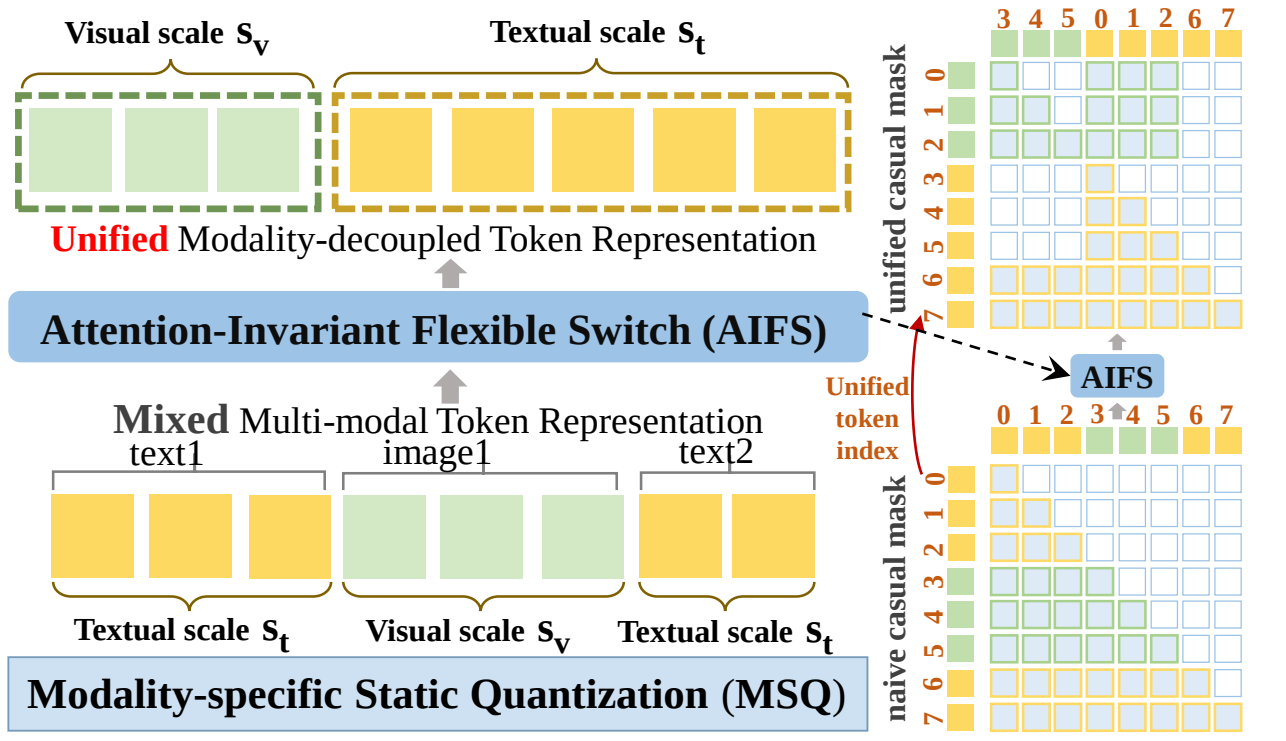
Visual scale S_v
Textual scale S_t
Unified Modality-decoupled Token Representation
Attention-Invariant Flexible Switch (AIFS)
Mixed Multi-modal Token Representation
text1
image1
text2
Textual scale S_t
Visual scale S_v
Textual scale S_t
Modality-specific Static Quantization (MSQ)
unified casual mask
naive casual mask
AIFS
Unified token index
3 4 5 0 1 2 6 7
0 1 2 3 4 5 6 7