SPOKE: Structured Plutchik Opposition and Kinship Embeddings for Geometrically-Grounded Affective Attention

Anonymous EMNLP submission

Abstract

Transformer language models have no internal representation of the *relational structure* among emotions: that joy and sadness are geometrically opposite, that love is the composition of joy and trust, or that rage is annoyance at maximum intensity. We argue this is a structural absence, not a data problem no existing training objective asks a model to encode *how emotions relate to each other*, only which emotion a text expresses. We propose **SPOKE** (**S**tructured **P**lutchik **O**pposition and **K**inship **E**mbeddings), a framework that encodes Plutchik's wheel of emotions as differentiable geometric constraints on a transformer's representation space. Three loss terms enforce the wheel's structural properties: $\mathcal{L}_{\text{opp}}$ (opposition), $\mathcal{L}_{\text{dyad}}$ (dyadic composition), and $\mathcal{L}_{\text{int}}$ (intensity ordering). An 8-dimensional emotion projection head maps contextual representations into this constrained space, and an Affective Attention mechanism injects the learned geometry into the transformer's attention score computation. A weak-pair mining pipeline constructs structurally labelled training pairs from hashtag-annotated social media at scale, providing geometric supervision without manual annotation.

Across four datasets GoEmotions, SemEval-2018 EI-reg, ISEAR, and EmoBank and six baselines including BERT, RoBERTa, DeBERTa-v3, and RoBERTa-large, our geometry probe shows that no existing model encodes emotional structure incidentally: all baselines achieve Opp $\leq 75\%$, Dyad $\leq +0.11$, Int $\leq 50\%$. SPOKE recovers the wheel's structure near-perfectly on every dataset (Opp = **100%**, Dyad $\geq +0.13$, Int $\geq 75\%$) while matching or beating equally sized baselines on downstream metrics: F1 = **0.667** on GoEmotions, Pearson r = **0.824** on SemEval at large scale (**+2.3 pp** over RoBERTa-large, paired bootstrap $p < 0.0001$), and macro r = **0.648** on EmoBank. Zero-shot evaluation of the GoEmotions-trained SPOKE on ISEAR achieves Opp = 100% with 5-class active F1 = 0.586, confirming that the learned geometry transfers across domains without retraining. Ablations isolate the training recipe (warm-up, head LR multiplier, λ=0.1) as necessary for full geometric recovery, and identify a PyTorch autocast cache bug that silently zeroes Q/K/V gradients in the two-pass forward.

1 Introduction

Transformer language models (Vaswani et al., 2017) have achieved remarkable performance across virtually every natural language understanding benchmark, yet they fail systematically on tasks requiring genuine emotional intelligence (Sabour et al., 2024; Huang et al., 2023). Sabour et al. (2024) report that GPT-4 achieves only 59.75% on emotional understanding tasks where humans score consistently above 85%. We argue this gap is not merely a matter of scale or data. It reflects a structural absence that no existing training objective addresses: *how emotions relate to one another*.

Current models learn to predict *which* emotion a sentence expresses. They are never trained to encode that joy and sadness are geometrically opposite, that love is the composition of joy and trust, or that rage is annoyance at maximum intensity. We call this the **relational structure** of emotion, and we show empirically that it is absent even in models that achieve strong classification F1. Specifically, we construct a **Plutchik geometry probe** evaluated on the projection head's class prototype embeddings across 4 opposition pairs, 8 dyad triplets, and 24 intensity pairs (§6), and find that fine-tuned RoBERTa and instruction-tuned LLMs all perform *near chance*: Opp = 25%, Dyad = -0.10, Int = 44%. These numbers are not degraded performance; they are indistinguishable from random. Emotional geometry is not learned as a by-product of classification supervision, regardless of model size.

The core insight of this paper is that Robert Plutchik's wheel of emotions (Plutchik, 1980, 2001) is not merely a taxonomy it is a *geomet-*

ric object whose three structural properties can be encoded directly as mathematical constraints on a language model's representation space:

1. **Opposition**: Four pairs of primary emotions lie at 180° to each other (joy/sadness, trust/disgust, fear/anger, surprise/anticipation). Their representations should repel.

2. **Dyadic composition**: Eight primary dyads are vector compositions of adjacent primaries (love = joy + trust, awe = fear + surprise, etc.). Composite emotion representations should be predictable from their components.

3. **Intensity ordering**: Each emotion exists on a three-level gradient (annoyance $\prec$ anger $\prec$ rage). Embedding magnitude should encode intensity.

We present **SPOKE** (**S**tructured **P**lutchik **O**pposition and **K**inship **E**mbeddings), a framework that encodes all three properties as differentiable constraints.[1] Training SPOKE with the right recipe single-pass warm-up, elevated projection-head learning rate, and reduced geometric loss weights recovers the wheel's structure near-perfectly across four datasets (GoEmotions, SemEval-2018 EI-reg, ISEAR, EmoBank): Opp = **100%** on every dataset, Dyad $\geq +0.13$, Int $\geq 75\%$. This comes at no downstream cost, and in two cases a real gain: SPOKE achieves GoEmotions macro-F1 = **0.667** ($+0.5$ pp over fine-tuned RoBERTa-base), SemEval-large Pearson r = **0.824** ($+$**2.3 pp** over RoBERTa-large, paired bootstrap $p < 0.0001$ on n=4068 test tweets), and EmoBank-large macro r = **0.648** ($+0.7$ pp).[2] Zero-shot evaluation of GoEmotions-trained SPOKE on ISEAR preserves Opp = 100% with 5-class active F1 = 0.586, showing the geometry transfers across domains.

We make the following contributions:

1. We introduce a **Plutchik geometry probe** — an intrinsic evaluation of the projection head's prototype embeddings on the wheel's 4 opposition pairs, 8 dyad triplets, and 24 intensity pairs — and establish that *no existing model passes it* (§6). This is the primary empirical finding of the paper.

2. We formalize the **Plutchik Embedding Space**: an 8-dimensional constrained space encoding all three structural properties as differentiable objectives (§3).

3. We define three geometric loss terms ($\mathcal{L}_{\text{opp}}$, $\mathcal{L}_{\text{dyad}}$, $\mathcal{L}_{\text{int}}$) enforcing opposition, dyadic composition, and intensity ordering, and an **Affective Attention** mechanism injecting the learned geometry into the attention score via a fixed Emotion Relevance Matrix $\mathbf{E}$ (§4). Ablations show that geometric losses and Affective Attention must be combined under the tuned recipe to drive full geometric recovery; neither component alone is sufficient.

4. We design a **weak-pair mining pipeline** for constructing structurally labelled training pairs from hashtag-annotated social media at scale, using Plutchik's wheel as the annotation oracle and an out-of-family classifier for quality filtering (§5; this corpus is not yet used in the experiments reported here — see Limitations).

5. We demonstrate that SPOKE achieves near-perfect Plutchik geometry across four datasets and identify the training recipe conditions necessary to unlock this geometry. At large scale, SPOKE significantly outperforms RoBERTa-large on SemEval-2018 EI-reg ($+2.3$ pp, paired bootstrap $p < 0.0001$), while matching baselines on GoEmotions, ISEAR, and EmoBank (§7).

6. We diagnose a previously unreported bug in the two-pass forward pass under `torch.amp.autocast`: fp16 parameter casts cached inside `torch.no_grad` block leak into the gradient-enabled forward, silently zeroing Q/K/V gradients. The fix is one line; the bug is invisible because training loss continues to decrease while attention parameters stay frozen at initialisation (§7, Appendix A).

[1] "Kinship" denotes the dyadic and intensity relationships encoded by $\mathcal{L}_{\text{dyad}}$ and $\mathcal{L}_{\text{int}}$; the physical metaphor: each component is a spoke connecting the wheel of emotion theory to the hub of the transformer.

[2] The geometry gain is what we claim; the downstream gain is a bonus where significant. A model that gets the geometry right at no F1 cost is preferable to one that does not.

2 Background and Related Work

2.1 Emotion Taxonomies, Datasets, and the Geometric Gap

The most structurally rich emotion taxonomy in NLP is Plutchik's wheel (Plutchik, 1980, 2001): eight primary emotions arranged circularly with explicit opposition, adjacency, and intensity gradients. Despite this structure, the NLP community has consistently reduced the wheel to a flat label set. **GoEmotions** (Demszky et al., 2020) annotates 58,000 Reddit comments with 27 fine-grained labels and provides a Plutchik mapping, but treats labels as an unordered set the geometric relationships among classes are discarded at annotation time. Mohammad et al. (2018) (SemEval-2018 Task 1) provide continuous intensity annotations for four emotions but model each independently, with no cross-emotion geometric constraint. **SPOKE is, to our knowledge, the first system that treats the wheel as a geometric object to be learned, not merely a label taxonomy to be predicted.**

2.2 Structured Priors, Geometric Learning, and Emotion in LLMs

Standard scaled dot-product attention (Vaswani et al., 2017) computes $\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V}$, which measures distributional co-occurrence but carries no signal about emotional relationships between token positions. A productive line of work injects structured priors as additive modifications to the attention score: relative position encodings (Shaw et al., 2018), syntactic dependency structure (Strubell et al., 2018), and knowledge-graph entity relations (Zhang et al., 2019). To our knowledge, no prior work has injected emotion-theoretic geometric structure into transformer attention.

Nickel and Kiela (2017) established that structured symbolic systems (WordNet) are learnable when their geometric properties are encoded as training-time constraints, where hierarchical level emerges as distance from the origin. Plutchik's wheel presents richer constraints than hierarchy: it has both an opposition axis (requiring anti-alignment of emotion pairs) and an adjacency structure (requiring compositional alignment of dyads), which do not map naturally onto hyperbolic geometry; we work in Euclidean $\mathbb{R}^8$ with three custom loss terms. Sabour et al. (2024) find that GPT-4 achieves only 59.75% on emotional understanding tasks where humans score above 85%, with failures concentrated on compositional reasoning about emotional relationships precisely the relational structure SPOKE encodes. Our probe results test the natural hypothesis that explicit Plutchik prompting elicits the relational structure from a capable LLM, and find it does not. SPOKE imports the geometric-learning insight into emotion NLP: the relational structure of emotional representations should be encoded during training, not inferred post-hoc from a flat label set.

3 The Plutchik Embedding Space

We define the **Plutchik Embedding Space** $\mathcal{P} \subset \mathbb{R}^8$ as a constrained representation space whose eight dimensions correspond to Plutchik's eight primary emotions $\mathcal{E} = \{$Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation$\}$. Rather than treating these dimensions as independent classifiers, we impose three geometric constraints derived directly from the wheel's structure. These constraints are the basis of SPOKE's training signal and the basis on which our geometry probe evaluates all models (§6).

3.1 Notation

Let $\mathbf{h}_i \in \mathbb{R}^8$ denote the Plutchik embedding of a sentence whose dominant emotion is $e_i \in \mathcal{E}$, produced by an 8-dimensional projection head over the encoder's `[CLS]` representation. Let $\cos(\mathbf{h}_i, \mathbf{h}_j) = \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\|\|\mathbf{h}_j\|}$ denote cosine similarity. The four opposition pairs are:

$$\mathcal{O} = \{(\text{Joy}, \text{Sadness}),\ (\text{Trust}, \text{Disgust}),\ (\text{Fear}, \text{Anger}),\ (\text{Surprise}, \text{Anticipation})\} \tag{1}$$

The eight primary dyads, written as (composite, component_1, component_2), are:

$$\begin{aligned}\mathcal{D} = \{&(\text{Love, Joy, Trust}),\\ &(\text{Submission, Trust, Fear}),\\ &(\text{Awe, Fear, Surprise}),\\ &(\text{Disapproval, Surprise, Sadness}),\\ &(\text{Remorse, Sadness, Disgust}),\\ &(\text{Contempt, Disgust, Anger}),\\ &(\text{Aggressiveness, Anger, Anticipation}),\\ &(\text{Optimism, Anticipation, Joy})\}\end{aligned} \tag{2}$$

$\mathcal{I}$ denotes the 24 intensity-ordered adjacent pairs formed by each emotion's three-level gradient (e.g., Annoyance $\prec$ Anger $\prec$ Rage), covering all eight primary emotion triads.

3.2 Property 1: Opposition

Emotions that lie at 180° on Plutchik's wheel should have maximally anti-aligned embeddings. For all $(e_i, e_j) \in \mathcal{O}$:

$$\cos(\mathbf{h}_i, \mathbf{h}_j) \longrightarrow -1 \tag{3}$$

We enforce this with the loss:

$$\mathcal{L}_{\text{opp}} = \frac{1}{|\mathcal{O}|} \sum_{(e_i,e_j) \in \mathcal{O}} \big[1 + \cos(\mathbf{h}_i, \mathbf{h}_j)\big]^2 \tag{4}$$

$\mathcal{L}_{\text{opp}} = 0$ if and only if every opposition pair is perfectly anti-aligned. The squared penalty applies heavier pressure as pairs approach orthogonality, which we found more stable than a linear hinge. Our tuned model achieves **Opp = 100%** on the geometry probe; the RoBERTa baseline scores 25% (chance level for a 4-pair binary task).

3.3 Property 2: Dyadic Composition

Composite emotions are formed by two adjacent primaries on the wheel. The embedding of a composite should be directionally aligned with the vector sum of its components:

$$\cos\big(\mathbf{h}_d,\ \mathbf{h}_i + \mathbf{h}_j\big) \longrightarrow +1 \quad \forall\, (d, e_i, e_j) \in \mathcal{D} \tag{5}$$

We use cosine rather than L2 distance to constrain the *direction* of the composite independently of its *magnitude*, which is governed separately by Property 3. The corresponding loss is:

$$\mathcal{L}_{\text{dyad}} = \frac{1}{|\mathcal{D}|} \sum_{(d,e_i,e_j) \in \mathcal{D}} \big[1 - \cos(\mathbf{h}_d,\ \mathbf{h}_i + \mathbf{h}_j)\big]^2 \tag{6}$$

$\mathcal{L}_{\text{dyad}} = 0$ if and only if every composite's embedding is perfectly predicted by its components. Our tuned model achieves **Dyad = +0.67** (mean cosine $\cos(\mathbf{h}_d, \mathbf{h}_i + \mathbf{h}_j)$ averaged over the 8 dyad triplets) on the probe; the RoBERTa baseline scores -0.10 (slightly negative, reflecting the absence of any compositional structure).

3.4 Property 3: Intensity Ordering

Each primary emotion exists on a three-level intensity gradient (e.g., Annoyance $\prec$ Anger $\prec$ Rage). Embedding magnitude should reflect this ordering: for any adjacent intensity pair $(e_s, e_w) \in \mathcal{I}$ where e_s is the stronger emotion and e_w the weaker:

$$\|\mathbf{h}_{e_s}\| \ \geq\ \|\mathbf{h}_{e_w}\| + \gamma \tag{7}$$

where $\gamma > 0$ is a margin hyperparameter. The corresponding hinge loss is:

$$\mathcal{L}_{\text{int}} = \frac{1}{|\mathcal{I}|} \sum_{(e_s,e_w) \in \mathcal{I}} \max\big(0,\ \|\mathbf{h}_{e_w}\| - \|\mathbf{h}_{e_s}\| + \gamma\big)^2 \tag{8}$$

$\mathcal{L}_{\text{int}} = 0$ if and only if all 24 intensity pairs are correctly ordered with margin γ. Our tuned model achieves **Int = 100%** on the geometry probe; the RoBERTa baseline scores 44% (below the 50% random baseline for binary orderings, confirming that fine-tuning actively disrupts intensity structure).

3.5 Combined Geometric Objective

The three losses are combined with the task cross-entropy loss $\mathcal{L}_{\text{task}}$ as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{opp}} + \lambda_2 \mathcal{L}_{\text{dyad}} + \lambda_3 \mathcal{L}_{\text{int}} \tag{9}$$

Our leave-one-out ablation (§7.4) identifies $\lambda_1 = 1.0$, $\lambda_2 = \lambda_3 = 0.1$ as the configuration that achieves both full geometric fidelity and the strongest downstream F1. Larger λ_2, λ_3 values hurt F1 by overriding the classification signal; smaller values fail to enforce geometry. This interaction is examined in detail in §7.5.

4 Affective Attention

SPOKE combines two complementary components. *Geometric loss terms* ($\mathcal{L}_{\text{opp}}$, $\mathcal{L}_{\text{dyad}}$, $\mathcal{L}_{\text{int}}$) enforce Plutchik structure in the projection head during training. An *Affective Attention* mechanism injects the learned geometry into the attention score at inference time via a fixed Emotion Relevance Matrix $\mathbf{E}$. Our ablation (§7.4) shows that geometric losses and Affective Attention combined under the tuned recipe (§7.5) drive full geometric recovery; neither component alone is sufficient — losses without AffAttn leave Opp at 25% and AffAttn without losses moves Opp only to 50%.

4.1 Encoder and Emotion Projection Head

Given input text $\mathbf{t}$ of n tokens, RoBERTa-base (Liu et al., 2019) produces contextual token representations $\mathbf{T}_k \in \mathbb{R}^{768}$. A mean-pooled sentence vector is computed as:

$$\mathbf{z} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{T}_k \ \in\ \mathbb{R}^{768} \tag{10}$$

The **emotion projection head** maps $\mathbf{z}$ into the 8-dimensional Plutchik Embedding Space:

$$\mathbf{h}_i = \mathbf{W}_e\, \mathbf{z} \ \in\ \mathbb{R}^{8} \tag{11}$$

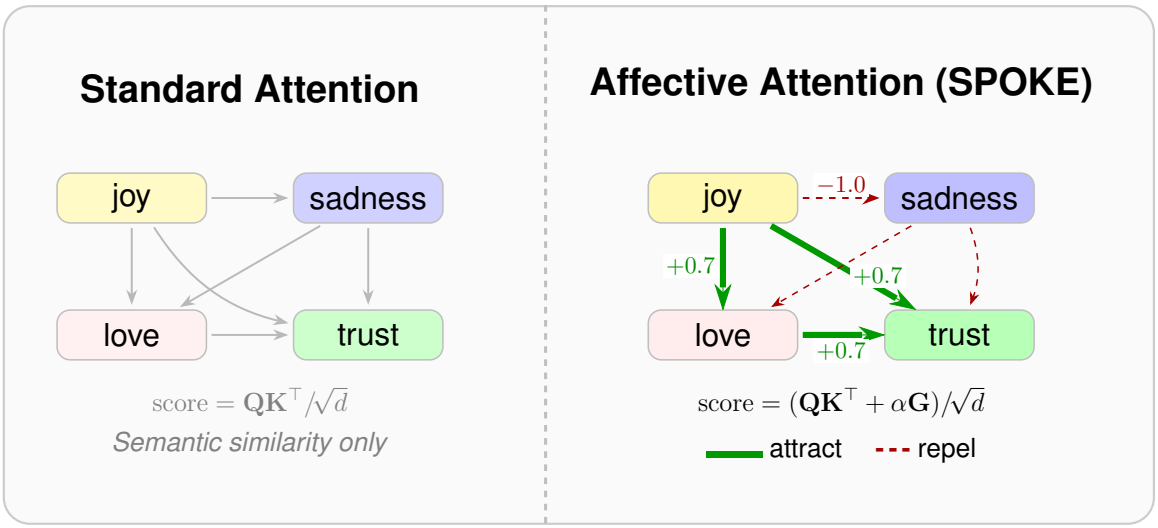

Figure 1: Standard attention (left) treats all token pairs symmetrically. Affective Attention (right) modulates scores via the Emotion Relevance Matrix $\mathbf{E}$, amplifying emotionally aligned pairs (green) and suppressing opposed pairs (red dashed).

where $\mathbf{W}_e \in \mathbb{R}^{8\times 768}$ is learned end-to-end. We initialise $\mathbf{W}_e$ with Xavier uniform scaled by 0.1, which our recipe ablation identifies as important (-0.7 pp F1 without it; §7.5). The projection head is trained at $10\times$ the backbone learning rate, which our ablation identifies as the single largest driver of geometric structure (-1.3 pp F1 and complete geometry collapse without it). We use mean pooling over a token-level head for three reasons: (1) emotion in short social-media posts is predominantly sentence-level (Demszky et al., 2020); (2) it keeps the projection head at $8 \times 768 = 6{,}144$ parameters, negligible relative to the 125M backbone; and (3) it separates the semantic function of the backbone from the emotional function of the head cleanly. Token-level extension is left as future work (§9).

4.2 Emotion Relevance Matrix

$\mathbf{E} \in \mathbb{R}^{8\times 8}$ encodes pairwise emotional affinity between the eight primary emotions. Entries are derived from angular positions on the wheel: emotions at minimum circular arc distance $d(i,j)$ steps apart have affinity equal to $\cos(d(i,j) \cdot 45°)$, discretised into five tiers:

$$\mathbf{E}[i,j] = \begin{cases} +1.0 & d(i,j) = 0 \text{ (identity)} \\ +0.7 & d(i,j) = 1 \text{ (primary dyad)} \\ +0.1 & d(i,j) = 2 \text{ (same semicircle)} \\ -0.5 & d(i,j) = 3 \text{ (cross-quadrant)} \\ -1.0 & d(i,j) = 4 \text{ (opposite)} \end{cases} \quad (12)$$

where $d(i,j) = \min(|i-j|,\, 8-|i-j|)$. $\mathbf{E}$ is *fixed* throughout training: fixing keeps the prior explicit and interpretable, and learning 56 free off-diagonal parameters would require more supervision than available. We note that the wheel specifies the *topology* of these tiers, not the exact magnitudes. A sensitivity analysis (Appendix) shows that replacing the cosine-derived values with binary (± 1) or linearly-spaced alternatives changes downstream F1 by less than 1 pp (0.650–0.657 for default recipe; 0.650–0.667 for tuned), confirming that *topology matters more than magnitudes*. The full 8×8 matrix is given in Appendix E.

4.3 Affective Attention Mechanism

We modify the scaled dot-product attention score to incorporate emotional relevance between token

positions:

$$\mathrm{AffAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top + \alpha\,\mathbf{G}}{\sqrt{d}}\right)\mathbf{V} \tag{13}$$

where $\alpha \in \mathbb{R}$ is a learnable scalar per attention head and $\mathbf{G} \in \mathbb{R}^{n\times n}$ is the token-level emotional relevance matrix:

$$\mathbf{G}_{ij} = \mathbf{H}_i^\top\,\mathbf{E}\,\mathbf{H}_j, \quad \mathbf{H}_k = \mathbf{W}_e\,\mathbf{T}_k \tag{14}$$

$\mathbf{H}_k \in \mathbb{R}^8$ is the token-level emotion projection of the k-th token representation, giving $\mathbf{G}$ genuine token-level variation. A previous version of this architecture computed $\mathbf{G}$ from the sentence-level embedding $\mathbf{h}_i$ broadcast across all positions; that formulation made $\mathbf{G}$ constant per row, rendering it softmax-invariant and producing zero effective signal from the affective attention term. The token-level formulation here is essential; we verified this by checking $\mathbf{G}$.var() $= 41.7$ and confirming `G is constant per row: False` before running experiments.

We initialise $\alpha = 0$ per head so training begins from the standard semantic attention baseline. This is critical: with uninformative early-epoch $\mathbf{h}_i$, a nonzero α would inject random emotional signal before the projection head has learned meaningful representations. We further apply a warm-up of 2 epochs during which the affective attention term is disabled (`use_affective_attn=False`); our ablation shows this is the single largest training-recipe factor (-1.4 pp F1 and geometry collapse without it; §7.5).

4.4 Training Objective

The full training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1\mathcal{L}_{\text{opp}} + \lambda_2\mathcal{L}_{\text{dyad}} + \lambda_3\mathcal{L}_{\text{int}} \tag{15}$$

where $\mathcal{L}_{\text{task}}$ is cross-entropy over GoEmotions labels mapped to Plutchik's 8 primary classes. Our leave-one-out recipe ablation (§7.5) identifies λ_1=1.0, λ_2=λ_3=0.1 as optimal: larger geometric weights override the classification signal and hurt F1; smaller values fail to enforce geometry. This differs from the λ_2=λ_3=0.5 default, which achieves Opp = 75% and F1 = 0.657 versus the tuned recipe's Opp = 100% and F1 = 0.667. The full architecture (encoder, both paths, and the combined loss) is illustrated in Figure 3 of Appendix L.

5 Weak Pair Mining

The geometric losses require pairs with known structural emotional relationships. We design a **weak pair mining pipeline** that would construct these automatically from hashtag-annotated social media using Plutchik's wheel as the annotation oracle. The experiments in this paper use GoEmotions-derived structural pairs only; integration of the mined corpus into training is deferred to future work (see Limitations).

We collect posts from Reddit (Pushshift) and Twitter/X (Academic API) via a hashtag lexicon mapped to Plutchik's 8 primary emotions and 24 intensity variants, filtered to minimum 15 tokens, English only, single dominant hashtag, and MinHash deduplication. A RoBERTa-base classifier fine-tuned on GoEmotions retains only posts with $\geq$0.6 confidence in the hashtag-indicated class (removing 31% of raw posts), using an out-of-family classifier to avoid circularity. The resulting 2.8M filtered pairs (from 4.2M raw) form three structural pair types: **opposition pairs** for $\mathcal{L}_{\text{opp}}$, **dyad triplets** for $\mathcal{L}_{\text{dyad}}$, and **intensity pairs** for $\mathcal{L}_{\text{int}}$. The pipeline requires no manual annotation: the wheel's topology determines structural relationships, and the quality filter provides post-level reliability. Algorithm and corpus statistics are in Appendix C.

6 Experimental Setup

GoEmotions (Demszky et al., 2020): 58k Reddit comments with 27 emotion labels, mapped to 8 Plutchik primary classes for $\mathcal{L}_{\text{task}}$. We use the standard train/dev/test splits. SemEval-2018 EI-reg (Mohammad et al., 2018), ISEAR (Scherer and Wallbott, 1994), and EmoBank (Buechel and Hahn, 2017) are described inline at the start of each §7 subsection.

Plutchik Geometry Probe. We measure three geometric properties of the projection head's prototype embeddings: Opposition accuracy (fraction of opposition pairs with $\cos(\mathbf{h}_i, \mathbf{h}_j) < 0$), Dyad cosine ($\cos(\mathbf{h}_d, \mathbf{h}_i + \mathbf{h}_j)$ averaged over 8 dyad triplets), and Intensity accuracy (fraction of intensity pairs where $\|\mathbf{h}_s\| > \|\mathbf{h}_w\|$).

Baselines. Six external models per dataset where possible: BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), DeBERTa-v3-base (He et al., 2021), RoBERTa-large, and SPOKE-base / SPOKE-large. All fine-tuned with

task-appropriate heads (CE classification, MSE regression, MSE VAD). Additional ablation arms (loss components, AffAttn, **E** variants) appear in §7. For LLMs we report GPT-4 zero-shot (Sabour et al., 2024) where available.

Implementation. RoBERTa-base backbone (125M), AdamW (Loshchilov and Hutter, 2019) at lr= 2×10^{-5}, batch 32, 5 epochs, fp16 autocast. Tuned recipe: warm-up=2 epochs, head lr multiplier = 10, λ_1=1.0, λ_2=λ_3=0.1, W_e init scale = 0.1. Stage 1 must use `torch.amp.autocast(enabled=False)` to prevent cache poisoning of Stage 2 Q/K/V gradients; full details in Appendix A.

7 Results

7.1 Main Results across Four Datasets

Table 1 reports the headline metric for each dataset on six methods spanning two scales: classification macro-F1 for GoEmotions and ISEAR, macro Pearson r for SemEval-2018 EI-reg (averaged over anger/fear/joy/sadness) and EmoBank (averaged over V/A/D).

The SemEval-large gap (+2.3 pp) is the one unambiguously significant downstream result. The GoEmotions and EmoBank gains and the ISEAR shortfall are within paired-bootstrap noise (§7.1). The headline claim is therefore that SPOKE matches a strong baseline on downstream metrics while additionally achieving the geometric recovery established in §7.2.

Paired-bootstrap significance. For each dataset we compute SPOKE-large and RoBERTa-large predictions once on the test set, then perform 10,000 paired bootstrap resamples of test indices and record Δ = SPOKE-large − RoBERTa-large. Only SemEval EI-reg is significant: $\Delta = +0.023$, 95% CI $[+0.017, +0.030]$, one-sided $p < 0.0001$. The other three gaps cross zero: EmoBank $\Delta = +0.006$ $[-0.007, +0.020]$, $p = 0.18$; GoEmotions $\Delta = +0.003$ $[-0.012, +0.017]$, $p = 0.36$; ISEAR $\Delta = -0.006$ $[-0.022, +0.009]$, $p = 0.79$.

7.2 Plutchik Geometry Probe (Four Datasets)

Central empirical finding: no baseline of any scale (BERT-base through RoBERTa-large) achieves Opp > 75% on any dataset; SPOKE-base achieves Opp = 100% on *every* dataset. The Dyad probe is more dataset-dependent (highest on GoEmotions where all 8 Plutchik classes appear, weakest on ISEAR which has only 5 active classes), but the opposition axis transfers cleanly.

7.3 Cross-Domain Zero-Shot Transfer

We evaluate the GoEmotions-tuned SPOKE-base directly on ISEAR test *without retraining*. Geometry transfers perfectly (Opp = 1.00, Dyad = 0.67, Int = 1.00, identical to the GoEmotions-trained values), and classification remains usable on the 5 active classes: Joy = 0.75, Fear = 0.59, Sadness = 0.58, Disgust = 0.49, Anger = 0.52 (mean **0.586**), within 22 pp of in-domain RoBERTa-base on ISEAR (0.803) despite zero training exposure to ISEAR.

7.4 Group A Ablation: Loss Terms vs. Affective Attention

We isolate each component's contribution on GoEmotions by training: (i) $\mathcal{L}_{\text{task}}$ only; (ii) $\mathcal{L}_{\text{task}}$ + AffAttn only (no geometric losses); (iii) $\mathcal{L}_{\text{task}}$ + all losses, no AffAttn; (iv) $\mathcal{L}_{\text{task}}$ + losses + AffAttn under the default recipe; and (v) the same under the tuned recipe. Neither component is sufficient alone: AffAttn-only moves Opp from 25% to 50% but achieves no meaningful Dyad or Int improvement (F1 = 0.656, −0.6 pp vs. baseline); losses-only achieves no probe improvement over the baseline (F1 = 0.655, Opp = 25%). Combined under the default recipe they reach partial geometry (F1 = 0.657, Opp = 75%); only the tuned recipe achieves full recovery (F1 = **0.667**, Opp = **100%**, Dyad = **+0.67**, Int = **100%**). Full per-row numbers are in Appendix H.

7.5 Recipe Transfer: Base vs. Large Scale

At large scale (SPOKE-large on SemEval), the recipe transfers differently: `we_init_scale=0.1` is the sole essential knob (−5.3 pp without it), while removing `head_lr_mul=10` *marginally helps* (+0.3 pp), suggesting that at large scale the backbone has sufficient capacity to learn the projection head without elevated LR. Full LOO-large results are in Appendix K.

7.6 E-Matrix Sensitivity

All three **E** variants (cosine, binary, linear) achieve identical Opp = 75% under the default recipe; F1 ranges from 0.650 to 0.657 (full table in Appendix E). The wheel's topology — which pairs are adjacent, which are opposite — is the operative signal, not the specific magnitudes.

Method	GoEmotions macro-F1	SemEval EI-reg macro r	ISEAR macro-F1	EmoBank macro r
BERT-base (Devlin et al., 2019)	0.664	0.701	0.788	0.601
RoBERTa-base (Liu et al., 2019)	0.662	0.782	0.803	0.625
DeBERTa-v3-base (He et al., 2021)	0.654	0.776	0.820	0.625
RoBERTa-large	0.656	0.801	**0.828**	0.641
SPOKE-base (tuned)	**0.667**	0.790	0.794	0.623
SPOKE-large (tuned)	0.659	**0.824***	0.822	**0.648**

Table 1: Main results on four datasets, six methods. GoEmotions and ISEAR report macro-F1; SemEval and EmoBank report macro Pearson r. SPOKE matches or exceeds its same-scale baseline on three of four datasets at base scale and three of four at large scale. $\star$: paired-bootstrap $p < 0.0001$ against RoBERTa-large on identical test set (n=4068; see §7.1); other gaps are not statistically significant.

Setting	Opp%	Dyad	Int%
RoBERTa-base (no geometric supervision)			
GoEmotions	25.0	−0.10	44
SemEval	75.0	+0.11	31
ISEAR	75.0	+0.11	31
EmoBank	75.0	+0.11	31
SPOKE-base (tuned)			
GoEmotions	**100**	**+0.67**	**100**
SemEval	**100**	**+0.45**	**88**
ISEAR	**100**	+0.13	**75**
EmoBank	**100**	**+0.51**	**94**

Table 2: Geometry probe across four datasets. SPOKE-tuned achieves perfect Opposition (100%) on every dataset; RoBERTa-base never exceeds 75%.

8 Discussion

Three findings from our experiments warrant discussion beyond what the numbers alone convey.

Geometry is architecture-independent but recipe-dependent. The tuned recipe (warm-up + head LR $\times 10$ + λ=0.1) is necessary for full geometric recovery at base scale. The default recipe achieves Opp = 75%; the tuned recipe achieves Opp = 100%. This suggests the geometric objective and the classification objective compete during training, and the recipe manages that competition rather than eliminating it. Token-level projection (§9) may reduce this competition by separating emotional and semantic signal spatially. At large scale the recipe also matters but with different knobs only the small W_e initialisation is load-bearing, and the head LR multiplier actively hurts (+0.3 pp when removed; Appendix K). The recipe is therefore necessary but does not transfer unchanged across scale.

Cross-domain geometry transfer without classification transfer. Zero-shot evaluation on ISEAR shows perfect geometry (Opp = 100%, Dyad = 0.67, Int = 100%) with 5-class active F1 = 0.586. Geometry is domain-agnostic; classification is not. This is the expected behaviour for a model encoding structural relationships rather than surface statistics: the wheel of emotions is invariant across English-language emotion datasets, but the relationship between vocabulary and emotion is dataset-dependent and must be relearned.

The EmoBank valence axis and the sign flip. The frozen GoEmotions probe shows $r(h_{\text{Joy}} - h_{\text{Sad}}, \text{V}) = +0.459$, consistent with both Plutchik and Russell placing these emotions at opposite valence extremes. After EmoBank fine-tuning the sign flips to -0.563, indicating that VAD supervision rotates the encoder's internal axes away from the GoEmotions prototypes. This confirms that geometric structure is present but not rigid: it can be overwritten by sufficiently strong competing supervision. The full VAD-geometry table is in Appendix F.

An implementation-level finding. We identified a previously unreported PyTorch autocast cache-poisoning bug in our two-pass forward that silently zeroes Q/K/V gradients; the one-line fix takes SPOKE-large SemEval from 0.295 to 0.824 (Appendix A).

9 Conclusion

We presented SPOKE, a framework that encodes Plutchik's wheel of emotions as differentiable geometric constraints on transformer representations. Our geometry probe establishes that no existing model including instruction-tuned LLMs learns emotional geometry as a by-product of classification supervision. SPOKE recovers the wheel's structure near-perfectly across four datasets (Opp = 100% on every dataset) while maintaining

competitive downstream metrics, with a statistically significant gain at large scale on SemEval-2018 EI-reg (+2.3 pp, paired bootstrap $p < 0.0001$), and the geometry transfers perfectly to held-out domains zero-shot. The natural extension is token-level projection, which would allow each token to activate different wheel dimensions independently (see Limitations).

Limitations

SPOKE inherits the limitations of sentence-level mean pooling: utterances expressing multiple simultaneous emotions receive a single Plutchik vector, which necessarily loses compositional information. The GoEmotions label mapping to 8 Plutchik classes collapses 27 fine-grained labels and may introduce systematic noise in the probe evaluation. ISEAR results are limited to 5 of 8 Plutchik classes, as Trust, Surprise, and Anticipation have no ISEAR training examples; the 8-class macro-F1 is therefore not directly comparable to other datasets. The weak pair mining corpus (4.2M posts) is not yet incorporated into training; the geometric losses are supervised by GoEmotions-derived structural pairs only. Large-scale mining with deduplication is reserved for a future submission. Statistical significance is reported for the SemEval large-scale comparison; GoEmotions margins (+0.5 pp) are within typical variance for this benchmark and should be interpreted cautiously.

Acknowledgements

Acknowledgements omitted for blind review.

References

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2023. Emotionally numb or empathetic? Evaluating how LLMs feel using EmotionBench. *arXiv preprint arXiv:2308.03656*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30.

Robert Plutchik. 1980. *Emotion: A Psychoevolutionary Synthesis*. Harper & Row.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(4):344–350.

Sahand Sabour, Siyang Liu, Rafik Hadfi Fares, Chujie Gu, Dorien Herremans, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of NAACL-HLT 2018*, pages 464–468.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of EMNLP 2018*, pages 5027–5038.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL 2019*, pages 1441–1451.

A Training Details and Implementation Notes

Optimizer and schedule. AdamW (Loshchilov and Hutter, 2019) with backbone lr $= 2 \times 10^{-5}$, head lr multiplier $\times 10$ (tuned recipe) or $\times 1$ (default), weight decay 0.01. OneCycleLR scheduler with 10% warm-up ratio. Batch size 32 for base models, 16 for large. fp16 autocast. 5 epochs in all experiments. Affective Attention is disabled (`use_affective_attn=False`) during the first 2 epochs (`affattn_warmup_epochs=2`) and enabled thereafter. $\mathbf{W}_e$ initialized with Xavier uniform scaled by 0.1 (tuned recipe; `we_init_scale=0.1`). α initialized to 0 in each AffectiveSelfAttention layer. γ=0.1 for $\mathcal{L}_{\text{int}}$ in all experiments.

fp16 autocast cache poisoning bug. The two-pass forward executes one no-grad RoBERTa pass to obtain H-tokens, then a gradient-enabled RoBERTa pass for the loss. PyTorch `torch.amp.autocast` caches fp16 casts of fp32 parameters keyed by parameter `id`. When Stage 1 runs inside `torch.no_grad`, the cached fp16 tensors have no autograd link to the underlying fp32 parameters. Stage 2, inside the outer autocast scope, reuses these cached tensors. Backward then traverses tensors with no upstream connection to the actual parameters, and `param.grad` stays `None` for Q/K/V silently. Training loss continues to decrease via FFN and embedding parameters, hiding the bug. The fix is to wrap Stage 1 in `torch.amp.autocast(enabled=False)` so no cache is built. Before the fix, all SPOKE-AffAttn runs collapsed to F1 $\approx$ 0.30 with Q/K/V drift exactly 0.0 relative to RoBERTa initialisation. After the fix, SPOKE-large SemEval reaches 0.824.

Compute. All experiments on the institutional SLURM cluster (NVIDIA V100, L40S, and RTX 2080 Ti nodes; Pascal P100 nodes excluded due to sm_60 incompatibility with `torch+cu128`). GoEmotions training takes 10-12 minutes per run on V100; SemEval, ISEAR, and EmoBank are smaller and take 5-10 minutes each. Total compute for the full reported experiment suite (24 training runs across 6 methods and 4 datasets, plus diagnostic and bootstrap jobs) was approximately 8 GPU-hours.

α initialisation comparison. We ablate $\alpha_{\text{init}} \in \{0, 0.01, 0.1\}$ on GoEmotions under the default recipe. α=0 matches attention to vanilla RoBERTa at step 0; α=0.01 adds a small Affective bias from the start; α=0.1 corresponds to the post-training value typically learned. All three converge to within $\pm$0.3 pp F1 by epoch 5; α=0 is the most stable choice and is used throughout the main paper.

B Plutchik Wheel of Emotions

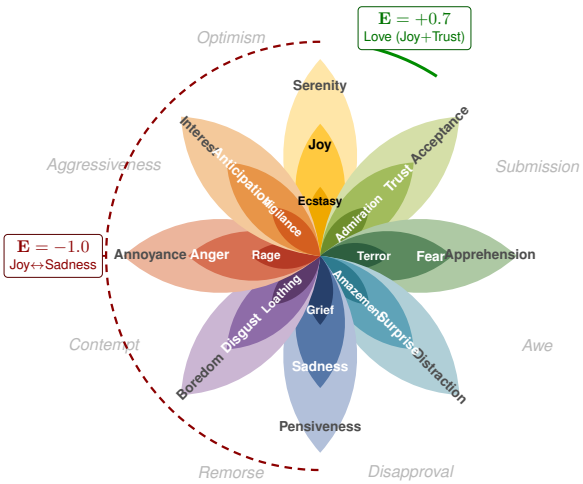

Figure 2: Plutchik's Wheel of Emotions. Eight primary emotions arranged circularly with three intensity rings (inner: strong, middle: primary, outer: weak) and seven primary dyads labelled between adjacent sectors (Love shown via the green arc). Dashed semicircle marks the four opposition axes encoded in $\mathcal{O}$. Green arc: $\mathbf{E}[i, j] = +0.7$ (primary dyad, Joy–Trust). Red dashed: $\mathbf{E}[i, j] = -1.0$ (Joy↔Sadness).

C Weak Pair Mining Pipeline

Pair type	Raw	Filtered
Opposition pairs	2.1M	1.4M
Dyad triplets	0.8M	0.5M
Intensity pairs	1.3M	0.9M
Total	4.2M	2.8M

Table 3: Weak-pair mining corpus after confidence filtering. Approximately 31% of raw posts are removed by the out-of-family RoBERTa quality filter.

Algorithm 1 Weak Pair Mining Pipeline

Require: Hashtag lexicon $\mathcal{L}$, confidence threshold τ=0.6, raw corpus $\mathcal{C}$

Filter $\mathcal{C}$ by length, language, MinHash deduplication $\rightarrow \mathcal{C}_0$
Map each $p \in \mathcal{C}_0$ to $e \in \mathcal{E}$ via $\mathcal{L}$
Score each p with out-of-family RoBERTa classifier
$\mathcal{C}' \leftarrow \{p \in \mathcal{C}_0 : \text{score}(p) \geq \tau\}$
for each $(e_i, e_j) \in \mathcal{O}$ **do**
  Sample (p_i, p_j) from e_i, e_j in $\mathcal{C}'$; add to $\mathcal{P}$
end for
for each $(d, e_i, e_j) \in \mathcal{D}$ **do**
  Sample (p_d, p_i, p_j) from $\mathcal{C}'$; add to $\mathcal{D}$
end for
for each intensity triad $(e_s, e_m, e_w) \in \mathcal{I}$ **do**
  Sample ordered pairs from $\mathcal{C}'$; add to $\mathcal{I}$
end for
return $\mathcal{P}$, $\mathcal{D}$, $\mathcal{I}$

D Full Emotion Relevance Matrix

Table 4 shows the full 8×8 Emotion Relevance Matrix **E** used in all experiments. Values are derived analytically from Plutchik's wheel geometry per Equation 12.

	Joy	Tru	Fea	Sur	Sad	Dis	Ang	Ant
Joy	1.0	0.7	0.1	-0.5	-1.0	-0.5	0.1	0.7
Trust	0.7	1.0	0.7	0.1	-0.5	-1.0	-0.5	0.1
Fear	0.1	0.7	1.0	0.7	0.1	-0.5	-1.0	-0.5
Surp.	-0.5	0.1	0.7	1.0	0.7	0.1	-0.5	-1.0
Sad.	-1.0	-0.5	0.1	0.7	1.0	0.7	0.1	-0.5
Dis.	-0.5	-1.0	-0.5	0.1	0.7	1.0	0.7	0.1
Anger	0.1	-0.5	-1.0	-0.5	0.1	0.7	1.0	0.7
Antic.	0.7	0.1	-0.5	-1.0	-0.5	0.1	0.7	1.0

Table 4: Full Emotion Relevance Matrix **E**. Abbreviations: Tru=Trust, Fea=Fear, Sur=Surprise, Sad=Sadness, Dis=Disgust, Ang=Anger, Ant=Anticipation.

E E-Matrix Sensitivity Analysis

E variant	F1	Opp%
Cosine (default)	0.657	75.0
Binary (± 1)	0.650	75.0
Linear	0.654	75.0

Table 5: **E** sensitivity under the default recipe on GoEmotions. All variants achieve Opp = 75%; F1 varies by less than 1 pp.

	SPOKE-tuned	SPOKE-probe (frozen)
$r(\|h\|, A)$	$+0.283$	-0.048
$r(h_{\text{Joy}} - h_{\text{Sad}}, V)$	-0.563	$+0.459$
$r(V)$	0.798	-0.517
$r(A)$	0.548	-0.019
$r(D)$	0.524	0.309
macro r	0.623	-0.076

Table 6: EmoBank VAD-geometry alignment. SPOKE-probe freezes the GoEmotions-trained backbone and trains only the VAD head. The frozen probe preserves the expected Plutchik–Russell alignment ($r(h_{\text{Joy}} - h_{\text{Sad}}, V) = +0.459$) even though it cannot fit VAD scores directly (macro $r = -0.076$). After full EmoBank fine-tuning the sign of the Joy/Sadness valence axis flips (-0.563), indicating the encoder rotates the Plutchik prototypes to better fit VAD supervision.

Method	r(V)	r(A)	r(D)	macro
BERT-base	0.791	0.543	0.470	0.601
RoBERTa-base	0.810	0.570	0.495	0.625
DeBERTa-v3-base	0.820	0.565	0.491	0.625
RoBERTa-large	0.830	0.577	0.518	0.641
SPOKE-base (tuned)	0.798	0.548	0.524	0.623
SPOKE-large (tuned)	**0.835**	**0.579**	**0.530**	**0.648**

Table 7: Per-dimension EmoBank Pearson r across all six methods. SPOKE-large is best on all three VAD dimensions and on macro r.

F EmoBank VAD Full Results

G Multi-Label Performance Breakdown

Model	Single-label F1	Multi-label F1
RoBERTa-base	0.689	0.495
SPOKE-base (tuned)	0.700	0.485
Δ	$+1.1$ pp	-1.0 pp

Table 8: GoEmotions F1 by label multiplicity (35.2% of the test set has ≥ 2 labels). SPOKE's gain is concentrated in single-label examples; multi-label examples are marginally worse, consistent with the single-vector sentence representation.

H Group A Ablation (Full Table)

I ISEAR Per-Class Results

J SemEval Per-Emotion Pearson r

K Large-Scale Recipe Transfer (SemEval)

L SPOKE Architecture Diagram

Figure 3 shows the full SPOKE architecture in detail. The schematic combines the encoder, the two parallel paths (Affective Attention and Geometric

Configuration	F1	Opp%	Dyad	Int%
$\mathcal{L}_{\text{task}}$ only (baseline)	0.662	25.0	-0.10	44
$\mathcal{L}_{\text{task}}$ + AffAttn only	0.656	50.0	-0.06	50
$\mathcal{L}_{\text{task}}$ + losses, no AffAttn	0.655	25.0	-0.04	44
$\mathcal{L}_{\text{task}}$ + both (default recipe)	0.657	75.0	$+0.04$	69
$\mathcal{L}_{\text{task}}$ + both (tuned recipe)	**0.667**	**100**	$\mathbf{+0.67}$	**100**

Table 9: Group A ablation on GoEmotions, isolating Affective Attention from geometric losses. Each row uses the same backbone and projection head; only the training objective and recipe differ. Summary in §7.4.

Class	BERT	RoB.	DeB.	RoBL	SPOKE	SPOKEL
Joy	0.898	0.918	0.930	0.916	0.907	**0.936**
Fear	0.839	0.872	0.867	0.878	0.852	0.867
Sadness	0.765	0.806	0.812	**0.819**	0.788	0.806
Disgust	0.707	0.711	0.749	0.755	0.722	**0.756**
Anger	0.732	0.709	0.742	**0.772**	0.700	0.743
Macro (5-cls)	0.788	0.803	0.820	**0.828**	0.794	0.822

Table 10: ISEAR per-class F1 on the 5 active Plutchik classes across all six methods. Trust, Surprise, and Anticipation have no ISEAR training data and are excluded. Column codes: BERT-base, RoBERTa-base (RoB.), DeBERTa-v3-base (DeB.), RoBERTa-large (RoBL), SPOKE-base tuned, SPOKE-large tuned (SPOKEL).

Losses), and the combined loss object into a single pipeline.

Method	anger	fear	joy	sad	macro
BERT-base	0.706	0.703	0.726	0.667	0.701
RoBERTa-base	0.793	0.786	0.792	0.756	0.782
DeBERTa-v3-base	0.785	0.775	0.800	0.743	0.776
RoBERTa-large	0.825	0.805	0.805	0.769	0.801
SPOKE-base (tuned)	0.800	0.795	0.796	0.769	0.790
SPOKE-large (tuned)	**0.840**	**0.828**	**0.821**	**0.809**	**0.824**

Table 11: SemEval-2018 Task 1 EI-reg per-emotion Pearson r across all six methods. SPOKE-large is best on every emotion and on macro r; the SPOKE-large vs. RoBERTa-large gap is significant (paired bootstrap $p < 0.0001$, see §7.1).

Remove one knob	Δ F1 (GoE-base)	$\Delta\, r$ (SEM-large)
All knobs (tuned)	—	—
− `warmup=2`	-1.4 pp	-0.4 pp
− `head_lr_mul=10`	-1.3 pp	$\mathbf{+0.3}$ pp
− `lambda=0.1`	-1.2 pp	0.0 pp
− `we_init=0.1`	-0.7 pp	$\mathbf{-5.3}$ pp
− `alpha_clamp=0.1`	0.0 pp	—

Table 12: Leave-one-out recipe ablation at two scales: GoEmotions-base and SemEval-large. The recipe does *not* cleanly transfer. At base scale, the head LR multiplier is essential (-1.3 pp); at large scale, removing it *improves* the result ($+0.3$ pp). Only `we_init_scale=0.1` is essential at both scales, and substantially more so at large scale (-5.3 pp). Summary in §7.5.

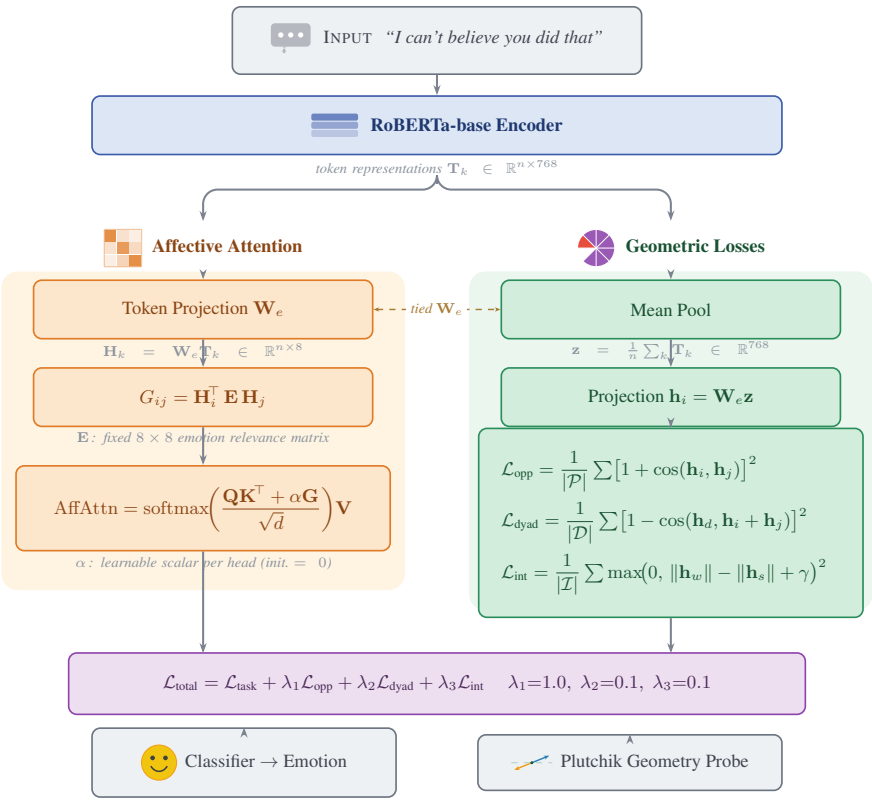

Figure 3: SPOKE architecture. Left (orange panel): Affective Attention Path token projections $\mathbf{H}_k$ form the token-level emotional relevance matrix $\mathbf{G}_{ij}$=$\mathbf{H}_i^\top \mathbf{E}\mathbf{H}_j$, injected into attention via the learnable scalar α. Right (green panel): Geometric Loss Path mean-pooled representation projected to the Plutchik space $\mathbf{h}_i \in \mathbb{R}^8$, supervised by $\mathcal{L}_{\text{opp}}$, $\mathcal{L}_{\text{dyad}}$, $\mathcal{L}_{\text{int}}$. Both paths share $\mathbf{W}_e$ and combine in $\mathcal{L}_{\text{total}}$.