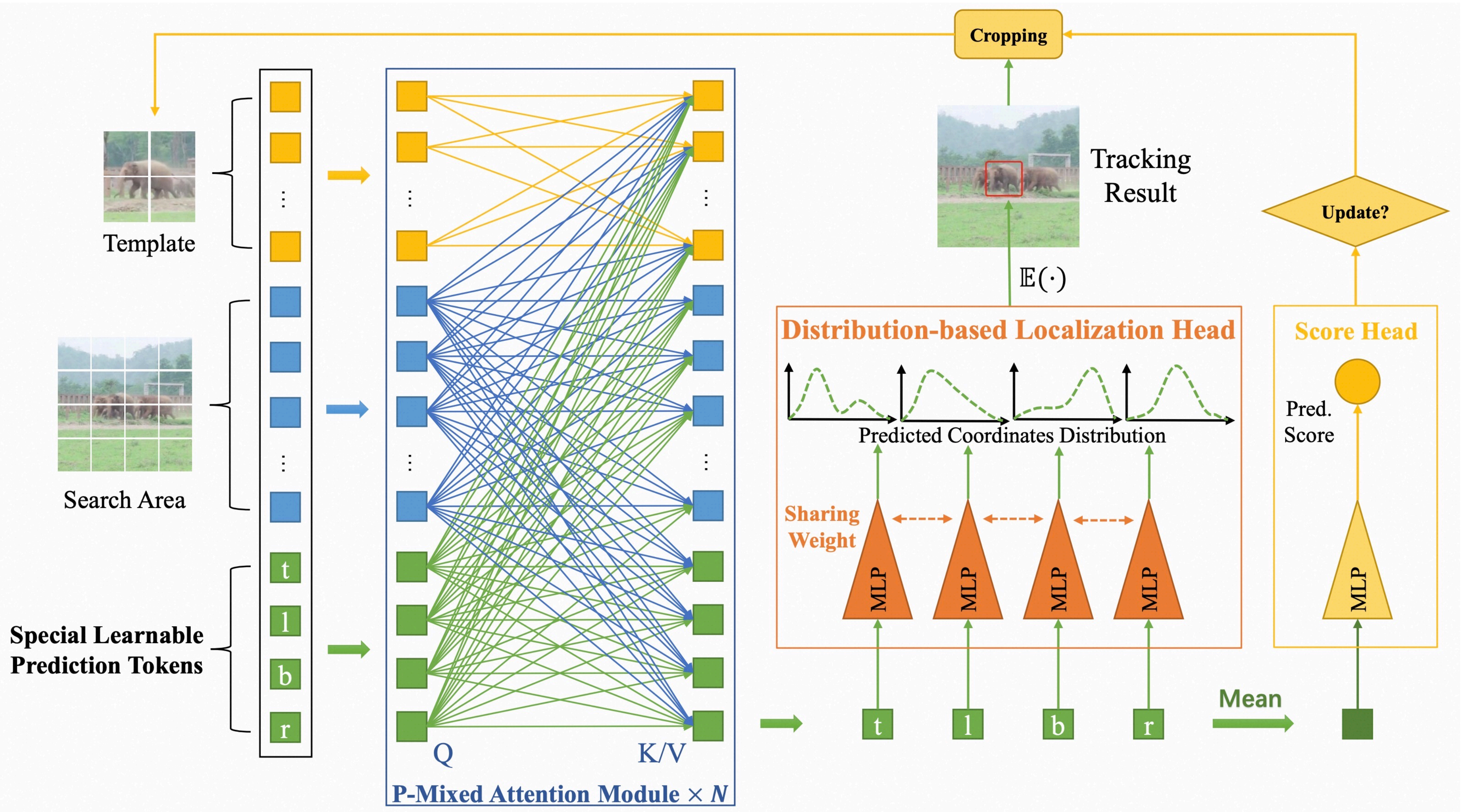
MixFormerV2: Efficient Fully Transformer Tracking.

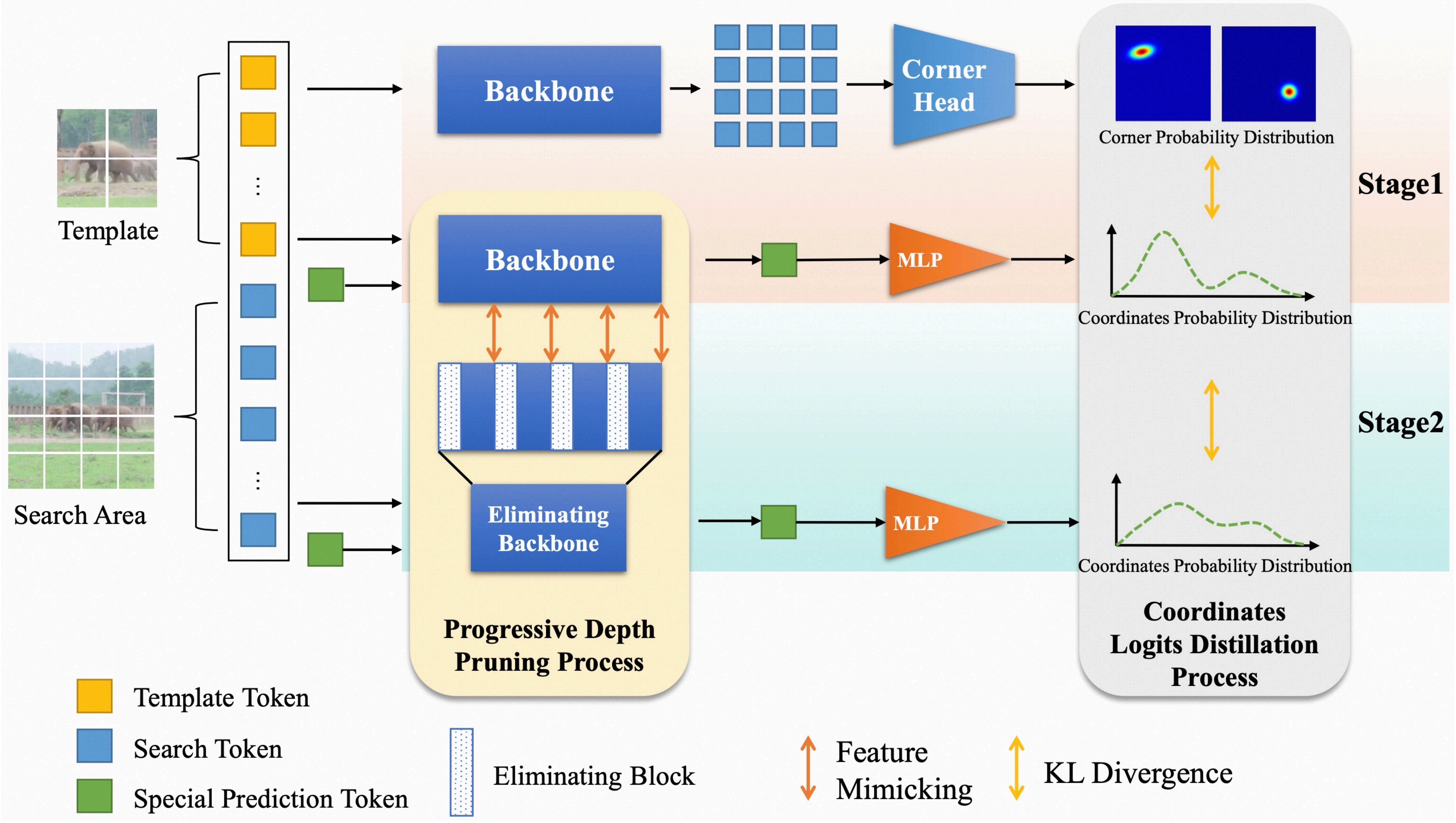
Yutao Cui*, Tianhui Song*, Gangshan Wu, Limin Wang
State Key Laboratory for Novel Software Technology, Nanjing University, China

Efficient Fully Transformer Tracking Framework.


Cropping
Tracking Result
Update?
Template
Search Area
Special Learnable Prediction Tokens
t
l
b
r
Q
K/V
P-Mixed Attention Module × N
𝔼(·)
Distribution-based Localization Head
Predicted Coordinates Distribution
Sharing Weight
MLP
Score Head
Pred. Score
Mean

- **Four Special Prediction Tokens** is proposed for both target localization and confidence estimation in unified architecture.
- **Distribution-based Prediction** is effective for improving accuracy and bridging distillation process.

Progressive Distillation Training Pipeline.



- **Dense-to-Sparse** distillation smoothly transfers knowledge between teacher and student models with different localization head.
- **Deep-to-Shallow** distillation efficiently prunes the backbone layers for model reduction.

Experimental results

- MixFormerV2-B For GPU High Speed

Method	LaSOT AUC(%)	GPU Speed (fps)
MixViT-B (Teacher)	69.6	75
MixViT (8 blocks, baseline)	66.5	90
w/ our architecture	65.0	165
w/ our distillation	69.7	90
w/ both, MixFormerV2-B	**70.6**	**165**

- MixFormerV2-S For CPU Real-time Speed

Method	LaSOT AUC(%)	GPU Speed	CPU Speed
MixFormerV2-S	**60.6**	**325**	**30**

- Visualization
  - Video Object Segmentation with SAM

  - Video Object Removal with E2FGVI

  - Visualization of attention maps of four prediction tokens