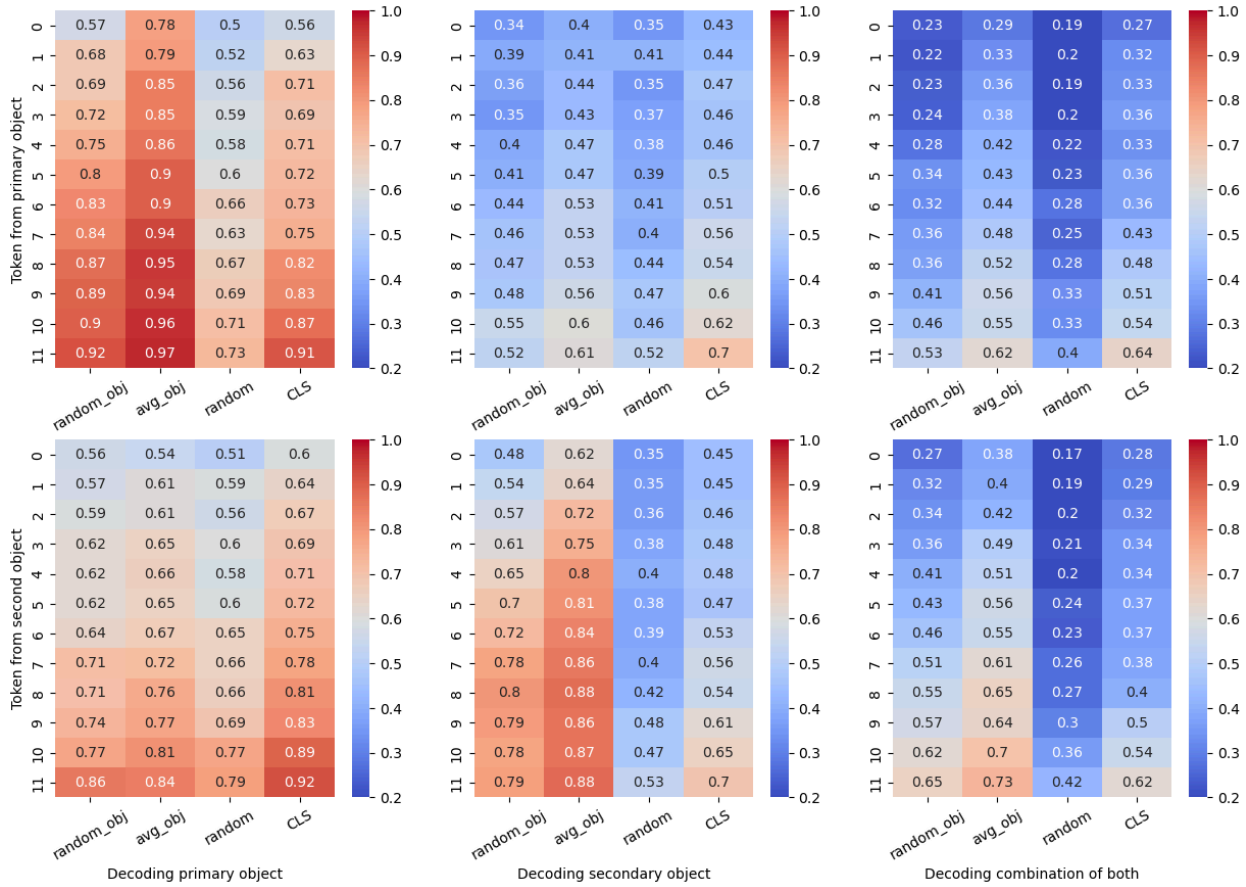
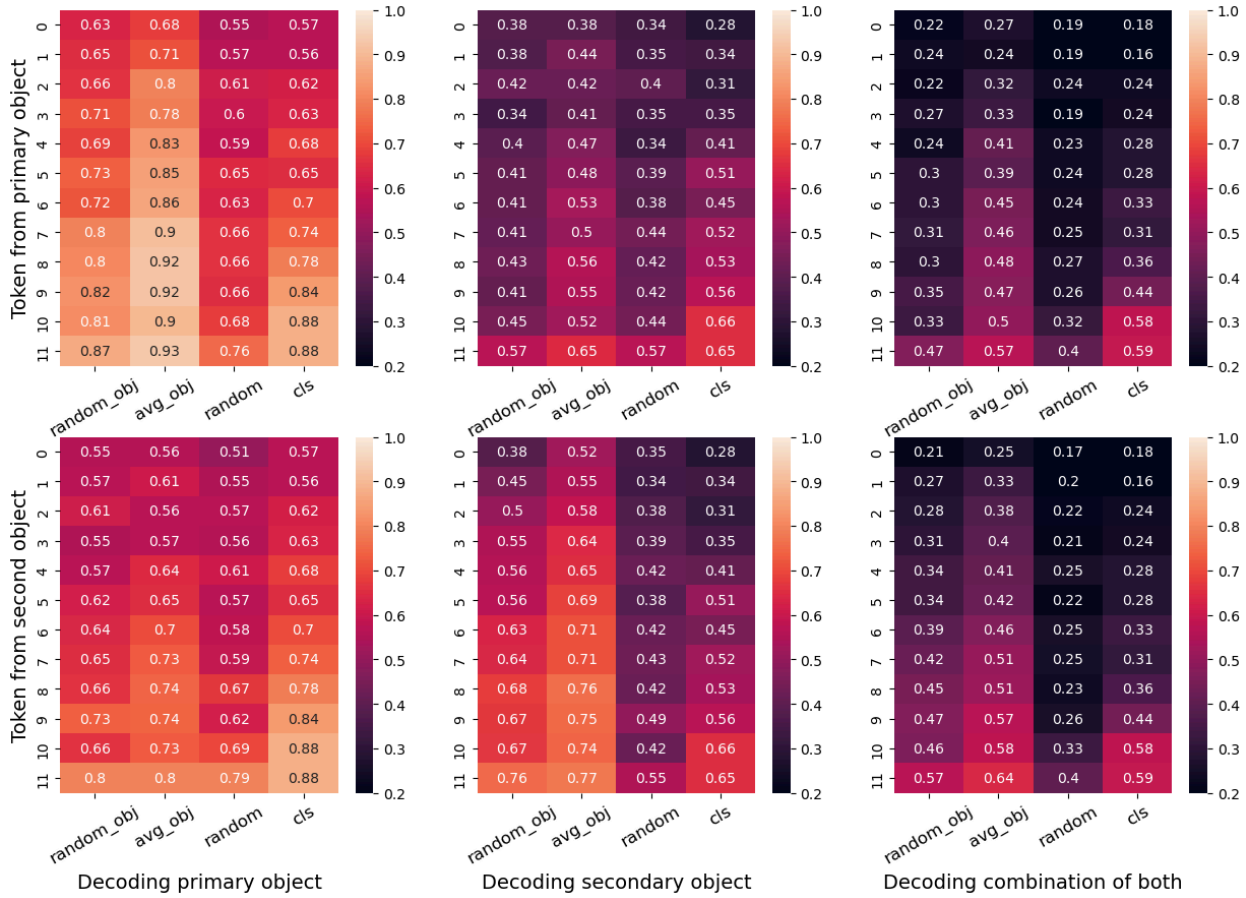
Supplementary material

Complex probe ablation (3 layers of 100,50,10 neurons in that order) for BLIP (showing same decoding pattern just higher numbers). Even though we re-iterate that a higher number for complex probes cannot be attributed to better representations as the probe itself can learn many relationships.

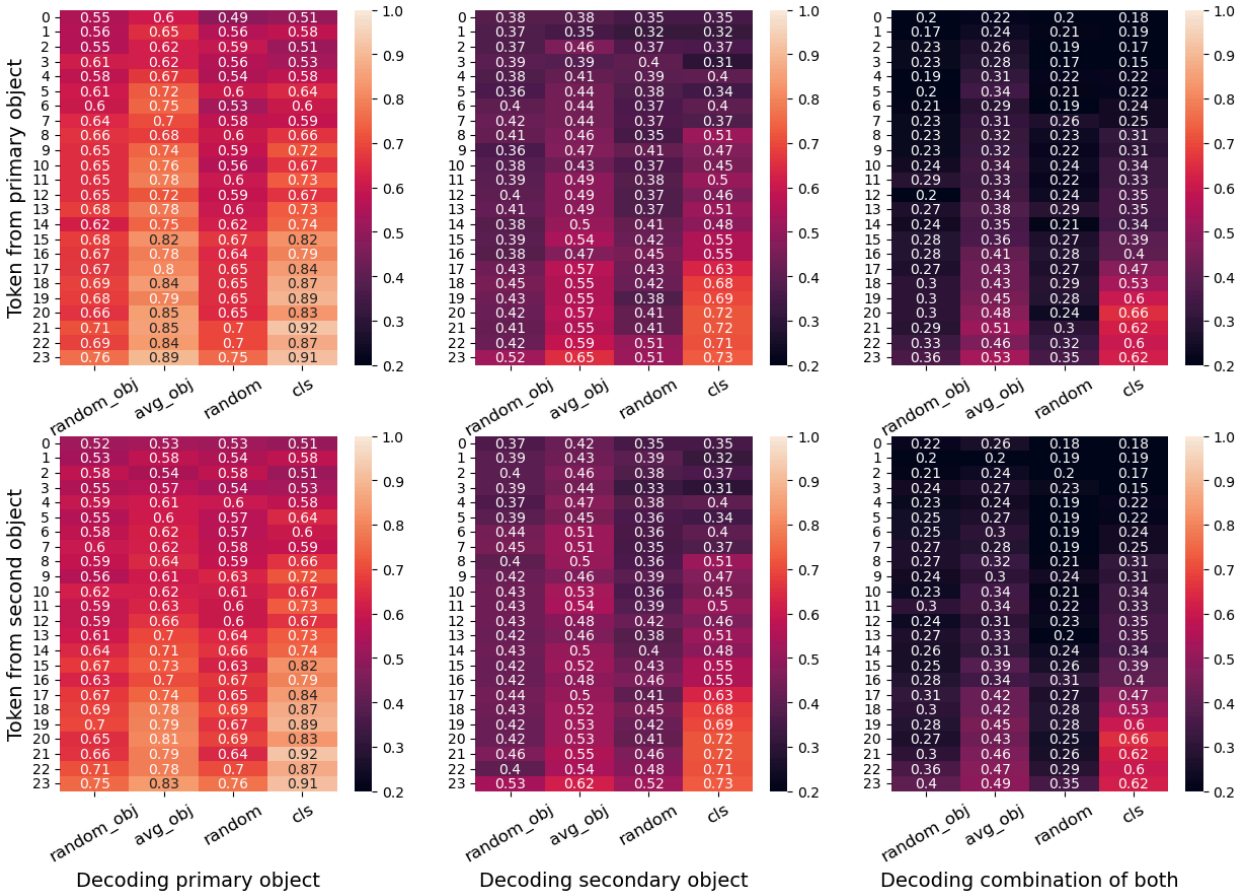


The detailed plot for the paired object decoding task for each model is given below:

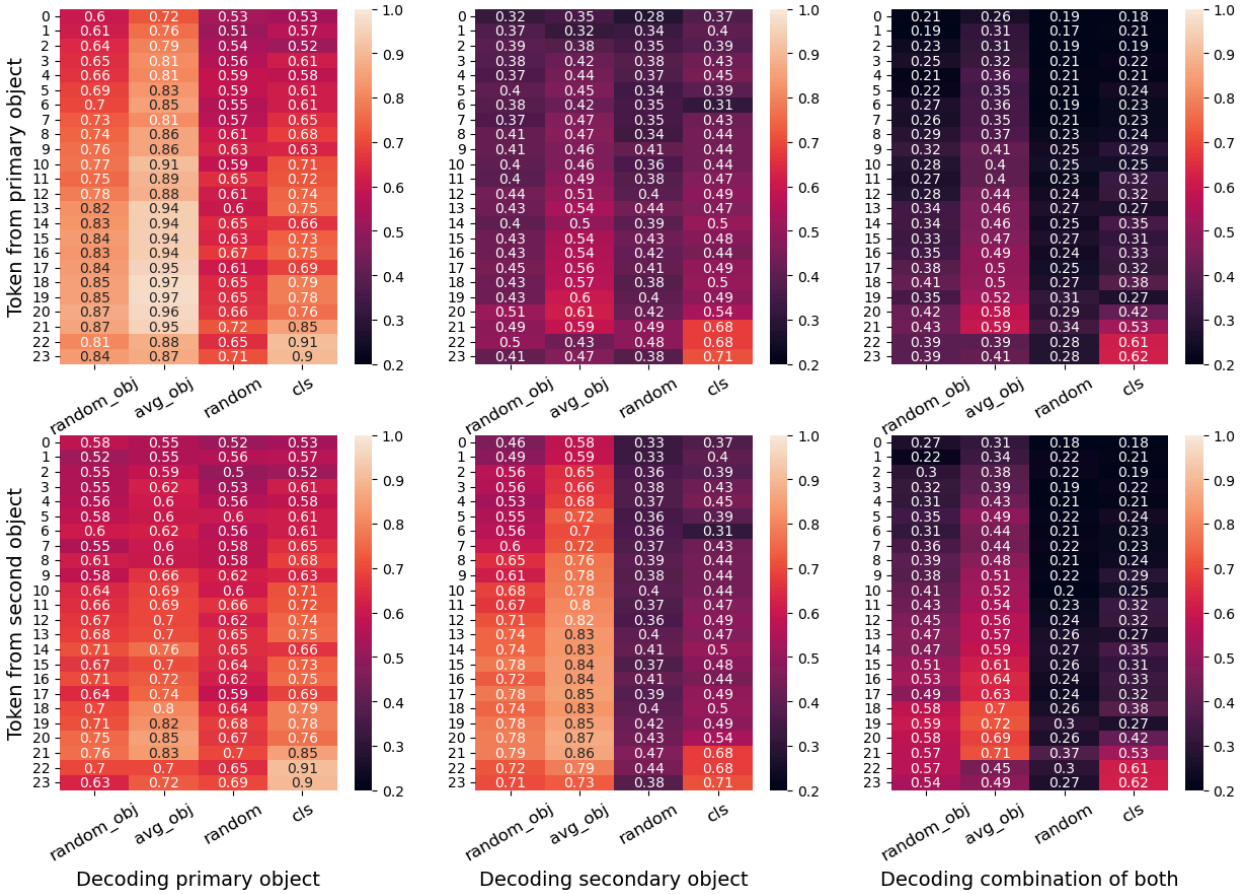
CLIP



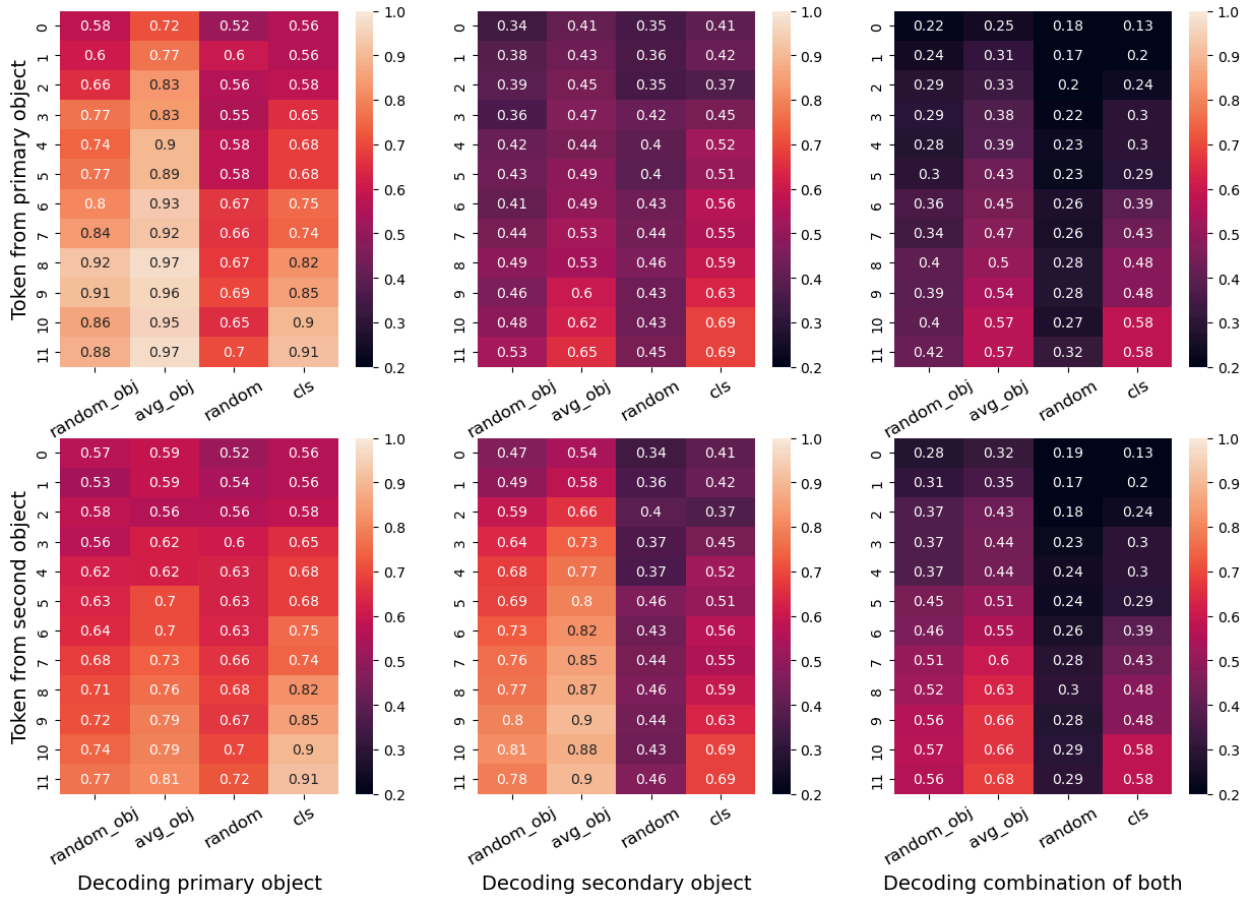
CLIP-Large


Token from primary object
Token from second object
random_obj
avg_obj
random
cls
Decoding primary object
Decoding secondary object
Decoding combination of both

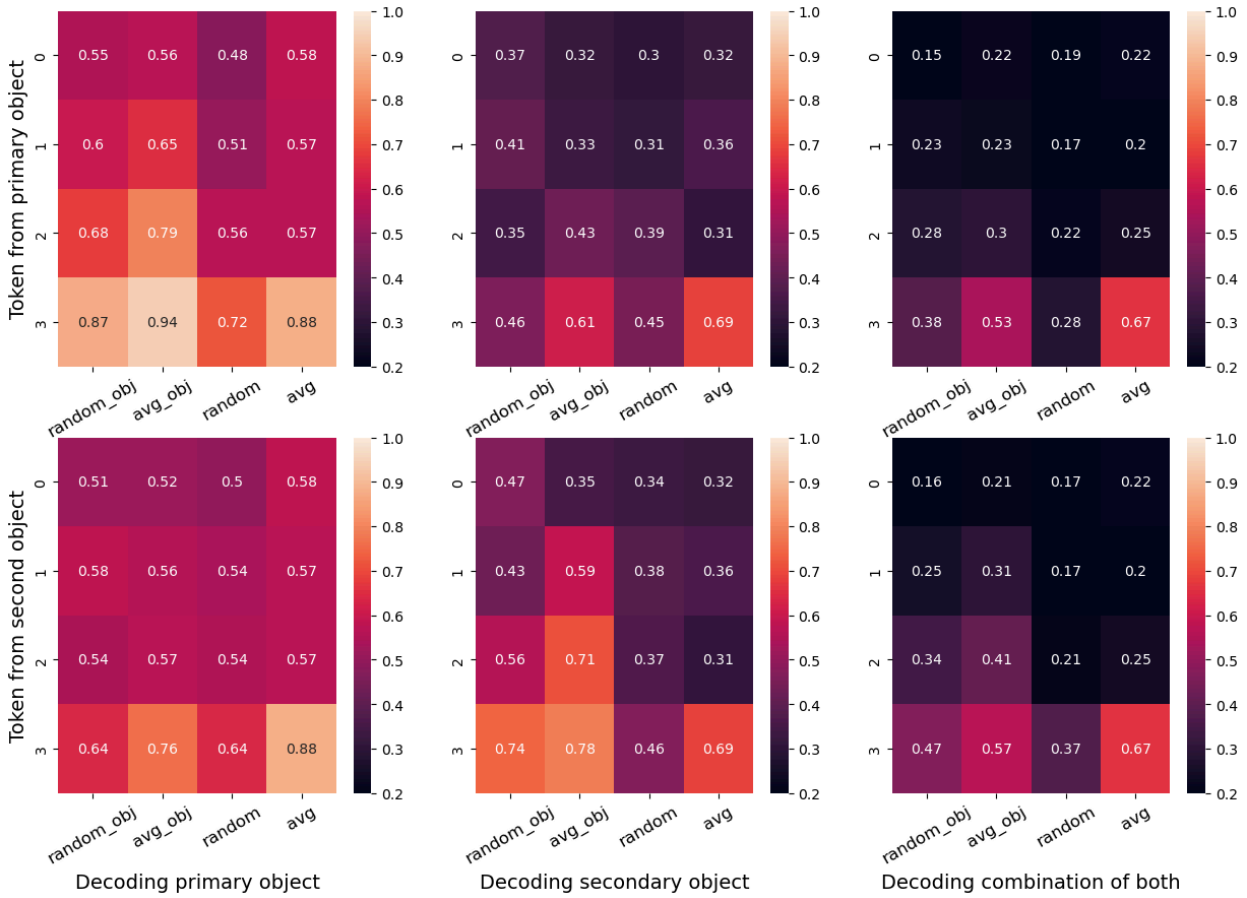
BLIP-Large


Token from primary object
Token from second object
random_obj
avg_obj
random
cls
Decoding primary object
Decoding secondary object
Decoding combination of both

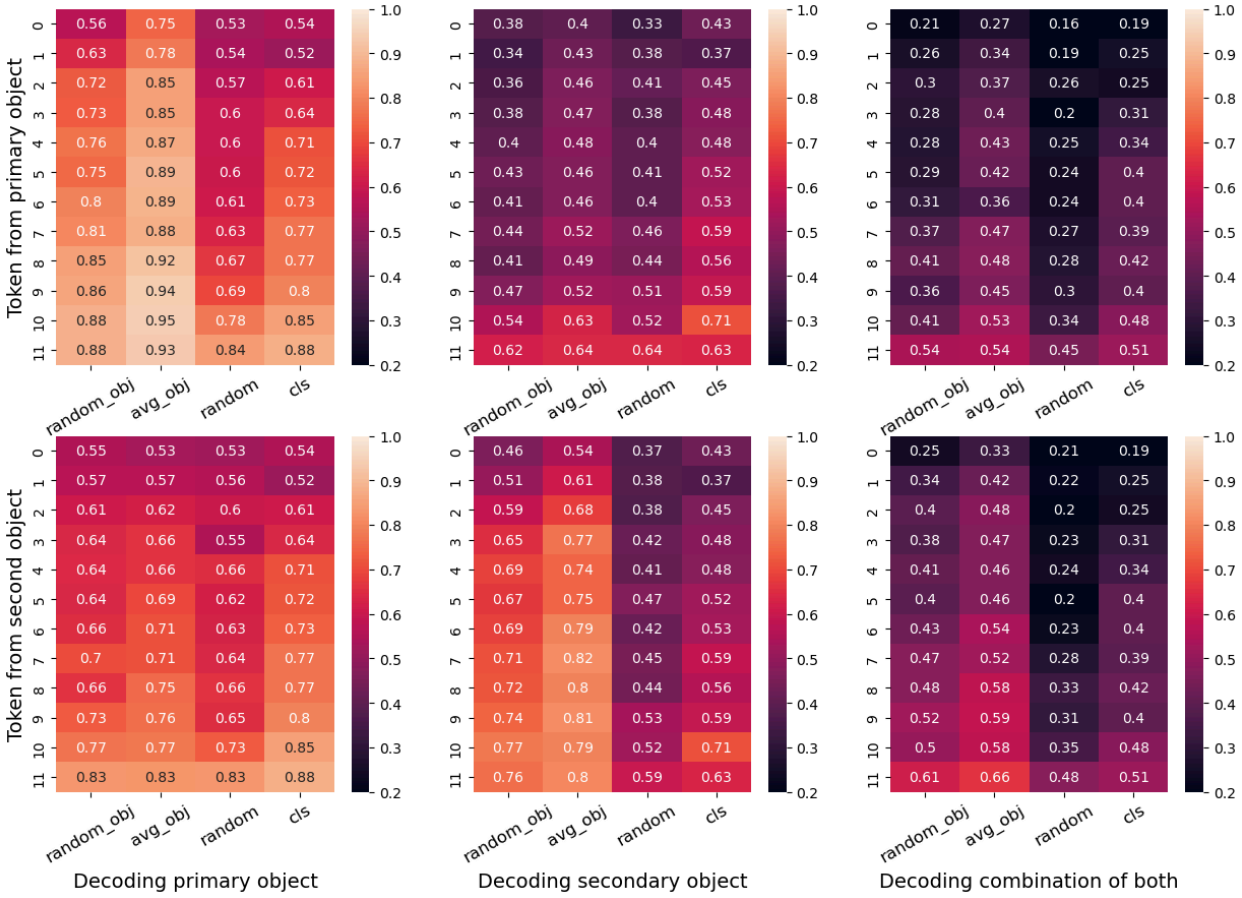
FLAVA



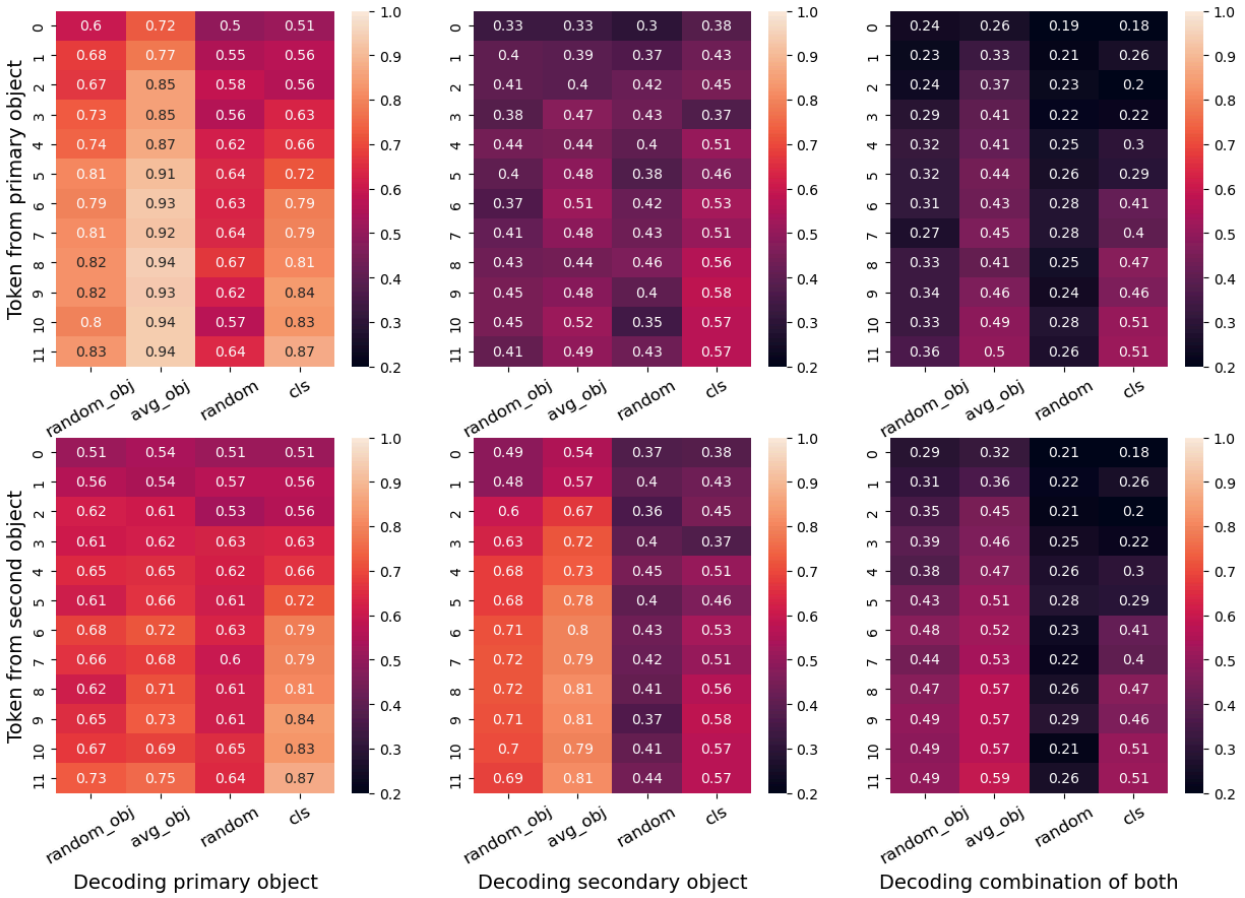
CLIP Renet X4



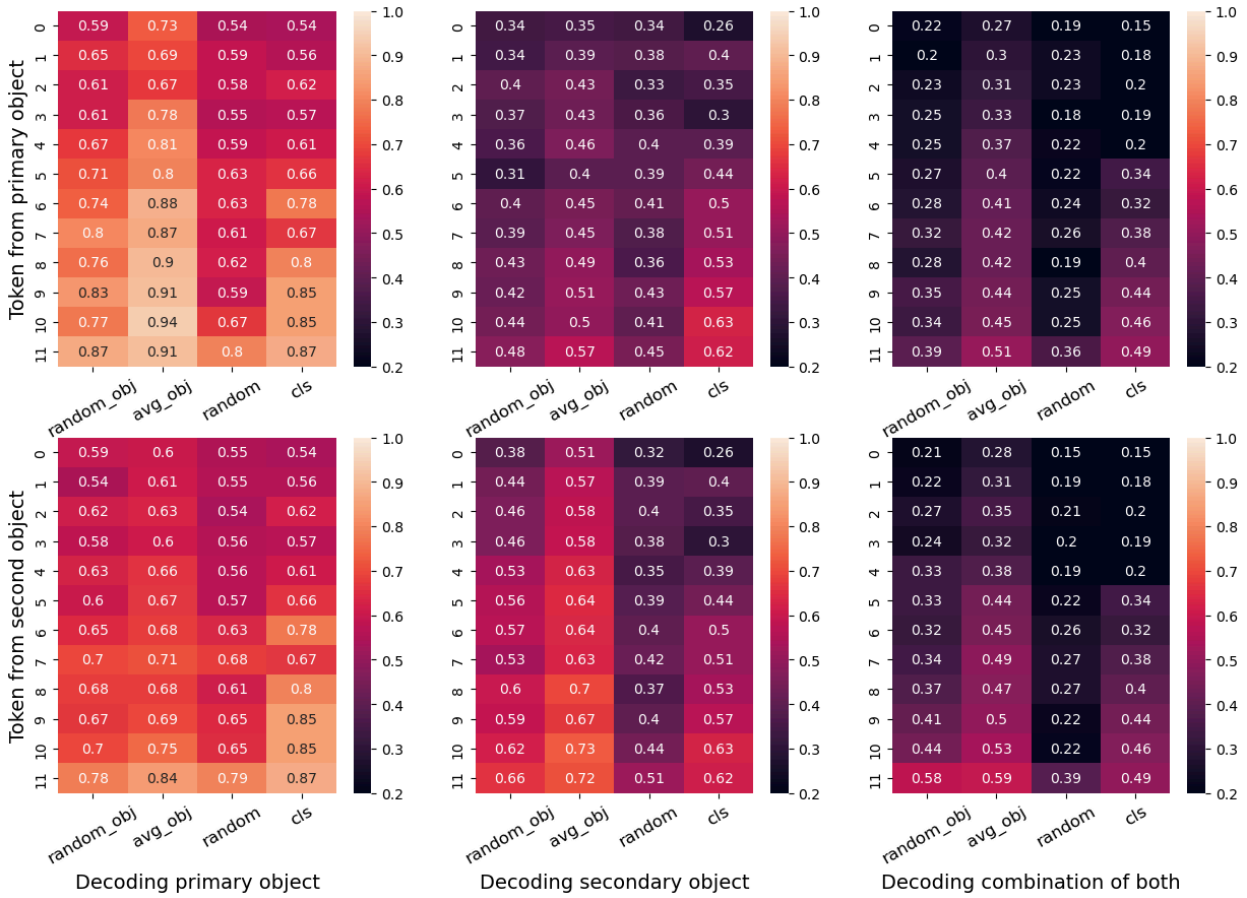
ViT (Imagenet 21k)


Token from primary object
Token from second object
random_obj
avg_obj
random
cls
Decoding primary object
Decoding secondary object
Decoding combination of both

DINO


Token from primary object
Token from second object
random_obj
avg_obj
random
cls
Decoding primary object
Decoding secondary object
Decoding combination of both

DINO2


Token from primary object
Token from second object
random_obj
avg_obj
random
cls
Decoding primary object
Decoding secondary object
Decoding combination of both