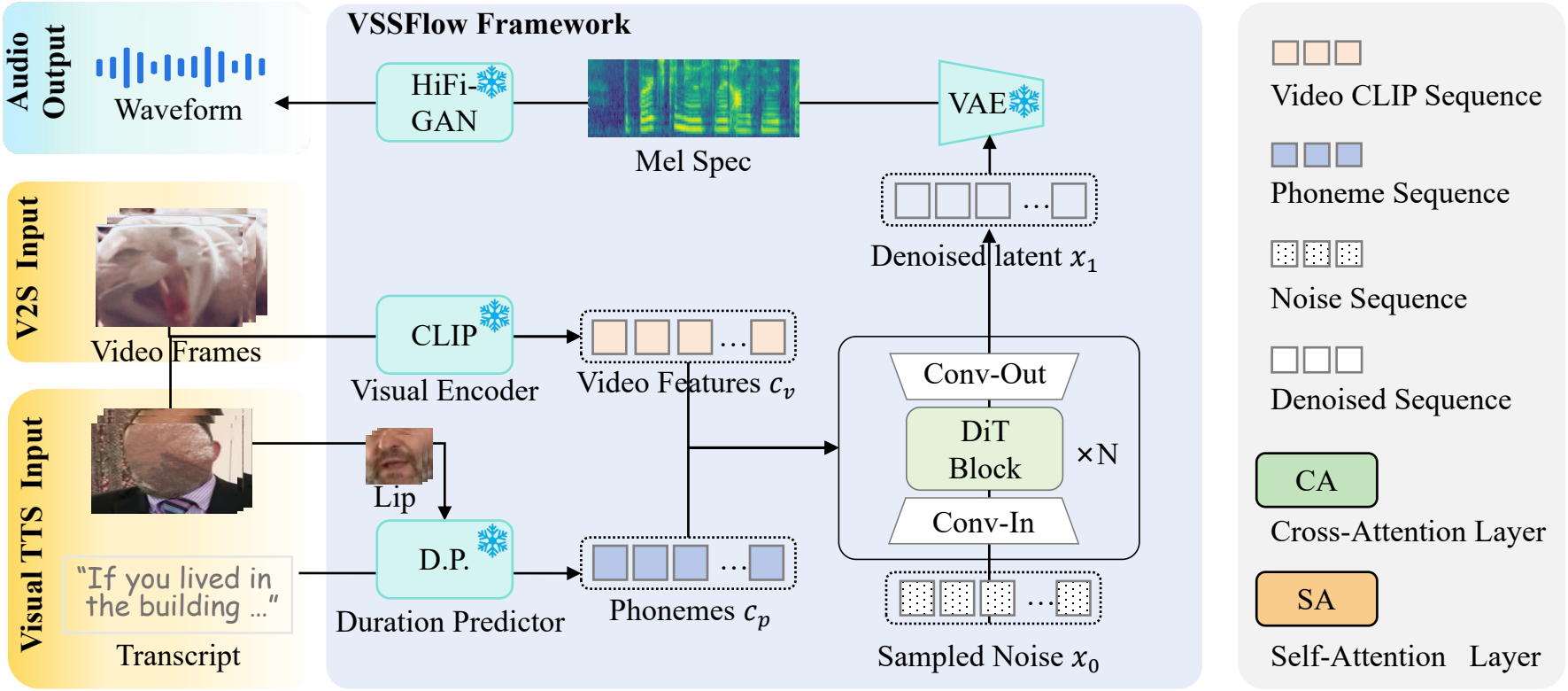
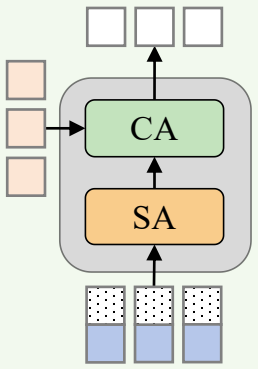
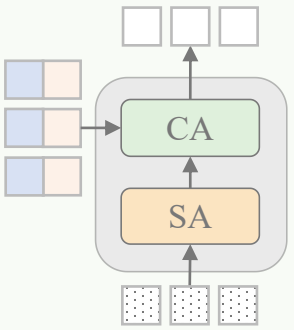


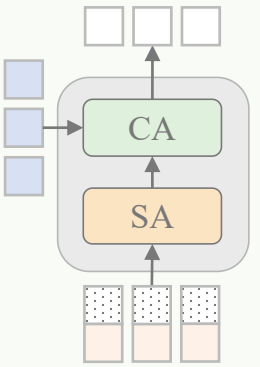
Variants of Condition Aggregation Mechanism



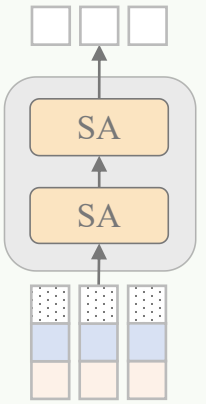
CrossV (Variant1):
- CrossAttn Video CLIP
- Concat Speech Phoneme
- 443M Param.
- sound ✓ speech ✓



CrossVS (Variant2):
- CrossAttn Video CLIP
- CrossAttn Speech Phoneme
- 443M Param.
- sound ✓ speech ✗



ConcatV (Variant3):
- CrossAttn Speech Phoneme
- Concat Video CLIP
- 447M Param.
- sound ✗ speech ✓



ConcatVS (Variant4):
- Concat Video CLIP
- Concat Speech Phoneme
- 486M Param.
- sound ✗ speech ✓