StreamLip: Audio-Centric Visual Speech Reconstruction with Weak Text Conditioning

Zihan Guo **2024011317**
Changxun Pan **2024011323**
Anqiao Fu **2024010753**

Abstract

This report summarizes the final StreamLip audio reconstruction system. The key finding is that high-quality visual speech audio generation should not be framed as a strict vision-to-text-to-audio cascade. In our final system, text is a weak semantic and alignment cue; the main predictive path is from lip motion, speaker information, and audio-derived timbre conditions to continuous Mimi audio latents. This design makes exact transcript accuracy non-critical. Replacing ground-truth text with an automatically decoded transcript changes full-validation latent correlation from 0.5818 to 0.5783, a drop of only 0.0035, while explicit timbre/audio prompt conditioning and residual Mimi-latent reconstruction provide the major gains. We additionally train StreamLip V5, an LM-based visual speech recognizer, to make the text branch self-contained and to show that language-model text quality is not the limiting factor for audio reconstruction. The strongest checkpoint reaches 0.58184180 mean correlation on the LRS3 validation split and supports raw-video and silent-video inference with optional reference audio. Project page: `https://2omegaxv.github.io/StreamLip/`. Code: `https://github.com/2omegaXv/StreamLip`.

1 Goal and Main Claim

The project goal is to generate speech audio from talking-face video. The final system is an offline research prototype: it predicts Mimi audio latents from visual speech conditions and decodes them into waveform. It is not positioned as a real-time product system or a full speaker-cloning model.

The central conclusion is architectural. A direct *vision* $\rightarrow$ *text* $\rightarrow$ *audio* pipeline is too dependent on transcript quality and discards acoustic information that is visible in lip dynamics but not represented by text. Our final route instead uses:

$$\text{lip/video} + \text{weak text} + \text{speaker/timbre} \longrightarrow \text{Mimi audio latent} \longrightarrow \text{waveform}. \tag{1}$$

Text helps stabilize content, but it is not the main representation used to generate audio. This explains why imperfect transcripts from either an external baseline or StreamLip V5 can still produce usable speech: the model is not forced to reconstruct audio from text alone.

The final technical contributions are:

1. **Audio-centric latent formulation.** We formulate visual speech generation as reconstruction of normalized continuous Mimi latents rather than as TTS from a transcript. This preserves acoustic information beyond text and gives a direct objective for waveform reconstruction.
2. **Residual endpoint reconstruction architecture.** The final model abandons stochastic FM sampling for a deterministic endpoint predictor with a frozen base reconstructor and a learned residual correction.
3. **Explicit timbre/audio prompt conditioning.** The first 38 Mimi frames, about 3.04 seconds, are used both as prompt tokens and as global mean/std timbre statistics. This provides speaker and recording-color cues that text cannot provide.

Preprint. Under review.

4. **Self-trained StreamLip V5 text branch.** We train an LM-based VSR model that injects frozen visual speech features into OLMo-1B through gated cross-attention. This makes the text branch an internal component of the project rather than only an external dependency.
5. **Weak text conditioning finding.** Ground-truth text, baseline decoded text, and StreamLip V5 text lead to very similar audio metrics when alignment conditions are matched. This shows that exact transcript accuracy is not the dominant bottleneck; the main load is carried by lip and audio-latent conditions.
6. **Data processing as model evidence.** The final results depend on a consistent data representation: StreamLip-compatible lip crops, visual speech encoder states, language-model hidden states, Mimi targets, speaker features, and prompt-derived timbre conditions.

2 Data Processing and Data Use

Data construction is a core part of the method. Each clip is converted into a matched set of visual, text, speaker, and audio-latent files. The model never uses raw text alone as the generation target; all training and evaluation are defined against Mimi Défossez et al. (2024) continuous latents.

2.1 Training and Validation Splits

The final checkpoint is trained on the StreamLip-compatible processed LRS3 Afouras et al. (2018) pretrain split with 59,144 training clips and evaluated on a fixed 1,000-clip validation split:

- train split: `configs/eval_splits/pretrain_lipavsr_train59144_seed44_plus_ready_remaining9144_excl_val1000_seed43.txt`
- validation split: `configs/eval_splits/pretrain_len80_260_lipavsr_val1000_seed43.txt`

Earlier 10k and 30k subsets were used to validate scale trends and ablations before moving to the final 59k setting.

2.2 Per-Clip Files

File	Shape	Use in the model
`lip_avsr.npy`	$(T_v, 96, 96)$	StreamLip-compatible grayscale mouth crop
`avsr_enc_lipavsr.npy`	$(T_v, 768)$	visual speech encoder state
`avsr_text_lipavsr.txt`	text	baseline transcript for compatibility checks
`streamlip_v5_text.txt`	text	StreamLip V5 decoded transcript
`smollm2_h_text_json.npy`	$(L, 960)$	GT-text hidden states for in-domain eval
`smollm2_h_v5.npy`	$(L, 960)$	StreamLip V5 hidden states for raw inference
`speaker_emb.npy`	(256)	face-based identity condition
`latent.npz`	$(T_a, 512)$	Mimi latent training target
`audio_prompt.npy`	$(38, 512)$	Mimi prompt tokens, first 3.04 s
`timbre_cond.npy`	(1024)	prompt mean/std timbre condition

Table 1: Final per-clip representation. The model consumes conditions and predicts the normalized Mimi latent target.

2.3 Visual Processing

The visual path must use the StreamLip-compatible grayscale mouth crop `lip_avsr.npy`. The older RGB lip crop is not the final representation. The correct crop is encoded by the frozen visual speech encoder used by StreamLip:

$$e^v_{1:T_v} = E_{\text{vis}}(V_{\text{lip}}), \qquad e^v_t \in \mathbb{R}^{768}. \tag{2}$$

In the current implementation, E_{vis} is initialized from the Auto-AVSR Ma et al. (2023) Conformer-style visual speech encoder and kept frozen. The project contribution is the downstream StreamLip text branch and the audio-latent reconstruction model built on this representation. For external videos

without ground-truth text or timestamps, decoded text can come from the baseline decoder or from StreamLip V5.

This processing choice is important for generalization. On external raw videos, switching from the old RGB lip path to the correct `lip_avsr` path improved latent correlation:

Video	Old RGB lip path	StreamLip visual path
Trump	0.30353260	0.38691377
HRX	0.25592437	0.27240886

Table 2: External-video checks showing why the final StreamLip visual representation matters. This is reported as data processing evidence rather than as a standalone preprocessing contribution.

2.4 Audio Latent Target

Mimi Défossez et al. (2024) provides the continuous audio latent target. For each audio-latent frame $z_t \in \mathbb{R}^{512}$, the training target is normalized per latent dimension:

$$y_t = \frac{z_t - \mu}{\sigma}. \tag{3}$$

The model predicts $\hat{y}_{1:T}$, then denormalizes and decodes:

$$\hat{z}_t = \hat{y}_t \sigma + \mu, \qquad \hat{x}_{\text{wav}} = \text{MimiDecoder}(\hat{z}_{1:T}). \tag{4}$$

All reported MSE, MAE, and correlation metrics are computed in normalized Mimi latent space, not waveform space.

2.5 Text and Timbre Conditions

Text is encoded by SmolLM2 Ben Allal et al. (2025) into hidden states. For LRS3 validation with ground-truth word timestamps, audio-latent frames are aligned to text positions using word timestamps. For raw videos, timestamps are unavailable, so the hidden states are uniformly resampled. This makes text usable even when it is not perfectly recognized or perfectly aligned.

The timbre condition is derived from the first 38 normalized Mimi frames $A \in \mathbb{R}^{38\times512}$:

$$q = [\text{mean}(A), \text{std}(A)] \in \mathbb{R}^{1024}. \tag{5}$$

The same A is also provided as prompt tokens. Listening examples remove this 3.04-second prompt region when the prompt comes from the same clip, because the first seconds are condition input rather than generated speech.

3 Architecture and Mathematical Formulation

3.1 Condition Fusion

Figure 1 shows the final inference graph. The model condition is:

$$C = \{e^v_{1:T_v}, h_{1:L}, s, A, q\}, \tag{6}$$

where e^v is the StreamLip visual speech sequence, h is the language-model text hidden sequence, s is the speaker embedding, A is the audio prompt token sequence, and q is the prompt statistics timbre vector.

The visual speech features are produced at about 25 Hz, while Mimi latents are used at 12.5 Hz. The visual condition is aligned by taking every second visual frame:

$$v_t = e^v_{2t}. \tag{7}$$

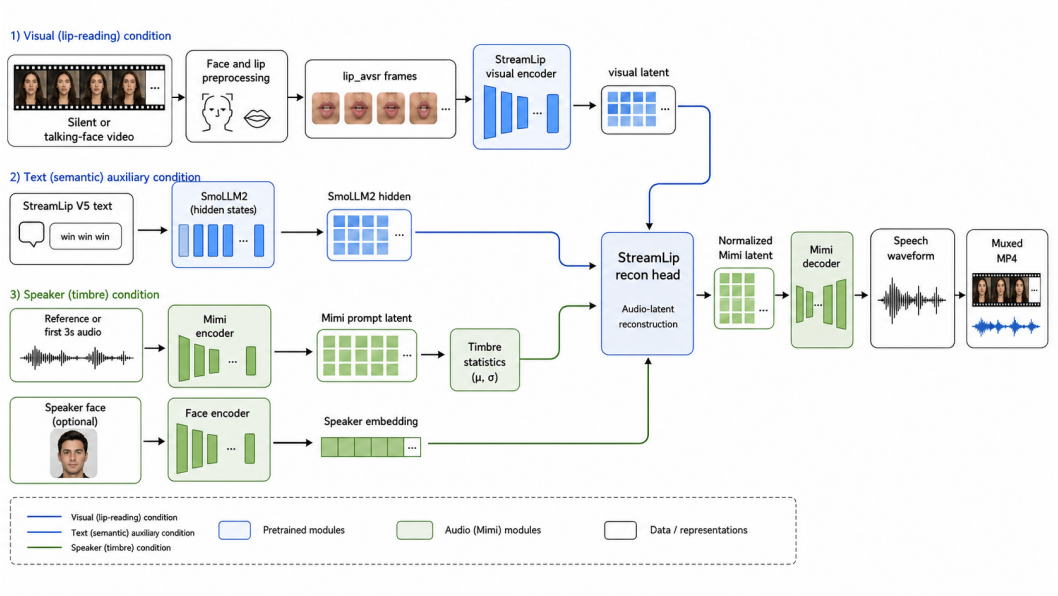

Figure 1: Final audio-centric visual speech reconstruction architecture. Lip features, weak text hidden states, speaker identity, and audio-prompt timbre conditions are fused to predict normalized Mimi latents.

3.2 Deterministic Residual Reconstruction

The explored flow-matching route learned a vector field from noise to target. It did not become competitive: sampling correlation stayed around 0.35, far below the reconstruction route. The final model therefore fixes the inference state to:

$$\tilde{x} = 0, \qquad \tau = 1, \tag{8}$$

and directly predicts the endpoint latent. There is no random noise input and no iterative denoising loop.

The final checkpoint uses a frozen base reconstructor and a residual head:

$$y_{\text{base}} = f_{\text{base}}(C), \qquad \Delta = f_{\text{residual}}(C), \qquad \hat{y} = y_{\text{base}} + \Delta. \tag{9}$$

This residual design is the main architecture-level improvement because it lets the final model refine a strong endpoint predictor instead of relearning the whole mapping from scratch.

3.3 Training Objective

The final loss combines reconstruction, sample-level correlation, and prompt timbre statistics:

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + 0.2\,\mathcal{L}_{\text{corr}} + 0.05\,\mathcal{L}_{\text{prompt\_stats}}. \tag{10}$$

The correlation term is computed per clip after skipping the first 38 prompt frames:

$$\mathcal{L}_{\text{corr}} = 1 - \text{corr}\left(\text{vec}(\hat{y}_{39:T}), \text{vec}(y_{39:T})\right). \tag{11}$$

The prompt-statistics term encourages the generated segment to stay close to the timbre statistics implied by the prompt.

4 StreamLip V5: LM-Based Visual Speech Recognition

StreamLip V5 is the project's self-trained visual speech recognition (VSR) branch. It replaces a small task-specific decoder with a large language model conditioned on frozen visual speech features. In the final system, V5 is used to demonstrate that the text branch can be internal to StreamLip, while the audio reconstructor remains robust even when transcript accuracy is imperfect.

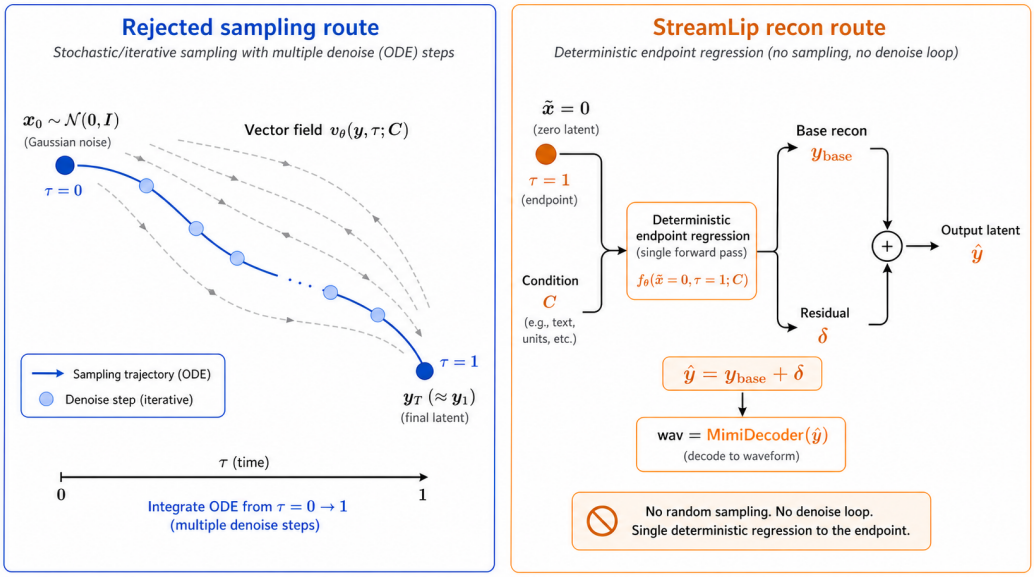

Figure 2: The final route is deterministic endpoint reconstruction, not stochastic FM sampling. The model fixes $\tilde{x} = 0$ and $\tau = 1$, then predicts Mimi latents through a frozen base plus residual correction.

4.1 Architecture

StreamLip V5 uses a two-stage design:

$$\text{lip frames} \xrightarrow{\text{frozen visual speech encoder}} e^v_{1:T} \in \mathbb{R}^{T \times 768} \xrightarrow{\text{proj+Gated Cross-Attn}} \text{OLMo-1B} \longrightarrow \hat{w}_{1:N} \tag{12}$$

The visual encoder is the frozen Conformer-style visual speech encoder used by our StreamLip preprocessing path. Its output features are projected to the OLMo-1B Groeneveld et al. (2024) hidden dimension (2048) and injected into every 4th transformer layer via Flamingo-style Gated Cross-Attention Alayrac et al. (2022):

$$x \leftarrow x + \tanh(\alpha) \cdot \text{CrossAttn}(x,\, e^v), \qquad \alpha \text{ initialized to } 0.1. \tag{13}$$

Only the OLMo-1B weights and the projection/cross-attention parameters are trained; the Conformer encoder is not updated.

Prior to V5 training, OLMo-1B is fine-tuned on the LRS3 transcript corpus (`lrs3_text.txt`, 147,930 lines, chunk length 256 tokens) using next-token prediction. A learning-rate and epoch sweep is run to find the best domain-adaptation checkpoint; Table 3 shows representative results. The base OLMo-1B reaches PPL 65.5 on the LRS3 validation split; the selected checkpoint (lr $= 3 \times 10^{-5}$, 2 epochs) achieves the best PPL of 25.9. High learning rates ($\geq 10^{-3}$) overfit quickly, while lr $= 10^{-5}$ plateau at PPL$\approx$26.1 regardless of epoch count, suggesting the model is near its adaptation limit on this data at that rate.

The full V5 model has approximately 1.24B parameters, of which 1.23B are trainable.

4.2 Training

Training uses the LRS3 pretrain split with cross-entropy loss over subword tokens. Table 4 summarises the ablation study that guided the final configuration.

EOS supervision. Without an explicit end-of-sequence target the model relies on `max_new_tokens` truncation; adding an EOS token to every target sequence has a dramatic effect on WER. On the 200-clip LRS3 test set with beam=3, the no-EOS model reaches WER 89.0% because it does not learn

LR	Epochs	Val PPL
10^{-3}	2	92.8
10^{-3}	3	71.8
3×10^{-4}	2	45.6
10^{-4}	3	31.2
10^{-5}	2	26.5
10^{-5}	5	26.1
3×10^{-5}	1	26.3
$\mathbf{3 \times 10^{-5}}$	**2**	**25.9**
3×10^{-5}	3	26.5

Table 3: OLMo-1B LRS3 domain-adaptation sweep. Base OLMo-1B PPL = 65.5. The selected checkpoint (lr $= 3 \times 10^{-5}$, 2 epochs, **bold**) is used for all V5 experiments.

to stop and keeps generating hallucinated tokens beyond the clip boundary (e.g. GT: *"and in many cases we don't"*; no-EOS output: *"and in many cases we don't the reason being is we're"*). Adding EOS supervision reduces WER to 31.4%, a drop of 57.6 pp, by teaching the model to terminate naturally. The val tok-accuracy improvement is modest (0.8585 $\rightarrow$ 0.8650), because per-token accuracy is not sensitive to where the sequence ends, but the WER impact is severe without it.

Maximum clip length. Raising `max_frames` from 150 to 400 drops the best validation CE from 0.675 to 0.553, a significant gain because the model sees longer and more varied visual contexts during training. Increasing further to 500 yields a smaller additional improvement (0.553 $\rightarrow$ 0.546), suggesting diminishing returns beyond 400 frames.

Learning rate. At max_frames $= 150$, lr $= 3 \times 10^{-6}$ outperforms lr $= 10^{-5}$ slightly (CE 0.664 vs. 0.675). At max_frames $= 500$, lr $= 10^{-6}$ gives the best results, consistent with larger batches (longer clips fill more GPU memory) requiring a smaller learning rate.

Warmup. All runs use a cosine decay schedule. Without warmup (warmup_epochs $= 0$) the model starts at full LR immediately; this actually achieves the lowest CE (0.527) and converges faster (best at step 1500 vs. 2000 for the default warmup of 3 epochs). Longer warmup (6 epochs) increases the warm-up phase tokens seen at low LR, which hurts early convergence (best CE 0.544 vs. 0.527).

Config	lr	max_frames	warmup	best val CE	best tok acc
no EOS	10^{-5}	150	3.0	0.6715	0.858
+ EOS	10^{-5}	150	3.0	0.6754	0.865
lower lr	3×10^{-6}	150	3.0	0.6636	0.867
+ max_frames 400	10^{-5}	400	3.0	0.5533	0.879
+ max_frames 400	3×10^{-6}	400	3.0	0.5563	0.883
max_frames 500, warmup=3 (default)	10^{-6}	500	3.0	0.5462	0.889
max_frames 500, warmup=0	10^{-6}	500	0.0	**0.5273**	0.887
max_frames 500, warmup=1	10^{-6}	500	1.0	0.5373	0.877
max_frames 500, warmup=6	10^{-6}	500	6.0	0.5440	0.882

Table 4: StreamLip V5 training ablation on LRS3 pretrain split. All runs use AdamW ($\beta = (0.9, 0.98)$, weight decay 0.01), batch size 256, 50 epochs, cosine schedule. The final checkpoint used for evaluation is `v5_olmo_lr1e-6_ep50_eos_frame500_warmup0`, step 1500.

The three parameter groups used in all runs are: pretrained OLMo layers (lr_{LM}, as shown above), new projection and cross-attention weights (3×10^{-4}), and the tanh gates (10^{-4}, ten times the LM rate to prevent saturation).

4.3 Recognition Performance

Table 5 shows a controlled VSR comparison on 200 LRS3 test clips (seed 42). StreamLip V5 uses beam search with beam width 3, which is the empirically optimal setting: larger beams allow the LM to produce fluent but visually under-grounded candidates that hurt WER.

System	WER	Word Acc	Notes
External VSR baseline	20.2%	79.8%	CTC+Att, beam=40
StreamLip V5 (warmup0, step 1500)	29.2%	70.8%	OLMo-1B, beam=3

Table 5: WER comparison on a 200-clip LRS3 test subset (seed=42). StreamLip V5 is not yet the strongest VSR recognizer, but it makes the text branch self-trained and exposes where LM-based visual decoding helps or fails.

The 9.0pp gap is primarily attributable to the frozen visual encoder and the lack of a CTC-style monotonic alignment head. The visual features are not optimized jointly with the LM, leaving a representational gap between the two components. However, per-sample analysis shows the intended advantage: V5 is competitive on longer sentences ($\geq$10 words, $\geq$4 s), where LM context accumulates and can correct substitution errors in rare words, fixed collocations, and contracted forms (e.g. *we're*, *they've*). Its large-vocabulary prior (50,280 BPE tokens) better preserves contracted forms and rare lexical choices. On shorter clips ($<$4 s), LM context is insufficient and insertion errors dominate because V5 lacks CTC monotonic alignment.

5 Why Text Does Not Need to Be Exact

The final architecture treats text as an auxiliary semantic cue, not as the central audio representation. This is a deliberate difference from a cascaded vision-to-text-to-audio system.

In a cascade, recognition errors in the text stage directly become content errors in the audio stage. In our model, lip motion and audio-latent prompt conditions remain visible to the reconstructor. The text hidden state provides coarse semantic guidance and helps stabilize alignment, but it does not have to carry phonetics, speaker timbre, timing, and acoustic detail by itself.

The strongest evidence is the text-source ablation:

Text condition	Alignment	Corr	MSE	MAE
Ground-truth text	word timestamps	0.58184	0.66763	0.60182
Baseline decoded text	word timestamps	0.57833	0.67162	0.60371
Baseline decoded text	uniform	0.28679	1.68438	1.02352
StreamLip V5 text	uniform	0.28609	1.68488	1.02375

Table 6: Full LRS3 validation metrics on 1,000 clips. The top two rows use word-level timestamps; the bottom two use uniform resampling. Under matched alignment conditions, replacing ground-truth text with baseline decoded text (WER 20.2%) drops correlation by only 0.00351, and replacing it with StreamLip V5 text (WER 29.2%) drops it by only 0.00070 further—a difference smaller than the GT-vs-decoded gap. The large drop from word-timestamps to uniform ($\Delta \approx 0.29$) dominates everything else, confirming that alignment quality is the primary bottleneck, not transcript accuracy. Closing the word-timestamps gap for V5 would require retraining the audio reconstructor on the V5 text distribution; simply supplying better timestamps while keeping the current model is insufficient, because the model's hidden-state conditioning was trained exclusively on the decoded baseline distribution.

This does not mean text is useless. It means transcript accuracy is not the limiting factor in this architecture. Even when the text is imperfect, generated audio can remain perceptually acceptable if the lip sequence and audio-latent conditions are matched.

The table also reveals a more important finding: **text alignment dominates text accuracy.** Under uniform resampling (no timestamps), correlation drops from $\approx$0.58 to $\approx$0.29 regardless of whether the text comes from ground-truth, the external decoded baseline, or StreamLip V5. The difference

between the external decoded baseline and V5 text under the same uniform alignment is only 0.00070, smaller than the GT-vs-decoded text gap of 0.00351. This means transcript accuracy at the current WER level is essentially not a factor; alignment quality is.

Therefore, the project focus should remain on visual speech representation, audio latent modeling, and timbre control.

6 Experimental Results

6.1 Final Checkpoint

- **Config:** `configs/fm_avsr_lipavsr_59144_timbre3s_audioprompt38_pool_promptstats005_residual_samplecorr02_lossstart38_from1500_recon_textjson_wordts.yaml`
- **Checkpoint:** `ckpt/recon/streamlip_recon_timbrefix_step_002000.pt`
- **Residual base:** `ckpt/recon/streamlip_recon_residual_base_step_005000.pt`

6.2 Scale and Timbre Ablations

Setup	Corr	MSE	MAE
10k visual-prior baseline	0.50725122	0.77212131	0.65607136
30k visual-prior baseline	0.53081070	0.74379597	0.64330638
30k no-timbre baseline	0.53298334	0.72177195	0.63384026
30k mean/std timbre	0.56328889	0.68858632	0.61404544
30k prompt tokens + pooled prompt + residual	0.57255184	0.67793650	0.60844724
Final 59k prompt-stats residual	0.58184180	0.66763106	0.60181897

Table 7: Main metric progression. Data scale improves the visual-prior model, and timbre/audio prompt conditioning gives a larger gain than text-source changes.

The 10k to 30k scale-up improved matched validation correlation by about 0.0236. Timbre conditioning gave a larger change: the 30k no-timbre baseline at 0.5330 moved to 0.5633 with mean/std timbre, and the prompt/residual route reached 0.5726. The final 59k residual prompt-statistics checkpoint reached 0.5818.

The shifted-condition negative control for the 30k audio-prompt model dropped to 0.00799 correlation, confirming that the model uses the matched conditions rather than predicting an average latent.

6.3 Raw and Silent Video Inference

External raw-video inference uses decoded text and uniform text alignment because no ground-truth transcript or word timestamps exist. The current demo path can use the compatibility decoder or StreamLip V5 as the text source; the central audio path is unchanged because text remains a weak auxiliary condition.

For silent video, the video supplies lip and face conditions while an optional reference audio supplies A and q. If no reference audio is provided, the audio prompt and timbre vector are zero. The current demo export drops the first 3.04 seconds because the model can copy reference-prompt audio into the beginning of the generated waveform. This is a temporary export-time hack and is listed as a model issue to fix.

7 Reproducibility Interfaces

The public project page is available at `https://2omegaxv.github.io/StreamLip/`. The code and model instructions are available at `https://github.com/2omegaXv/StreamLip`.

The primary command-line entry point is:

```
python scripts/run_raw_video_avsr_recon_pipeline.py \
  --input data/trump.mov --exp trump_cleanup_verify --force
```

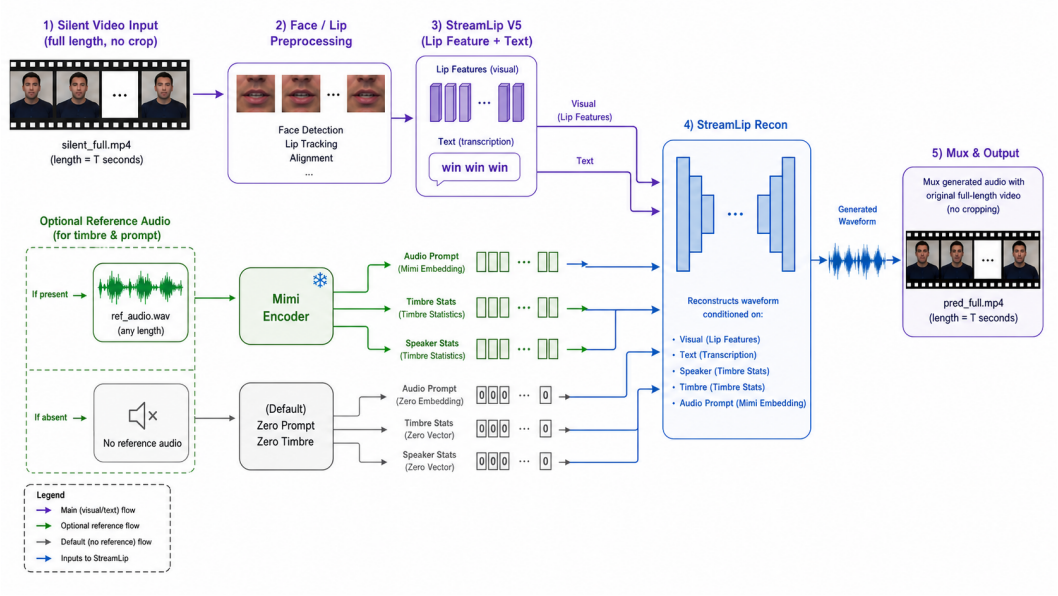

Figure 3: Silent-video inference with optional reference audio. This is an application of the same audio-prompt/timbre conditioning mechanism used in training. Current listening exports skip the first 3.04 seconds to avoid reference-prompt leakage at the start.

For silent video with reference audio:

```
python scripts/run_raw_video_avsr_recon_pipeline.py \
  --input silent.mp4 --ref_audio ref.wav \
  --silent_input --exp demo --force
```

The Gradio interface at `scripts/gradio_avsr_gui.py` exposes the same backend with video upload, optional reference audio, and silent-input controls.

These interfaces are not counted as research contributions; they are included so the result can be reproduced and demonstrated.

8 Limitations

- The system is offline. Earlier timing showed that face alignment and lip cropping dominate runtime; the FM head and Mimi decoder are not the primary bottlenecks.
- The model is deterministic. It improves stability, but it can average speaker-specific acoustic detail, which explains some mixed-voice artifacts.
- Timbre control is based on prompt latents and prompt statistics. It is useful, but it is not equivalent to a high-fidelity speaker cloning model.
- The current silent-reference export skips the first 3.04 seconds because the model can copy prompt/reference audio at the beginning. A proper fix should prevent prompt leakage inside the model rather than relying on export-time cropping.
- External videos remain harder than LRS3 because of domain shift, crop quality, missing word timestamps, and imperfect decoded text.
- Latent correlation is a useful paired-condition metric, but it is not a complete subjective audio-quality score.

9 Conclusion

The final result is best understood as StreamLip audio-centric visual speech reconstruction. The model does not rely on a perfect intermediate transcript. It uses lip evidence, weak LM text, speaker

information, and audio-prompt timbre conditions to reconstruct Mimi latents directly. The strongest empirical evidence is that replacing ground-truth text with decoded text only slightly changes validation correlation, while timbre/audio-prompt conditioning and residual endpoint reconstruction produce the major gains. StreamLip V5 further shows that the text branch can be trained inside the project, and that remaining audio quality is governed more by visual/audio latent representation and alignment than by raw transcript WER. This supports the chosen architecture over a strict vision-to-text-to-audio cascade and identifies future improvement targets: better lip/audio representations, stronger timbre control, V5 alignment heads, and higher-fidelity latent decoding.

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifi, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Lewis Tunstall, Agustín Piqueres, et al. SmolLM2: A family of small language models. *arXiv preprint arXiv:2502.02737*, 2025.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.