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

Climate tokens projected to Llama dimension

Dimension Alignment Layer

Climate Projector
W_c: 512 → 4096
Linear(512, 4096) + LayerNorm

Projected Climate
X'_climate = [B, 64, 4096]
Aligned with Llama dim

32 attention heads parallel computation

Multi-Head Cross-Attention Computation

Query Projection
Q = X_text · W_Q
Q: [B, 128, 4096]

Key Projection
K = X'_climate · W_K
K: [B, 64, 4096]

Value Projection
V = X'_climate · 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 = softmax(Q_h K_h^T / √d_k)
Per-head attention weights

Mathematical Formulation

Multi-Head Cross-Attention:

1. Projection: Q = X_text W_Q, K = V = X'_climate W_K,V

2. Multi-Head: Q_h = Q W_h^Q, K_h = K W_h^K, V_h = V W_h^V

3. Attention: A_h = softmax(Q_h K_h^T / √(d_k))

4. Output: O_h = A_h V_h

5. Concatenate: O = Concat(O_1, ..., O_32) W_O

6. Residual: Y = LayerNorm(X_text + O)

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

Cross-modal attention matrix

Output Fusion & Integration

Head Concatenation
Concat(O_1,...,O_32)
[B, 128, 4096]

Output Projection
W_O: 4096 → 4096
Linear + Dropout

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

AIFS Cross-Attention Specifications:

• Input Dimensions:
- Climate: [B, 64, 512] → 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 ÷ 32)
- Total parameters: ~67M for attention layers

• Attention Mechanism:
- Query: from text embeddings
- Key/Value: from projected climate embeddings
- Temperature scaling: learnable τ ∈ [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