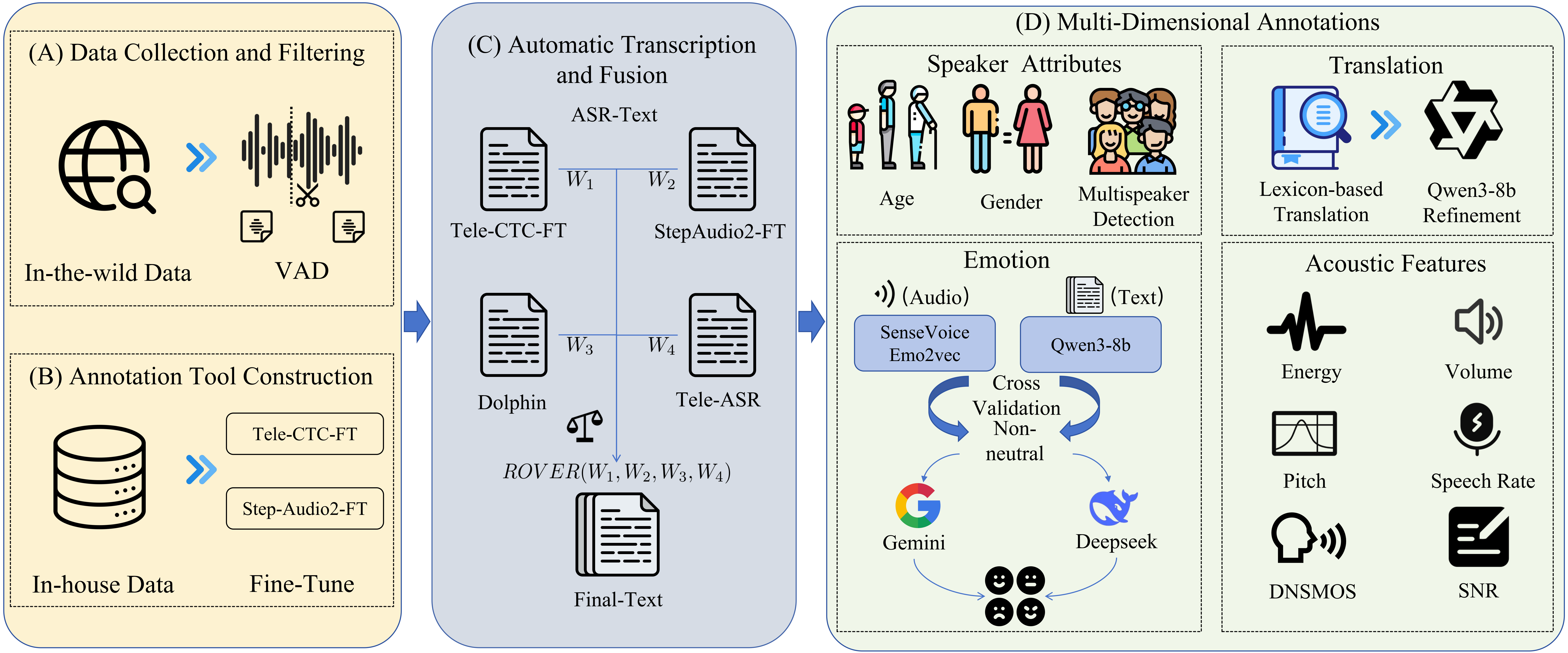
(A) Data Collection and Filtering
In-the-wild Data
VAD
(B) Annotation Tool Construction
Tele-CTC-FT
Step-Audio2-FT
In-house Data
Fine-Tune
(C) Automatic Transcription and Fusion
ASR-Text
W_1
W_2
Tele-CTC-FT
StepAudio2-FT
W_3
W_4
Dolphin
Tele-ASR
$ROVER(W_1, W_2, W_3, W_4)$
Final-Text
(D) Multi-Dimensional Annotations
Speaker Attributes
Age
Gender
Multispeaker Detection
Translation
Lexicon-based Translation
Qwen3-8b Refinement
Emotion
(Audio)
(Text)
SenseVoice Emo2vec
Qwen3-8b
Cross Validation
Non-neutral
Gemini
Deepseek
Acoustic Features
Energy
Volume
Pitch
Speech Rate
DNSMOS
SNR