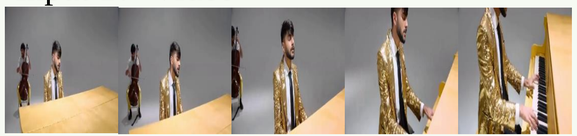
AVVS

Input 1: Video



Input 2: Audio

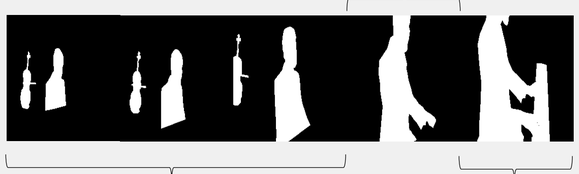
Frame 1~3: Violin / Singing

Frame 4: Singing

Frame 5: Piano/Singing

Output:

Frame 4: Mask of person



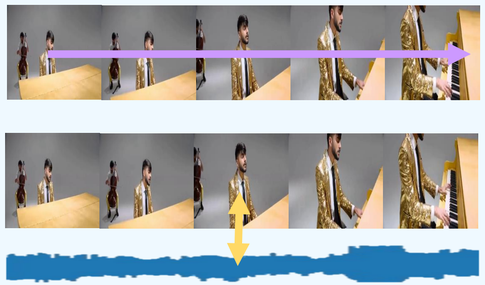
Frame 1~3: Mask of violin and person

Frame 5:Mask of piano and person

(a)

Previous Method

(i) Seperate-Dependence Fusion



video's temporal dependence

+

Audio-visual interaction

(ii) Object-Limited Queryless Decoding

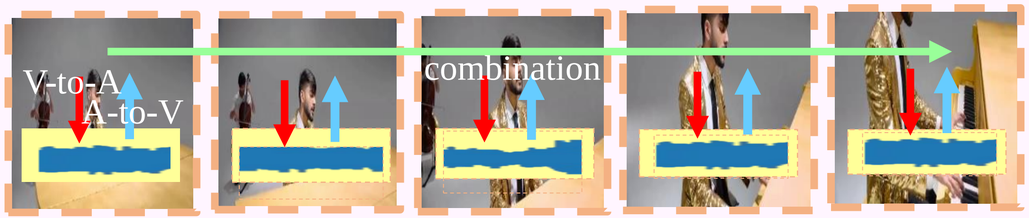
(Sound of Violin and Singing)

Piano segmented by mistake

(b)

CATR

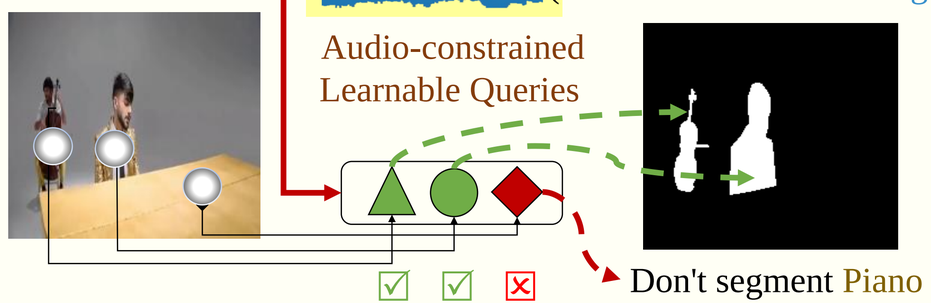
(i) Combinatorial-Dependence Fusion



The audio-video combinatorial spatial-temporal dependence

(ii) Object-Aware Audio-Queried Decoding

(Sound of Violin and Singing)



(c)