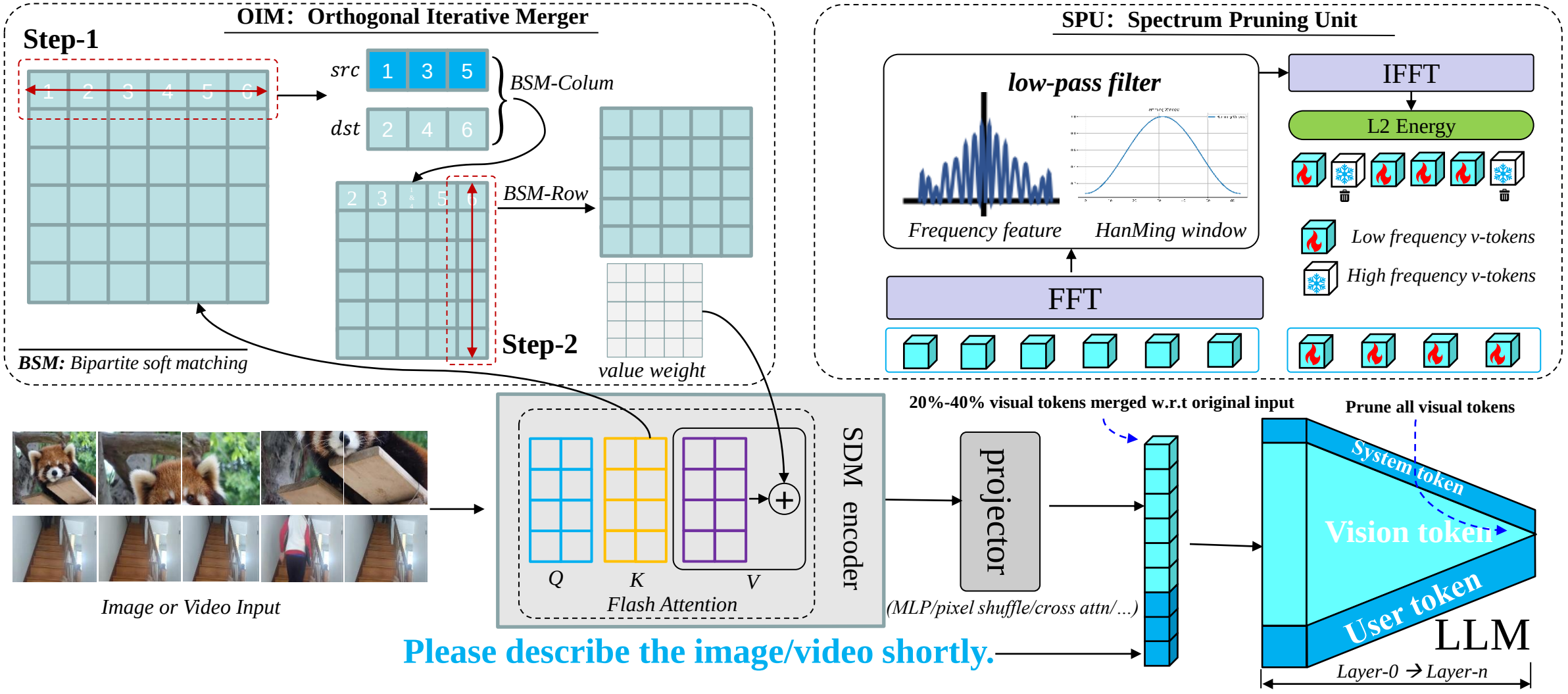
OIM: Orthogonal Iterative Merger
Step-1
src
dst
BSM-Colum
BSM-Row
Step-2
value weight
BSM: Bipartite soft matching
SPU: Spectrum Pruning Unit
low-pass filter
Frequency feature
HanMing window
IFFT
L2 Energy
Low frequency v-tokens
High frequency v-tokens
FFT
Image or Video Input
Q
K
V
Flash Attention
SDM encoder
20%-40% visual tokens merged w.r.t original input
projector
(MLP/pixel shuffle/cross attn/...)
Prune all visual tokens
System token
Vision token
User token
LLM
Layer-0 → Layer-n
Please describe the image/video shortly.