KTF-Transformer Detailed Architecture with Mathematical Formulas

Kinodynamic Trajectory Forecasting Transformer

Agent Embedding

$\mathbf{A} \in \mathbb{R}^{T \times 4}$

$\mathbf{E}_A = \text{Dropout}(\text{LN}(\text{Linear}(\mathbf{A})))$

$\mathbf{a}_t = [x_t, y_t, v_{x,t}, v_{y,t}]^T$

$T \times 4 \rightarrow T \times 256$

Map Embedding

$\mathbf{Map} \in \mathbb{R}^{N_m \times 2}$

$\mathbf{E}_M = \text{Dropout}(\text{LN}(\text{Linear}(\mathbf{Map})))$

$\mathbf{m}_i = [x_i, y_i]^T$ (polyline points)

$N_m \times 2 \rightarrow N_m \times 256$

Context Projection

$\mathbf{z}_{MMF} \in \mathbb{R}^{256}$

$\mathbf{E}_C = \text{Dropout}(\text{GELU}(\text{LN}(\text{Linear}(\mathbf{z}_{MMF}))))$

From MMF-GNN output

$256 \rightarrow 256$

Positional Encoding

$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}), PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$

KTF-Transformer Specifications:

- Parameters: $\theta_{KTF} = 5.3M$ (33.7%)
- Model Dimension: $d_{model} = 256$
- Encoder Layers: $L_{enc} = 4$
- Decoder Layers: $L_{dec} = 4$
- Attention Heads: $h = 8$
- FFN Dimension: $d_{ff} = 1024$
- Max Sequence: $N_{max} = 512$
- Prediction Horizon: $H = 30$ steps
- Trajectory Modes: $M = 6$
- Time Step: $\Delta t = 0.1s$
- Physics Constraints: 4 types
- Dropout: $p = 0.2$ (all layers)

Encoder Layer 1

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_{l-1}^{enc} + \text{MHA}(\mathbf{H}_{l-1}^{enc}))$

$\text{MHA}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$

$\text{head}_i = \text{Attention}(\mathbf{H}\mathbf{W}_i^Q, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V)$

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_l^{enc} + \text{FFN}(\mathbf{H}_l^{enc}))$

$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

Encoder Layer 2

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_{l-1}^{enc} + \text{MHA}(\mathbf{H}_{l-1}^{enc}))$

$\text{MHA}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$

$\text{head}_i = \text{Attention}(\mathbf{H}\mathbf{W}_i^Q, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V)$

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_l^{enc} + \text{FFN}(\mathbf{H}_l^{enc}))$

$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

Encoder Layer 3

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_{l-1}^{enc} + \text{MHA}(\mathbf{H}_{l-1}^{enc}))$

$\text{MHA}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$

$\text{head}_i = \text{Attention}(\mathbf{H}\mathbf{W}_i^Q, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V)$

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_l^{enc} + \text{FFN}(\mathbf{H}_l^{enc}))$

$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

Encoder Layer 4

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_{l-1}^{enc} + \text{MHA}(\mathbf{H}_{l-1}^{enc}))$

$\text{MHA}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O$

$\text{head}_i = \text{Attention}(\mathbf{H}\mathbf{W}_i^Q, \mathbf{H}\mathbf{W}_i^K, \mathbf{H}\mathbf{W}_i^V)$

$\mathbf{H}_l^{enc} = \text{LN}(\mathbf{H}_l^{enc} + \text{FFN}(\mathbf{H}_l^{enc}))$

$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

Decoder Layer 1

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_{l-1}^{dec} + \text{Masked-MHA}(\mathbf{H}_{l-1}^{dec}))$

$\text{Mask}_{i,j} = 0 \text{ if } j \leq i, -\infty \text{ if } j > i$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{Cross-Attn}(\mathbf{H}_l^{dec}, \mathbf{H}_L^{enc}))$

$\text{Cross-Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{FFN}(\mathbf{H}_l^{dec}))$

with Dropout($p = 0.2$) in all sub-layers

Decoder Layer 2

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_{l-1}^{dec} + \text{Masked-MHA}(\mathbf{H}_{l-1}^{dec}))$

$\text{Mask}_{i,j} = 0 \text{ if } j \leq i, -\infty \text{ if } j > i$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{Cross-Attn}(\mathbf{H}_l^{dec}, \mathbf{H}_L^{enc}))$

$\text{Cross-Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{FFN}(\mathbf{H}_l^{dec}))$

with Dropout($p = 0.2$) in all sub-layers

Decoder Layer 3

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_{l-1}^{dec} + \text{Masked-MHA}(\mathbf{H}_{l-1}^{dec}))$

$\text{Mask}_{i,j} = 0 \text{ if } j \leq i, -\infty \text{ if } j > i$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{Cross-Attn}(\mathbf{H}_l^{dec}, \mathbf{H}_L^{enc}))$

$\text{Cross-Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{FFN}(\mathbf{H}_l^{dec}))$

with Dropout($p = 0.2$) in all sub-layers

Decoder Layer 4

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_{l-1}^{dec} + \text{Masked-MHA}(\mathbf{H}_{l-1}^{dec}))$

$\text{Mask}_{i,j} = 0 \text{ if } j \leq i, -\infty \text{ if } j > i$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{Cross-Attn}(\mathbf{H}_l^{dec}, \mathbf{H}_L^{enc}))$

$\text{Cross-Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\mathbf{H}_l^{dec} = \text{LN}(\mathbf{H}_l^{dec} + \text{FFN}(\mathbf{H}_l^{dec}))$

with Dropout($p = 0.2$) in all sub-layers

Trajectory Head

$\mathbf{T} = \text{Linear}(\text{Dropout}(\text{GELU}(\text{LN}(\text{Linear}(\mathbf{H}_L^{dec})))))$

$\mathbf{T} \in \mathbb{R}^{M \times H \times 2}$ where $M = 6, H = 30$

$\mathbf{t}_{m,h} = [x_{m,h}, y_{m,h}]^T$ (future positions)

Risk Head

$r = \sigma(\text{Linear}(\text{Dropout}(\text{ReLU}(\text{LN}(\text{Linear}(\mathbf{h}_{cls}))))))$

$r \in [0, 1]$ (collision risk probability)

$\sigma(x) = \frac{1}{1+e^{-x}}$ (sigmoid activation)

Intention Head

$\mathbf{p} = \text{softmax}(\text{Linear}(\text{Dropout}(\text{ReLU}(\text{LN}(\text{Linear}(\mathbf{h}_{cls}))))))$

$\mathbf{p} \in \mathbb{R}^4$ (intention probabilities)

$p_i \in$ [lane_keep, left_change, right_change, turn]

Physics Constraints & Kinodynamic Rules

Velocity Continuity: $||\mathbf{v}_{t+1} - \mathbf{v}_t|| \leq a_{max} \cdot \Delta t$

Steering Limits: $|\theta_{t+1} - \theta_t| \leq \omega_{max} \cdot \Delta t$

Speed Bounds: $0 \leq ||\mathbf{v}_t|| \leq v_{max}$

Acceleration Limits: $||\mathbf{a}_t|| = ||\frac{\mathbf{v}_{t+1} - \mathbf{v}_t}{\Delta t}|| \leq a_{max}$