

AIFS Cross-Attention Mechanism Detail

TimeSeries Tokens ↔ Llama 3-8B Embeddings Fusion

Input Token Representations

AIFS TimeSeries Tokens
 $X_{climate} = [B, 64, 512]$
 $B=2, seq_len=64, d_model=512$
From 5D climate data

Llama Text Tokens
 $X_{text} = [B, 128, 4096]$
 $B=2, seq_len=128, d_model=4096$
From climate queries

Dimension Alignment Layer

Climate Projector
 $W_c: 512 \rightarrow 4096$
Linear(512, 4096) + LayerNorm

Projected Climate
 $X'_{climate} = [B, 64, 4096]$
Aligned with Llama dim

Climate tokens
projected to
Llama dimension

Multi-Head Cross-Attention Computation

Query Projection
 $Q = X_{text} \cdot W_Q$
 $Q: [B, 128, 4096]$

Key Projection
 $K = X'_{climate} \cdot W_K$
 $K: [B, 64, 4096]$

Value Projection
 $V = X'_{climate} \cdot W_V$
 $V: [B, 64, 4096]$

Multi-Head Split
32 heads × 128 dim each
 Q_h, K_h, V_h per head

Attention Computation
 $A_h = \text{softmax}(Q_h K_h^T / \sqrt{d_k})$
Per-head attention weights

32 attention heads
parallel computation

Mathematical Formulation

Multi-Head Cross-Attention:

- Projection: $Q = X_{text} W_Q, K = V = X'_{climate} W_{K,V}$
- Multi-Head: $Q_h = Q W_h^Q, K_h = K W_h^K, V_h = V W_h^V$
- Attention: $A_h = \text{softmax}(Q_h K_h^T / \sqrt{d_k})$
- Output: $O_h = A_h V_h$
- Concatenate: $O = \text{Concat}(O_1, \dots, O_{32}) W_O$
- Residual: $Y = \text{LayerNorm}(X_{text} + O)$

Attention Matrix
 $A: [B, 32, 128, 64]$
Text pos × Climate pos
Temperature-scaled
 $\tau = 0.1$ (learnable)

Cross-modal
attention matrix

AIFS Cross-Attention Specifications:

- Input Dimensions:
 - Climate: $[B, 64, 512] \rightarrow$ projected to $[B, 64, 4096]$
 - Text: $[B, 128, 4096]$ (native Llama dimension)
- Multi-Head Configuration:
 - Heads: 32 (same as Llama 3-8B)
 - Per-head dimension: 128 ($4096 \div 32$)
 - Total parameters: ~67M for attention layers
- Attention Mechanism:
 - Query: from text embeddings
 - Key/Value: from projected climate embeddings
 - Temperature scaling: learnable $\tau \in [0.01, 1.0]$
 - Dropout: 0.1 during training
- Performance:
 - FLOPs: ~2.1T per forward pass
 - Memory: ~4.2GB for batch_size=2
 - Latency: ~45ms on A100 GPU

Output Fusion & Integration

Head Concatenation
 $\text{Concat}(O_1, \dots, O_{32})$
 $[B, 128, 4096]$

Output Projection
 $W_O: 4096 \rightarrow 4096$
Linear + Dropout

Fused Embeddings
 $Y = [B, 128, 4096]$
Text enhanced with climate context
Ready for Llama decoder