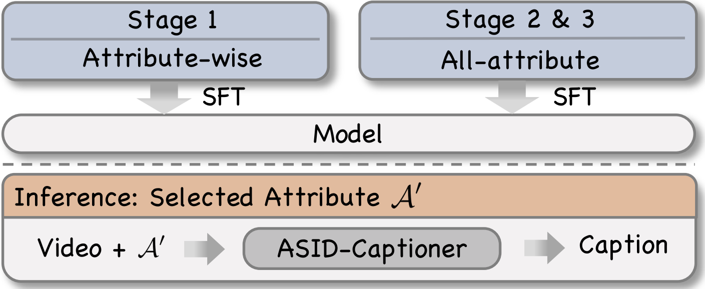
Stage 1
Attribute-wise
Stage 2 & 3
All-attribute
SFT
SFT
Model
Inference: Selected Attribute $\mathcal{A}'$
Video + $\mathcal{A}'$
ASID-Captioner
Caption