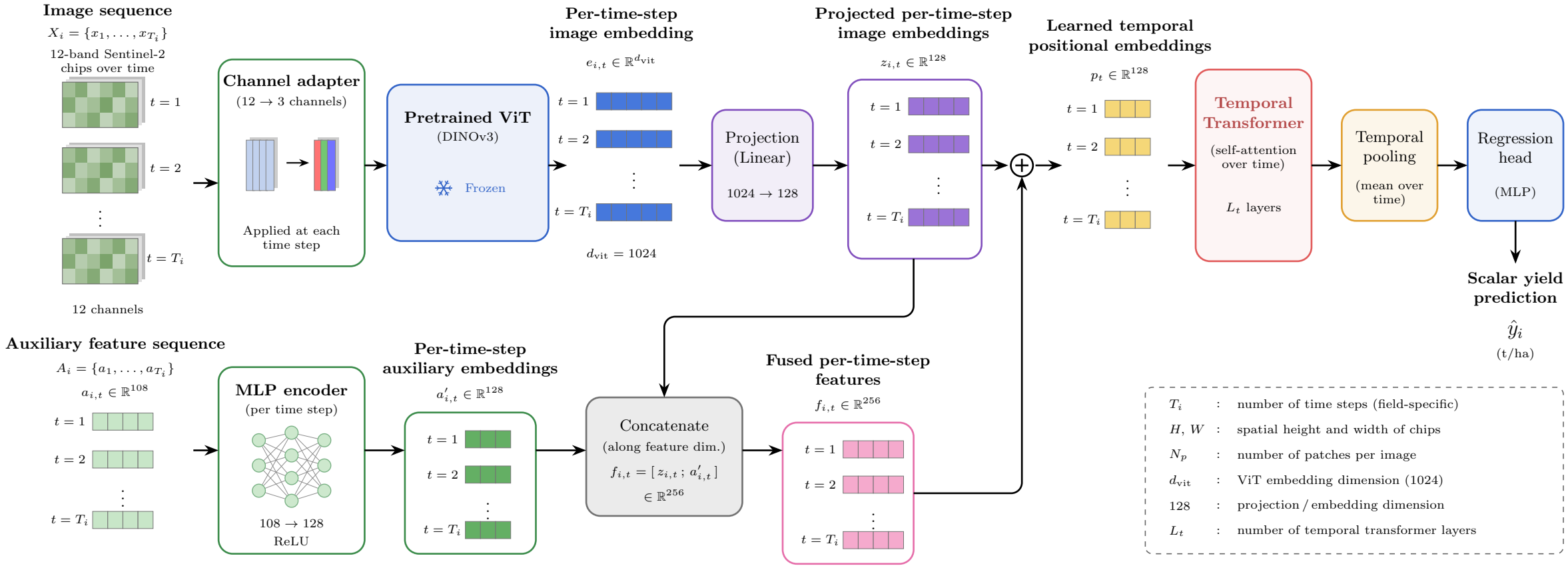
Image sequence
$X_i = \{x_1, \ldots, x_{T_i}\}$
12-band Sentinel-2 chips over time
$t = 1$
$t = 2$
$t = T_i$
12 channels
Channel adapter
(12 → 3 channels)
Applied at each time step
Pretrained ViT
(DINOv3)
Frozen
Per-time-step image embedding
$e_{i,t} \in \mathbb{R}^{d_{\text{vit}}}$
$d_{\text{vit}} = 1024$
Projection (Linear)
1024 → 128
Projected per-time-step image embeddings
$z_{i,t} \in \mathbb{R}^{128}$
Learned temporal positional embeddings
$p_t \in \mathbb{R}^{128}$
Temporal Transformer
(self-attention over time)
L_t layers
Temporal pooling
(mean over time)
Regression head
(MLP)
Scalar yield prediction
$\hat{y}_i$
(t/ha)
Auxiliary feature sequence
$A_i = \{a_1, \ldots, a_{T_i}\}$
$a_{i,t} \in \mathbb{R}^{108}$
MLP encoder
(per time step)
108 → 128
ReLU
Per-time-step auxiliary embeddings
$a'_{i,t} \in \mathbb{R}^{128}$
Concatenate
(along feature dim.)
$f_{i,t} = [\, z_{i,t} \,;\, a'_{i,t} \,]$
$\in \mathbb{R}^{256}$
Fused per-time-step features
$f_{i,t} \in \mathbb{R}^{256}$
T_i : number of time steps (field-specific)
H, W : spatial height and width of chips
N_p : number of patches per image
d_{vit} : ViT embedding dimension (1024)
128 : projection / embedding dimension
L_t : number of temporal transformer layers