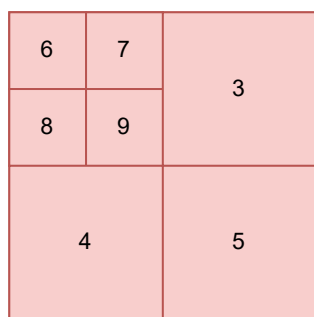




Patching into size $S \times S$



W

Linear projection

S

S

block

block

Prediction = 1

1

Block aggregation

T

• • • (#block/16, seqlen, d)

2

3

4

5

Block aggregation

Block function T (MSA_{NesT}, FFN, LN)

T

• • • (#block/4, seqlen, d)

6

7

8

9

• • • (#block, seqlen, d)