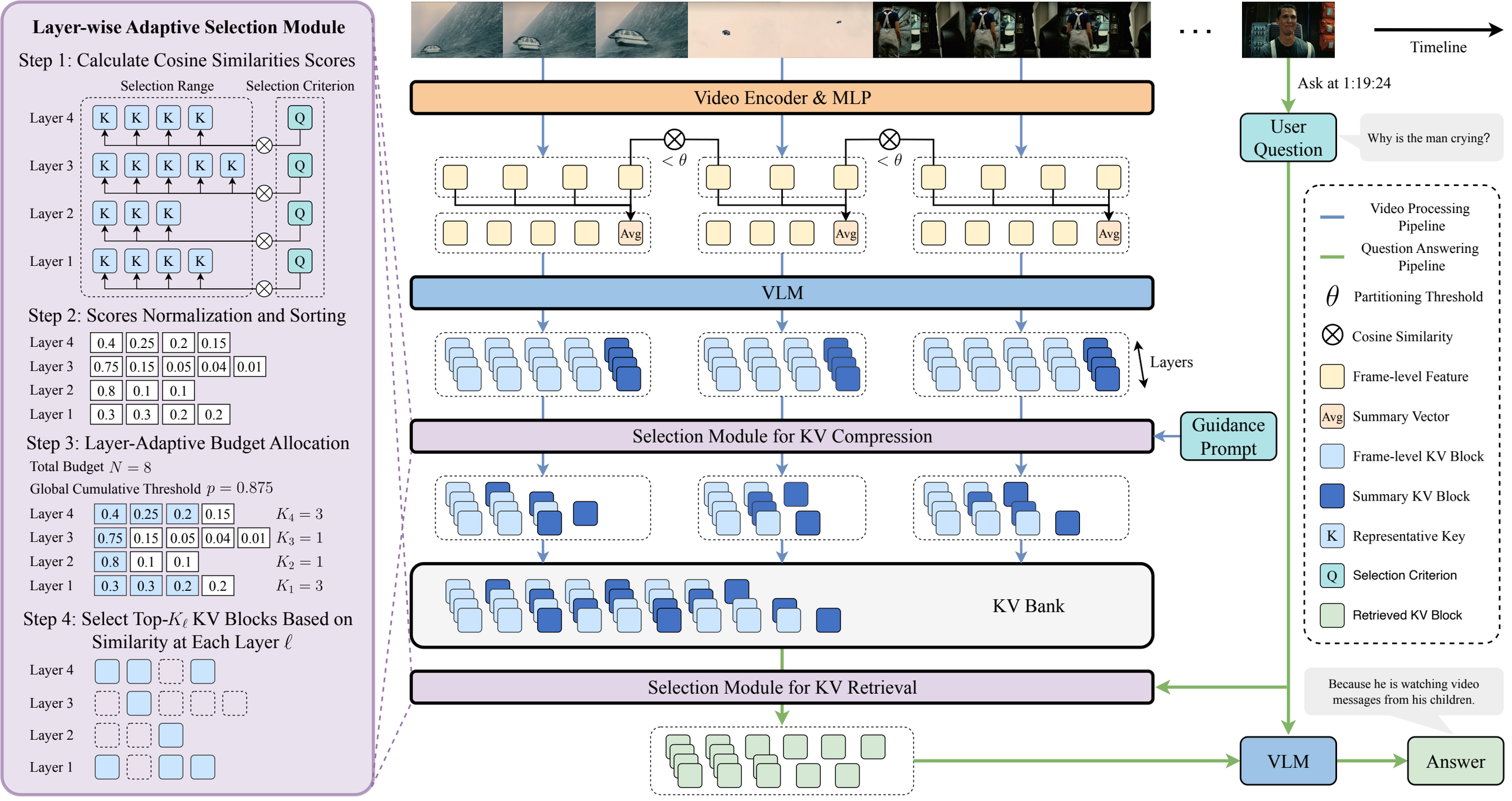
Layer-wise Adaptive Selection Module
Step 1: Calculate Cosine Similarities Scores
Selection Range
Selection Criterion
Layer 4
Layer 3
Layer 2
Layer 1
K
Q
Step 2: Scores Normalization and Sorting
Layer 4 0.4 0.25 0.2 0.15
Layer 3 0.75 0.15 0.05 0.04 0.01
Layer 2 0.8 0.1 0.1
Layer 1 0.3 0.3 0.2 0.2
Step 3: Layer-Adaptive Budget Allocation
Total Budget $N = 8$
Global Cumulative Threshold $p = 0.875$
Layer 4 0.4 0.25 0.2 0.15 $K_4 = 3$
Layer 3 0.75 0.15 0.05 0.04 0.01 $K_3 = 1$
Layer 2 0.8 0.1 0.1 $K_2 = 1$
Layer 1 0.3 0.3 0.2 0.2 $K_1 = 3$
Step 4: Select Top-K_ℓ KV Blocks Based on Similarity at Each Layer ℓ
Layer 4
Layer 3
Layer 2
Layer 1
Video Encoder & MLP
$< \theta$
$< \theta$
Avg
VLM
Layers
Selection Module for KV Compression
Guidance Prompt
KV Bank
Selection Module for KV Retrieval
Timeline
Ask at 1:19:24
User Question
Why is the man crying?
Video Processing Pipeline
Question Answering Pipeline
Partitioning Threshold
Cosine Similarity
Frame-level Feature
Summary Vector
Frame-level KV Block
Summary KV Block
Representative Key
Selection Criterion
Retrieved KV Block
Because he is watching video messages from his children.
VLM
Answer