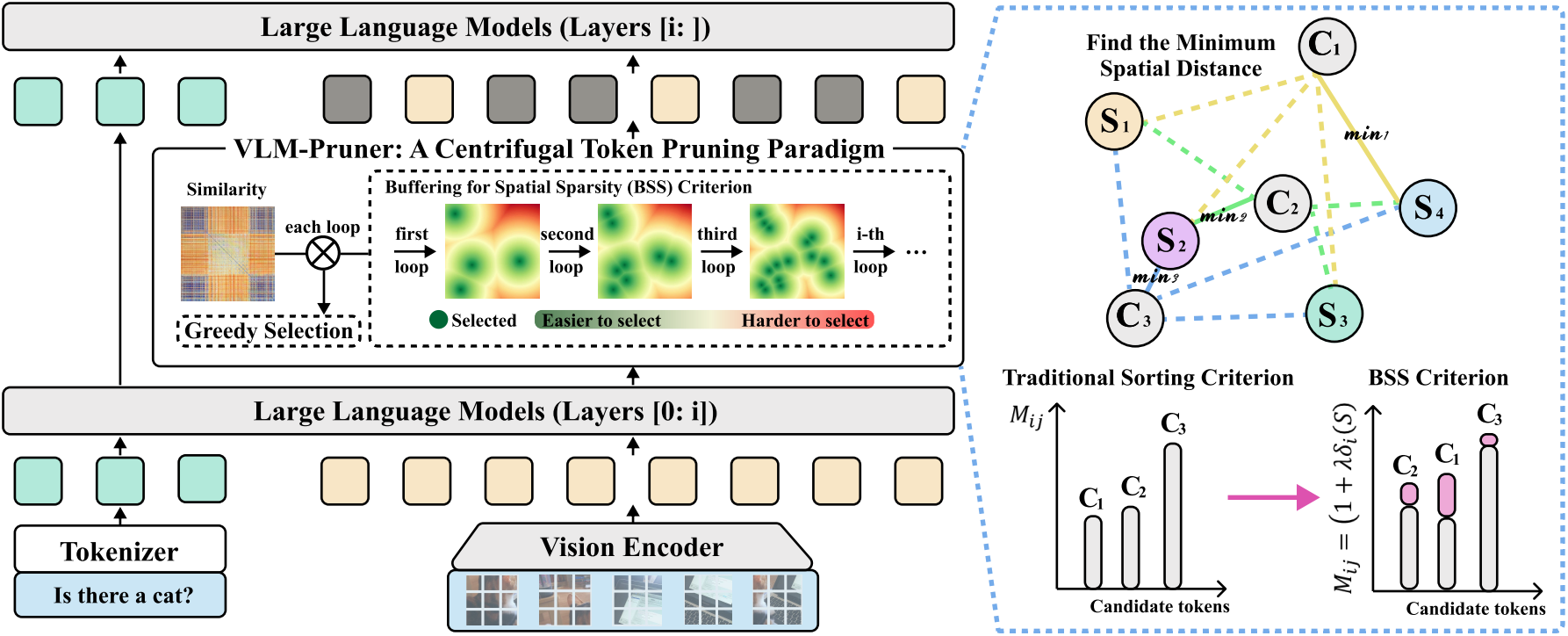
Large Language Models (Layers [i:])
VLM-Pruner: A Centrifugal Token Pruning Paradigm
Similarity
each loop
Greedy Selection
Buffering for Spatial Sparsity (BSS) Criterion
first loop
second loop
third loop
i-th loop
...
Selected
Easier to select
Harder to select
Large Language Models (Layers [0: i])
Tokenizer
Is there a cat?
Vision Encoder
Find the Minimum Spatial Distance
C1
C2
C3
S1
S2
S3
S4
min1
min2
min3
Traditional Sorting Criterion
BSS Criterion
Mij
Candidate tokens
Candidate tokens