps

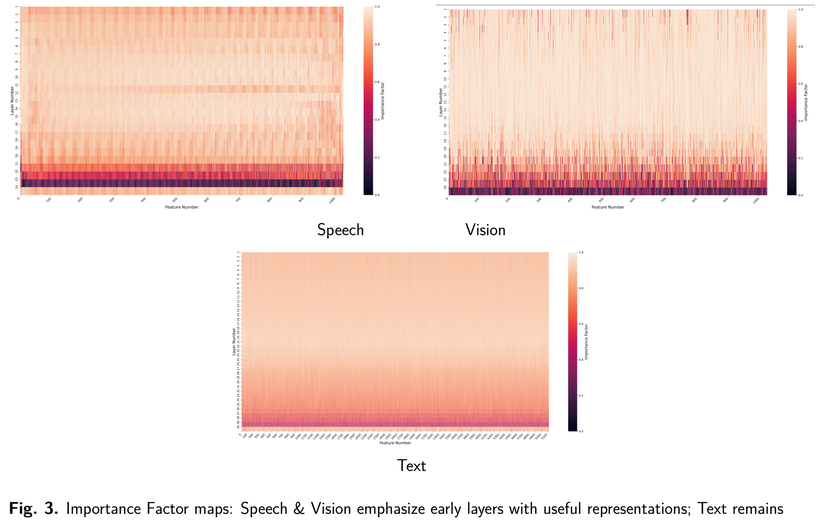

Fig. 3. Importance Factor maps: Speech & Vision emphasize early layers with useful representations; Text remains more uniformly distributed across depth.

Validating IF: ablations

We retrain using only *high-IF* vs *low-IF* layers:

Dataset	Layers	Acc/F1	Δ
IEMOCAP	High-IF (L9)	64.0	+1.7
	Low-IF (L23)	62.3	
CIFAR-100	High-IF (L10–12)	72.6	+7.9
	Low-IF (L22–24)	64.7	
AG News	High-IF (L19–21)	94.9	+0.2
	Low-IF (L37–39)	94.7	

High-IF layers consistently outperform low-IF ones, confirming that IF measures *task-aligned* importance, unlike raw self-attention weights that mainly capture correlation.

Takeaways

- DAAM provides **density-aware** attention that adapts to each modality.
- Gains are largest on **non-stationary** data (speech, vision).
- IF gives a **usable interpretability signal** for pruning and analysis.
- Integration with yields strong performance with **minimal** parameter cost.

Code & repo

GitHub: `https://github.com/gioannides/DAAM-paper-code`