

Video visualization of segmentation results.

- We provided 5 examples from testing set, and 3 examples from in-the-wild videos. All raw videos download from Youtube.
- Test methods: **MISA+MUSEUM** vs. Sa2VA-1B (ft.)
- Testing set
 - test_1_volume.mp4
 - test_2_rhythm.mp4
 - test_3_duration.mp4
 - test_4_firstlast.mp4
 - test_5_badquality.mp4
- In-the-wild videos
 - wild_1_rhythm.mp4
 - wild_2_volume.mp4
 - wild_3_sound.mp4

test_1_volume.mp4

The object making the loudest sound.
MISA+MUSEUM



The object making the lowest sound.
Sa2VA-1B (ft.)



GT: boy; ukulele

test_2_rhythm.mp4

The object making the fastest rhythm.

MISA+MUSEUM



The object making the slowest rhythm.

Sa2VA-1B (ft.)



GT: marimba; piano

test_3_duration.mp4

The object making the longest sound duration

MISA+MUSEUM



The object making the shortest sound duration

Sa2VA-1B (ft.)



GT: **cello**; **piano**

test_4_firstlast.mp4

The first object to make a sound.

MISA+MUSEUM



The last object to make a sound.

Sa2VA-1B (ft.)



GT: **ukulele**; **boy**

test_5_badquality.mp4

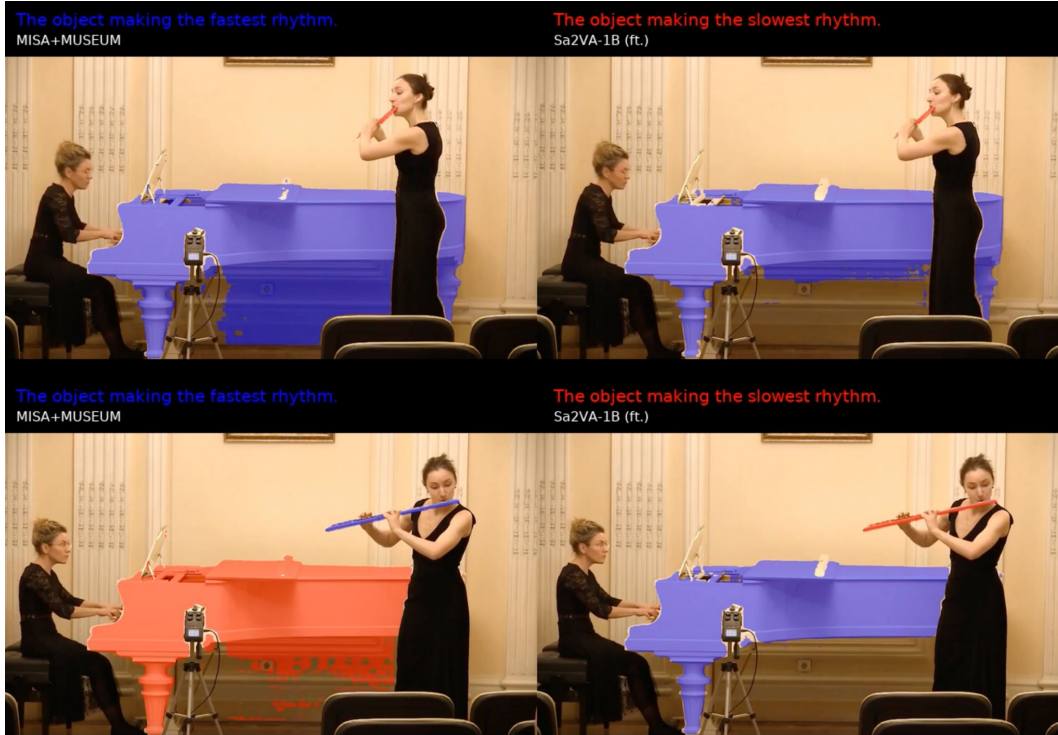
The object making lower sound than the clarinet.
MISA+MUSEUM



GT: **piano**

Showing a correct target, but bad quality example.

wild_1_rhythm.mp4



Piano might be playing in faster rhythm while **flute** might be playing in slower rhythm.

Flute might be playing in faster rhythm while **piano** might be playing in slower rhythm.

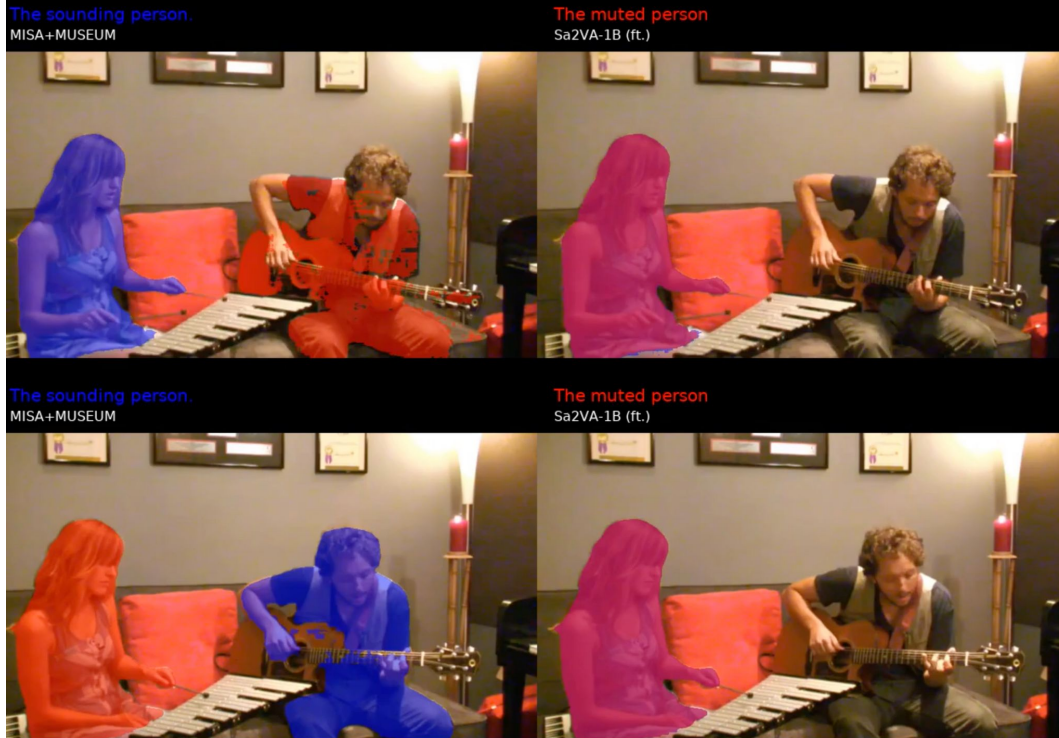
wild_2_volume.mp4



Violin might be playing in loudest sound while **cello** might be playing in lowest sound.

Cello might be playing in loudest sound while **violin** might be playing in lowest sound.

wild_3_sound.mp4



The **woman** should be the one who sings in the scene.

The **man** should be the one who sings in the scene.