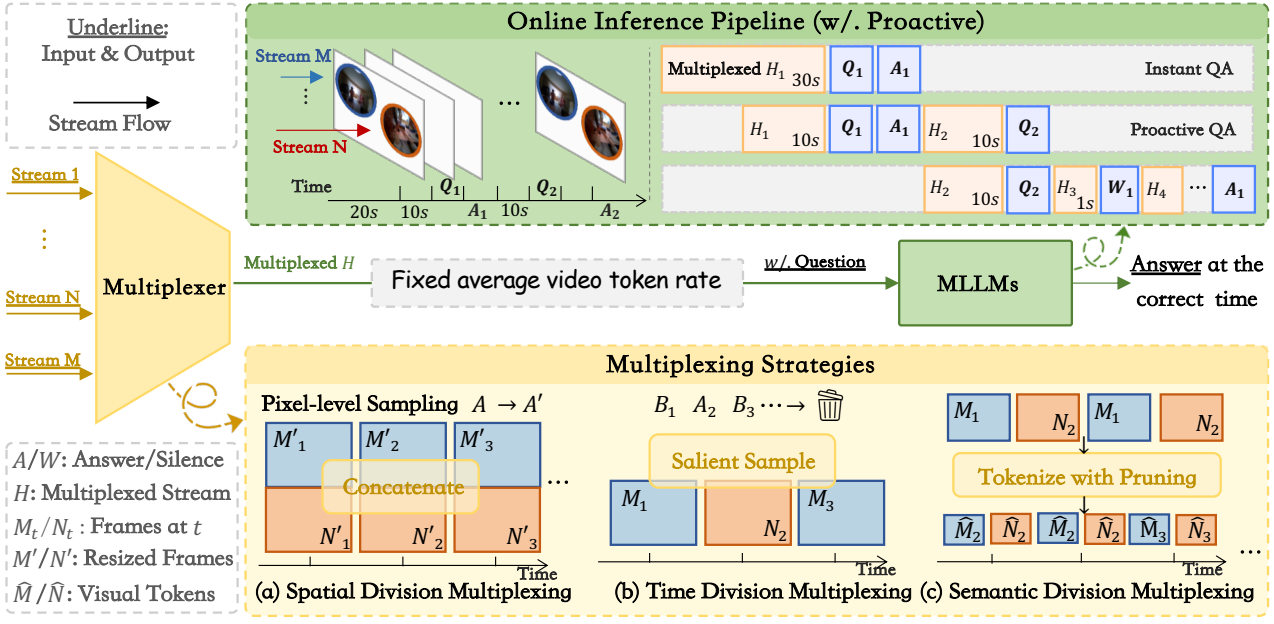
Underline:
Input & Output
Stream Flow
Online Inference Pipeline (w/. Proactive)
Stream M
Stream N
Time
20s
10s
Q_1
A_1
10s
Q_2
A_2
Multiplexed H_1 30s
Q_1
A_1
Instant QA
H_1 10s
Q_1
A_1
H_2 10s
Q_2
Proactive QA
H_2 10s
Q_2
H_3 1s
W_1
H_4
A_1
Stream 1
Stream N
Stream M
Multiplexer
Multiplexed H
Fixed average video token rate
w/. Question
MLLMs
Answer at the correct time
Multiplexing Strategies
Pixel-level Sampling $A \rightarrow A'$
M'_1
M'_2
M'_3
Concatenate
N'_1
N'_2
N'_3
Time
(a) Spatial Division Multiplexing
B_1 A_2 $B_3 \cdots \rightarrow$
Salient Sample
M_1
N_2
M_3
Time
(b) Time Division Multiplexing
M_1
N_2
M_1
N_2
Tokenize with Pruning
$\widehat{M}_2$
$\widehat{N}_2$
$\widehat{M}_2$
$\widehat{N}_2$
$\widehat{M}_3$
$\widehat{N}_3$
Time
(c) Semantic Division Multiplexing
A/W: Answer/Silence
H: Multiplexed Stream
M_t/N_t : Frames at t
M'/N': Resized Frames
$\widehat{M}/\widehat{N}$: Visual Tokens