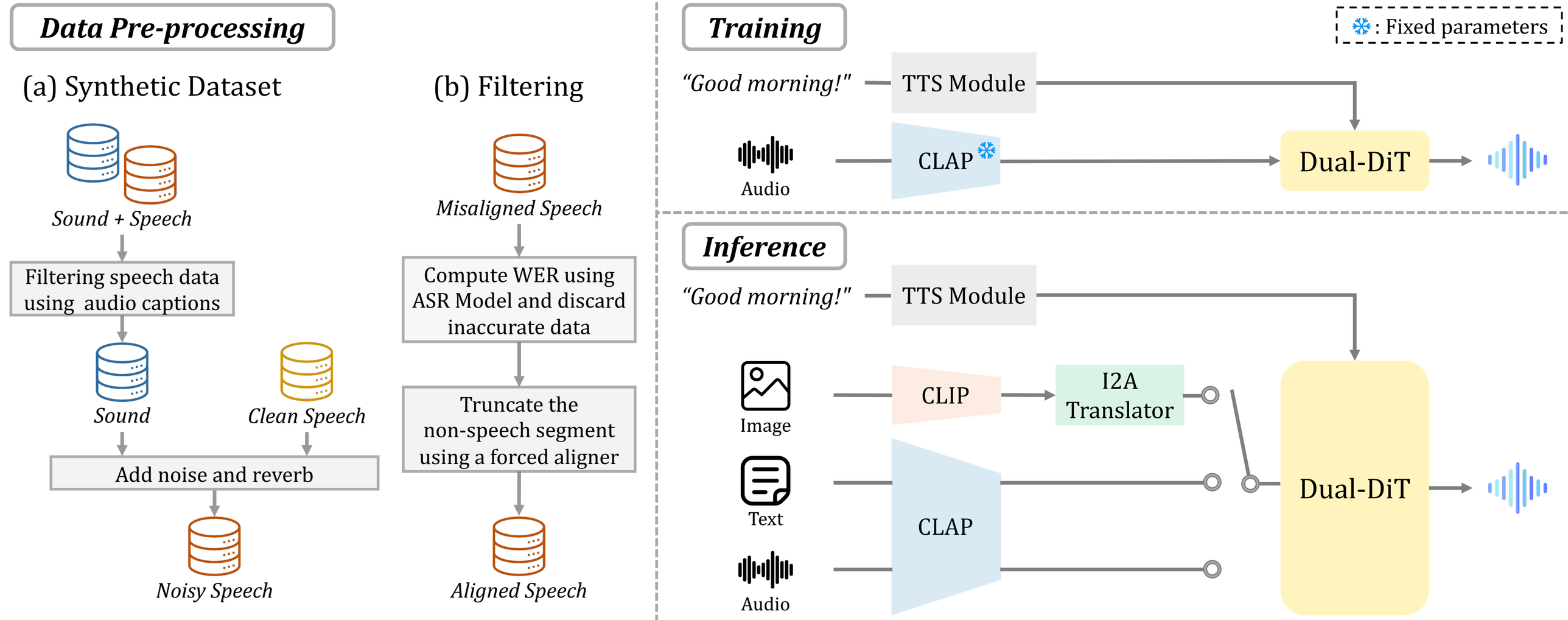
Data Pre-processing
(a) Synthetic Dataset
Sound + Speech
Filtering speech data using audio captions
Sound
Clean Speech
Add noise and reverb
Noisy Speech
(b) Filtering
Misaligned Speech
Compute WER using ASR Model and discard inaccurate data
Truncate the non-speech segment using a forced aligner
Aligned Speech
Training
: Fixed parameters
"Good morning!"
TTS Module
Audio
CLAP
Dual-DiT
Inference
"Good morning!"
TTS Module
Image
CLIP
I2A Translator
Text
CLAP
Audio
Dual-DiT