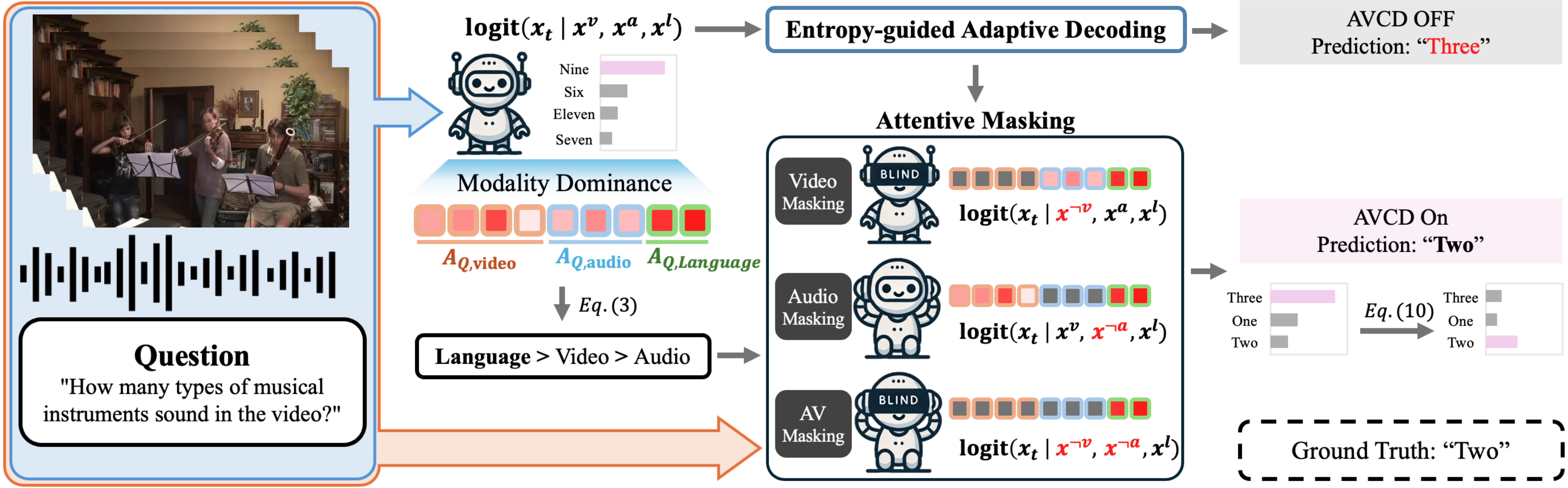
Question
"How many types of musical instruments sound in the video?"
$\mathrm{logit}(x_t \mid x^v, x^a, x^l)$
Nine
Six
Eleven
Seven
Modality Dominance
$A_{Q,video}$
$A_{Q,audio}$
$A_{Q,Language}$
$Eq.(3)$
Language > Video > Audio
Entropy-guided Adaptive Decoding
Attentive Masking
Video Masking
BLIND
$\mathrm{logit}(x_t \mid x^{\neg v}, x^a, x^l)$
Audio Masking
$\mathrm{logit}(x_t \mid x^v, x^{\neg a}, x^l)$
AV Masking
BLIND
$\mathrm{logit}(x_t \mid x^{\neg v}, x^{\neg a}, x^l)$
AVCD OFF
Prediction: "Three"
AVCD On
Prediction: "Two"
Three
One
Two
$Eq.(10)$
Three
One
Two
Ground Truth: "Two"