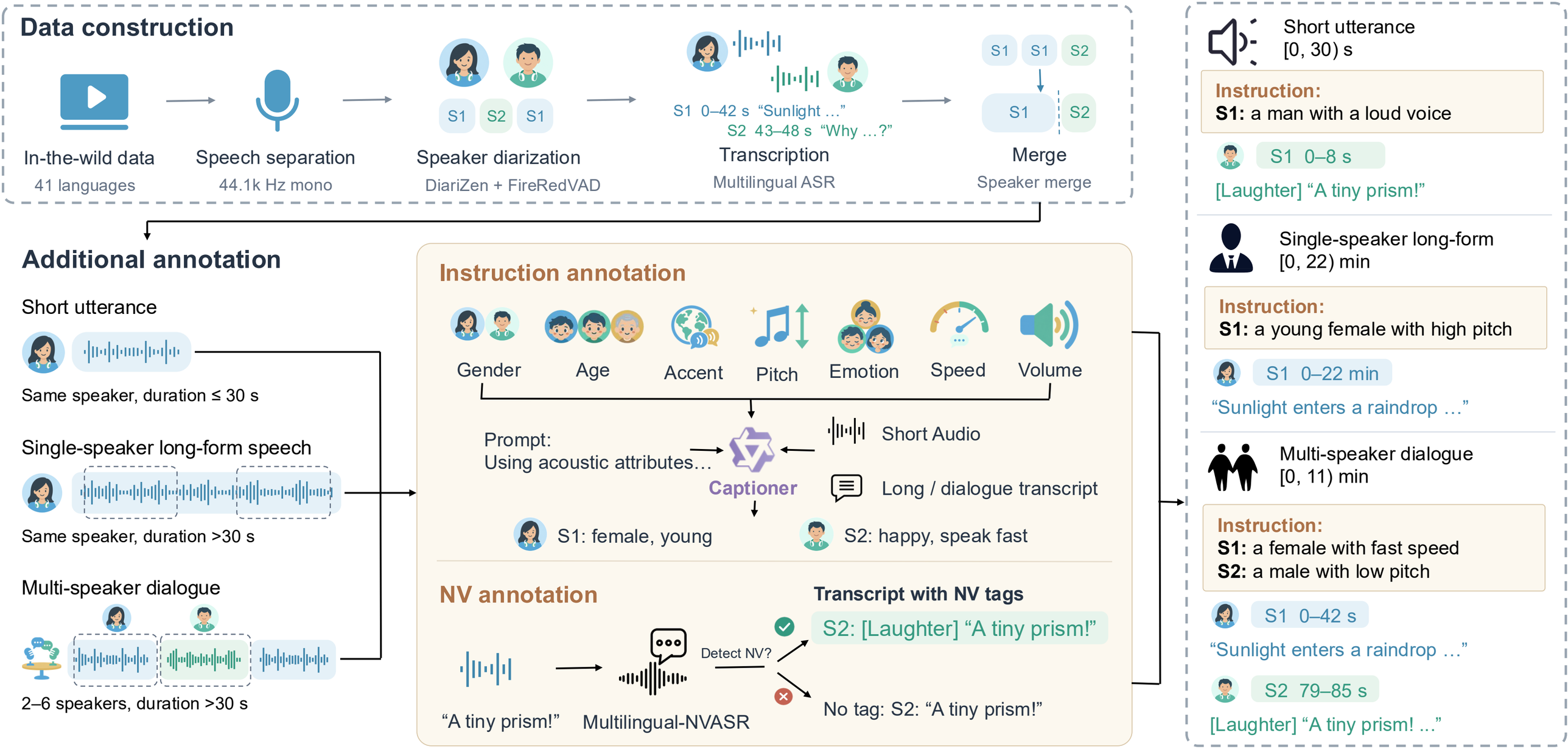

Data construction
In-the-wild data
41 languages
Speech separation
44.1k Hz mono
S1
S2
S1
Speaker diarization
DiariZen + FireRedVAD
S1 0–42 s "Sunlight …"
S2 43–48 s "Why …?"
Transcription
Multilingual ASR
S1
S1
S2
S1
S2
Merge
Speaker merge
Additional annotation
Short utterance
Same speaker, duration ≤ 30 s
Single-speaker long-form speech
Same speaker, duration >30 s
Multi-speaker dialogue
2–6 speakers, duration >30 s
Instruction annotation
Gender
Age
Accent
Pitch
Emotion
Speed
Volume
Prompt:
Using acoustic attributes…
Captioner
Short Audio
Long / dialogue transcript
S1: female, young
S2: happy, speak fast
NV annotation
Transcript with NV tags
Detect NV?
S2: [Laughter] "A tiny prism!"
No tag: S2: "A tiny prism!"
"A tiny prism!"
Multilingual-NVASR
Short utterance
[0, 30) s
Instruction:
S1: a man with a loud voice
S1 0–8 s
[Laughter] "A tiny prism!"
Single-speaker long-form
[0, 22) min
Instruction:
S1: a young female with high pitch
S1 0–22 min
"Sunlight enters a raindrop …"
Multi-speaker dialogue
[0, 11) min
Instruction:
S1: a female with fast speed
S2: a male with low pitch
S1 0–42 s
"Sunlight enters a raindrop …"
S2 79–85 s
[Laughter] "A tiny prism! ..."