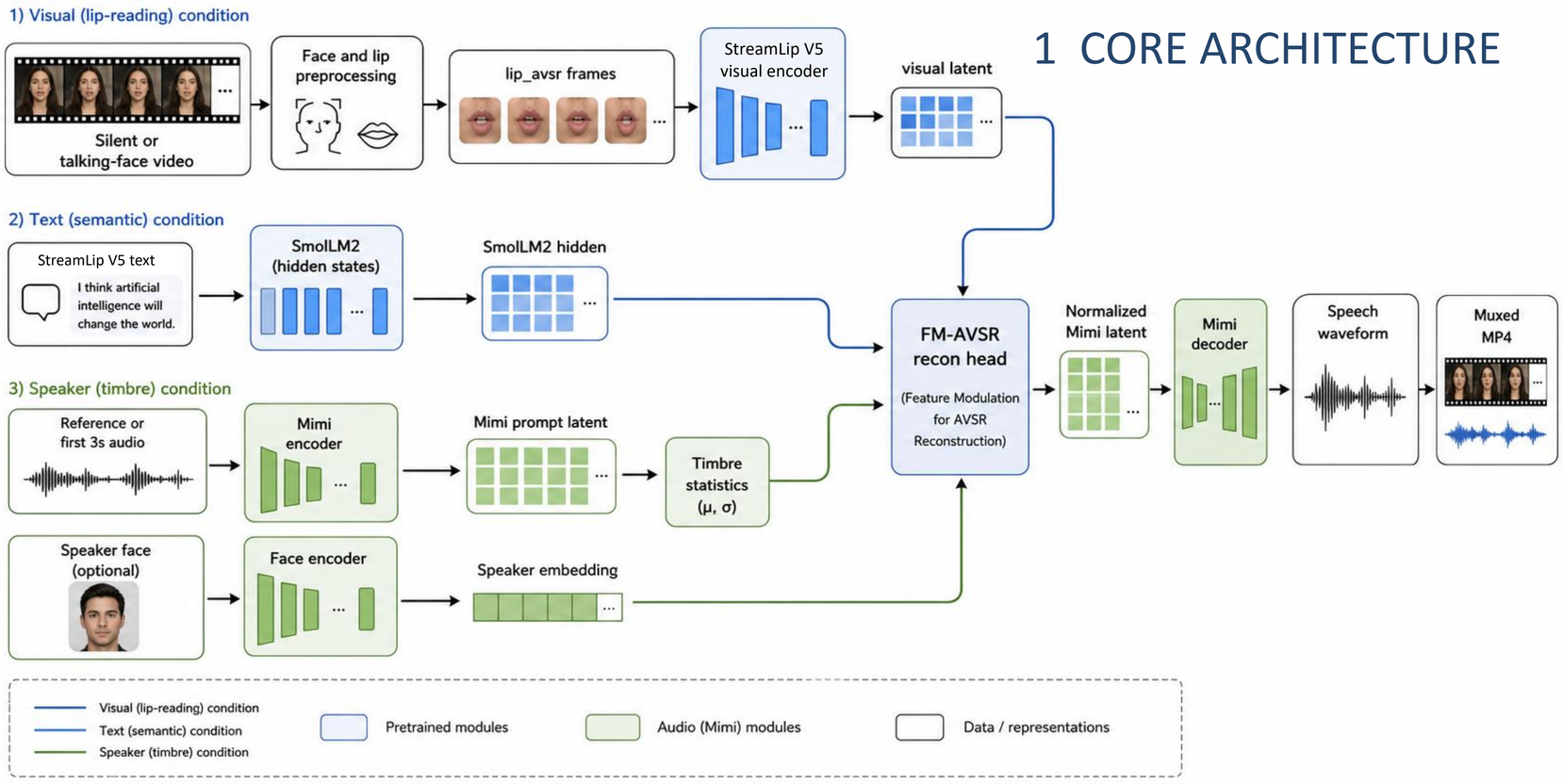
清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

Audio-Centric Visual Speech Reconstruction

Changxun Pan, Zihan Guo, Anqiao Fu

Generate speech by reconstructing Mimi audio latents, not by cascading lip → text → TTS

1 CORE ARCHITECTURE



- Conditions: lip/video | Weak text hidden states | Speaker embedding | 3.04s Mimi audio prompt | Prompt timbre stats
- Target: normalized Mimi latents → waveform via Mimi decoder (offline, 24 kHz mono)
- Key: text is NOT the main generation path — lip and audio-prompt latents carry most of the signal

2 PROJECT OVERVIEW

- **Main Claim**

Do not turn visual speech generation into a strict lip → text → TTS cascade. Text is a weak semantic/alignment cue.

- **Route**

Lip motion, weak LM text, speaker identity, and reference-timbre prompt are fused to reconstruct continuous Mimi audio latents.

- **Evidence**

Trained on 59,144 LRS3 lip-AVSR clips; final checkpoint reaches corr 0.5818 / MSE 0.6676 / MAE 0.6018.

3 DATA + TRAINING SIGNALS

59,144 lip-AVSR clips | fixed 1,000 val

- lip_avsr.npy — Auto-AVSR lip crop
- latentavsr_enc_lipavsr.npy — AVSR visual encoder state
- avsr_enc_lipavsr.npy — AVSR visual encoder state
- avsr_text_lipavsr.txt — noisy text, weak semantic aid
- SmolLM2 hidden — aligned text-token hidden states
- speaker/timbre — speaker_emb + Mimi prompt stats
- target — Mimi audio latent to reconstruct
- lip crops + weak text + timbre prompt

4 TEXT ABLATION

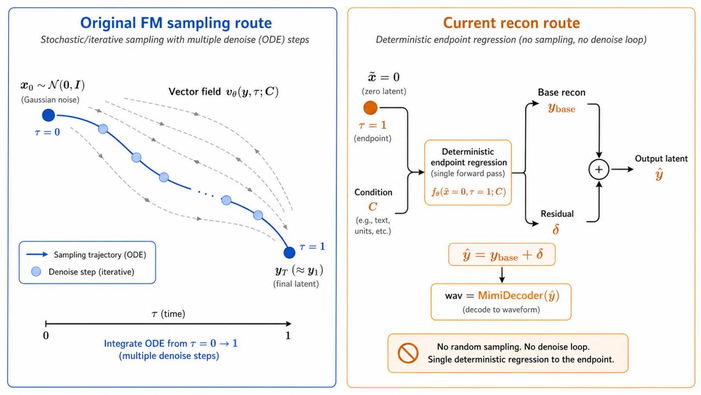
GT transcript is not the core bottleneck.

Text source	Corr	MSE	MAE
GT text + timestamps	0.5818	0.6676	0.6018
AVSR text + timestamps	0.5783	0.6716	0.6037
AVSR text + uniform	0.2868	1.684	1.024
V5 text + uniform	0.2861	1.685	1.024

- AVSR vs GT: Δcorr = 0.0035 only.
- V5 (WER 29%) vs AVSR (WER 20%): Δcorr = 0.0007.
- Alignment quality (timestamps vs uniform) = Δ0.29.

alignment > accuracy

5 RESIDUAL RECON

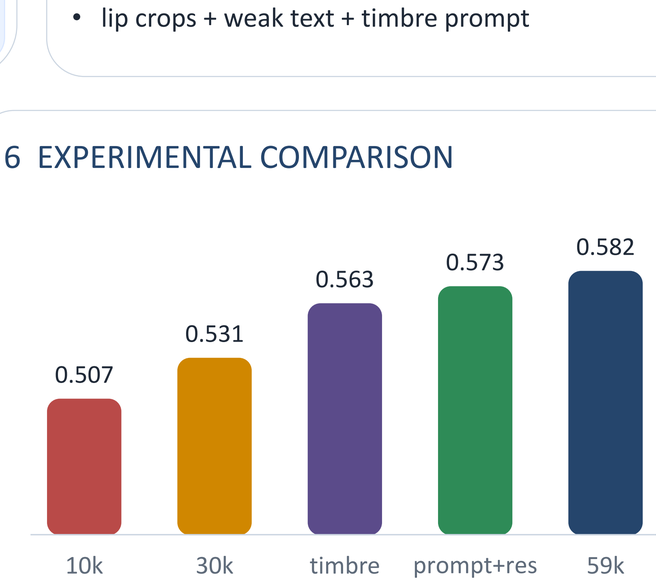


FM sampling (NFE=10) stalls at corr ≈ 0.35.
Fix: set x̃=0, τ=1 — predict endpoint directly.

- Frozen base reconstructor predicts latent.
- Trainable residual head corrects errors. Doubles corr vs naive endpoint.

ŷ = f_base(C) + Δ(C)

6 EXPERIMENTAL COMPARISON



Scale + timbre + residual progression (corr)

- 10k visual prior: 0.507.
- Final 59k prompt-stats residual: 0.582.
- Timbre/audio-prompt conditioning and residual recon produce the larger gains.

Final corr 0.5818 | MSE 0.6676 | MAE 0.6018

7 STREAMLIP V5: SELF-TRAINED VSR BRANCH

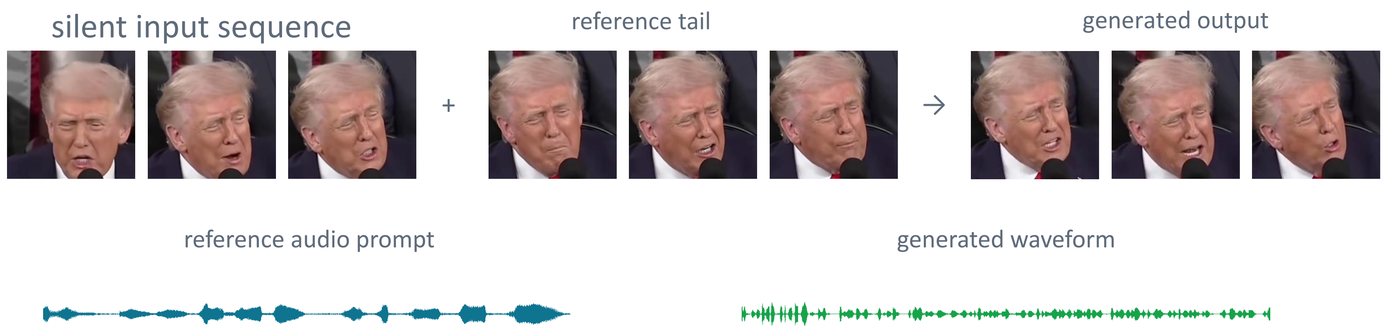
Architecture

lip frames (96×96 gray) → Conformer ×12 (frozen) → proj + Gated CrossAttn ×4 → OLMo-1B (fine-tuned) → transcript (beam=3)

- Data: LRS3 pretrain split ~115k, 102k left after filtering out the blurry samples.
- Model: pretrained AV-HuBERT video conformer encoder + OLMo-1B language model. Text tokens uses Flamingo-styled Gated Cross-Attention to query the conformer KV pairs and to integrate visual information.
- Training: first fine-tune the LM on texts of samples in LRS3 pretrain split, then train the GCA weights and fine-tune the LM using standard CE loss.
- Test: test on the LRS3 test split. StreamLip V5 shows capability on long sentences, grammar correctness and low frequency words.

System	WER	Word Acc	Notes
Auto-AVSR (beam=10)	20.2%	79.8%	CTC+Att, 5k vocab
StreamLip V5 (beam=10)	28.6%	71.4%	OLMo-1B, 50k vocab

8 SILENT → VOICED DEMO



Silent video + reference audio → AVSR text → Mimi recon → Voiced MP4
Trump corr: 0.3035 → 0.3869 with final lip-AVSR preprocessing path

9 TAKEAWAY

Do not cascade lip → text → TTS.

- Lip motion provides phonetic and timing evidence.
- Mimi latents preserve codec-compatible acoustic structure.
- Weak text stabilizes content — does not need to be perfect.
- Timbre/reference conditions carry speaker and recording color.
- StreamLip V5: LM-based VSR can serve as self-trained text branch.

10 REPRODUCE / ARTIFACTS

- One-command raw-video demo:

python scripts/run_raw_video_avsr_recon_pipeline.py --input data/trump.mov --exp demo --force

- Gradio GUI

scripts/gradio_avsr_gui.py

- Checkpoint

Timbre-fix residual recon: step_002000.pt
Fine-tuned OLMo 1B: olmo-1b-lrs3-lr3e-5_ep2
StreamLip V5 ckpt: v5_olmo_lr1e-6_ep50_eos_frame500_warmup0/step_001500.pt