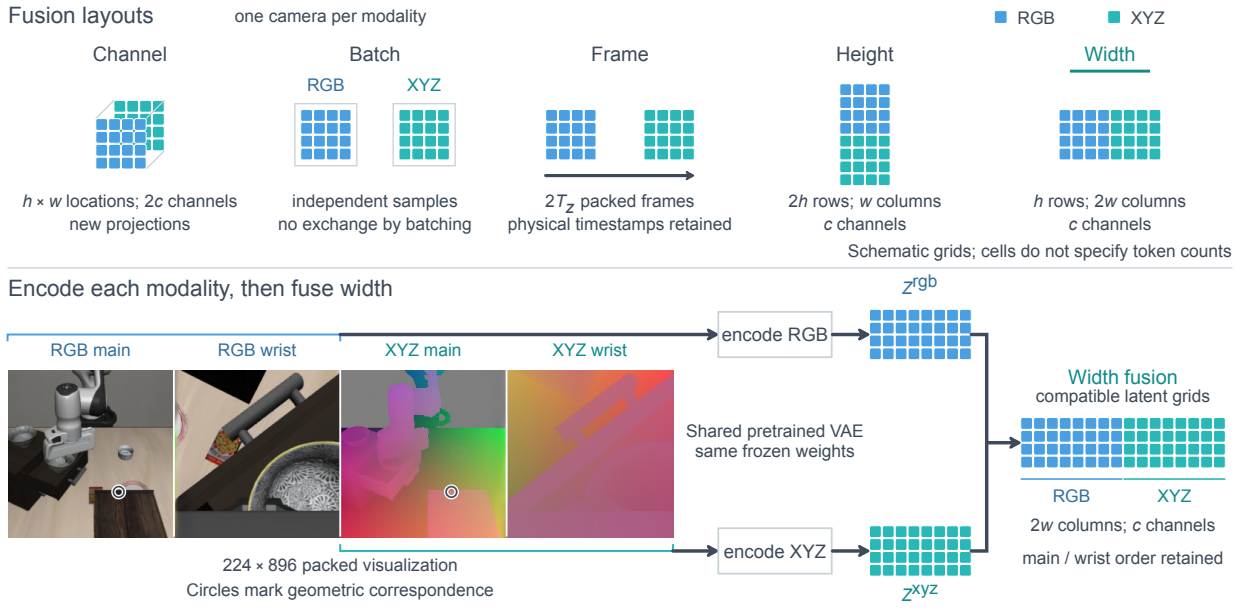

Fusion layouts
one camera per modality
RGB
XYZ
Channel
Batch
RGB
XYZ
Frame
Height
Width
h × w locations; 2c channels
new projections
independent samples
no exchange by batching
2T_z packed frames
physical timestamps retained
2h rows; w columns
c channels
h rows; 2w columns
c channels
Schematic grids; cells do not specify token counts
Encode each modality, then fuse width
z^rgb
RGB main
RGB wrist
XYZ main
XYZ wrist
encode RGB
Shared pretrained VAE
same frozen weights
Width fusion
compatible latent grids
RGB
XYZ
2w columns; c channels
main / wrist order retained
encode XYZ
z^xyz
224 × 896 packed visualization
Circles mark geometric correspondence