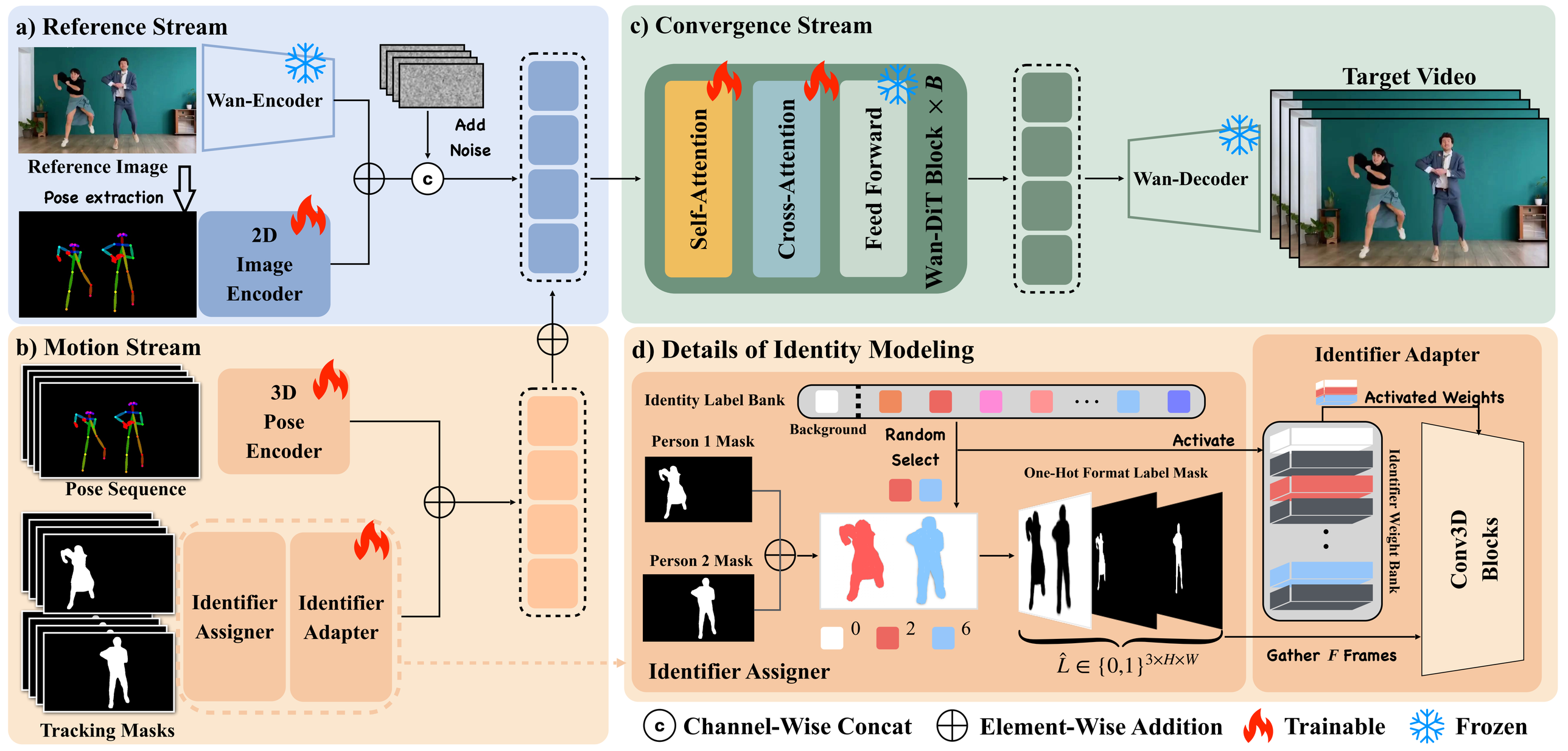
a) Reference Stream
Reference Image
Pose extraction
Wan-Encoder
2D Image Encoder
Add Noise
c
b) Motion Stream
Pose Sequence
3D Pose Encoder
Tracking Masks
Identifier Assigner
Identifier Adapter
c) Convergence Stream
Self-Attention
Cross-Attention
Feed Forward
Wan-DiT Block × B
Wan-Decoder
Target Video
d) Details of Identity Modeling
Identity Label Bank
Background
Random Select
Activate
Person 1 Mask
Person 2 Mask
0
2
6
One-Hot Format Label Mask
$\hat{L} \in \{0,1\}^{3\times H\times W}$
Identifier Assigner
Identifier Adapter
Activated Weights
Identifier Weight Bank
Gather F Frames
Conv3D Blocks
c Channel-Wise Concat
Element-Wise Addition
Trainable
Frozen