Automatic Processing

Manual Processing

Step 1. Conversational Video Collection

14 K Videos Collected

Step 2. Basic Filtering

Raw Videos

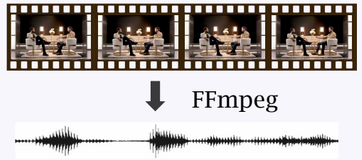
Filtering Rules

1. Category-based filtering
2. Resolution ≥ 480p
3. Language: English

Step 3. Scene Cut

Basic Video Processing

Step 4. Audio Extraction



Step 5. Speech Recognition & Speaker Diarization

- **Text**: *"When we were kids, it's all animation." "There's one thing I really love about…"*
- **Timestamp**: *"start": 0.15, "end": 1.805; "start": 2.315,"end": 2.99*
- **Speaker**: *"SPEAKER_00", …*

Step 6. Audio Filtering

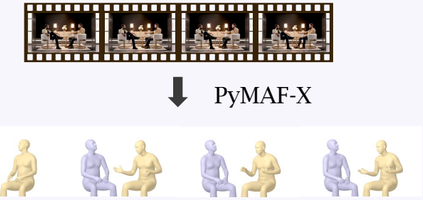
Clips

Filtering Rules

1. Audios are too noisy
2. No word detected

Audio Extraction and Filtering

Step 7. Pose Estimation



Step 8. Pose Filtering

Poses

Filtering Rules

1. Only two people
2. No speaker looking back
3. No missing joint of the upper body
4. Appropriately-sized speakers
5. Relatively noticeable gesture without change of the speaker positions

Step 9. Smoothing

Pose Estimation and Filtering

Step 10. Video Filtering

✗ Non-conversational

✗ Dual-person shots with external voice

Step 11. Pose Filtering

✗ Unnatural

✗ Unnatural

✗ Jittery…

Step 12. Audio&Speaker Alignment

Text	Timestamps	Speaker
"When we were kids, it's all animation."	*"start": 0.15, "end": 1.805*	*"SPEAKER_00"*
"There's one thing I really love about…"	*"start": 2.315, "end": 2.99*	*"SPEAKER_01"*

1. Manual Correction

"SPEAKER_00"

2. Manual Alignment

"SPEAKER_00" : right