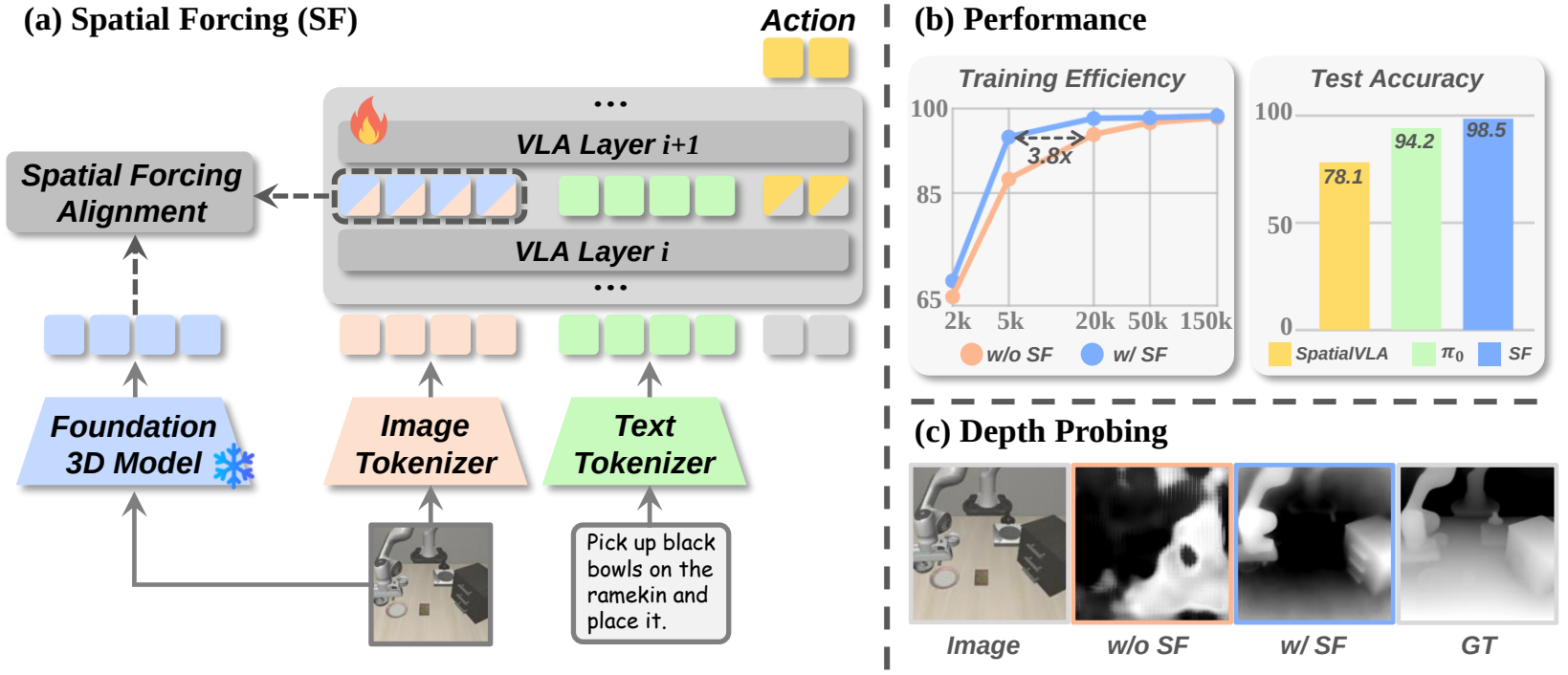
(a) Spatial Forcing (SF)
Action
VLA Layer i+1
VLA Layer i
Spatial Forcing Alignment
Foundation 3D Model
Image Tokenizer
Text Tokenizer
Pick up black bowls on the ramekin and place it.
(b) Performance
Training Efficiency
100
85
65
3.8x
2k
5k
20k
50k
150k
w/o SF
w/ SF
Test Accuracy
100
50
0
78.1
94.2
98.5
SpatialVLA
π_0
SF
(c) Depth Probing
Image
w/o SF
w/ SF
GT