XL-LoRA: Cross-Lingual Low-Rank Adaptation for Audio-Visual Speech Recognition

Lusi A[1,2,3]* **Zhiyong Duan**[4] **Feilong Bao**[1,2,3]†
[1]College of Computer Science, Inner Mongolia University, China
[2]National & Local Joint Engineering Research Center of
Intelligent Information Processing Technology for Mongolian, China
[3]Inner Mongolia Key Laboratory of Multilingual
Artificial Intelligence Technology, China
[4]College of Arts and Technology, Inner Mongolia Arts University, China

Abstract

Audio-visual speech recognition (AVSR) improves robustness to acoustic corruption by exploiting complementary visual speech cues. Extending AVSR to multiple languages, however, requires balancing transferable audio-visual patterns against language-specific phonological and orthographic variation. We present XL-LoRA, a unified parameter-efficient framework that organizes decoder adaptation into a Shared LoRA and a Language-specific LoRA (L-S LoRA) Bank within a frozen language-model backbone. The Shared LoRA is optimized across all languages, while an utterance-level language identifier selects one adapter from the L-S LoRA Bank for residual specialization, avoiding separate language-specific encoders or decoders. Across eight non-English MuAViC languages, XL-LoRA achieves average word error rates of 20.3% on clean speech and 41.3% under 0-dB babble noise with 217M trainable adaptation parameters and 88.6M active adaptation parameters per utterance, while keeping the encoders and language-model backbone frozen. Component ablations and adapter swapping support complementary roles for the Shared LoRA and L-S LoRA Bank; we further characterize their learned updates through parameter-space analysis. These results establish XL-LoRA as an effective shared and language-specific adaptation strategy for multilingual and noise-robust AVSR.

1 Introduction

Audio-visual speech recognition (AVSR) combines acoustic evidence with visible articulation and is therefore markedly more robust than audio-only recognition when speech is corrupted by noise (Tamura et al., 2015; Petridis et al., 2018; Makino et al., 2019). Its progress has been driven by large-scale resources such as LRS3 (Afouras et al., 2018), the transition from recurrent sequence models (Elman, 1990) to Transformers (Vaswani et al., 2017), and advances in self-supervised audio-visual learning (Shi et al., 2022) and corruption-aware training (Hong et al., 2023a). Recent systems further couple speech encoders with large language models (LLMs) to exploit stronger linguistic priors (Cappellazzo et al., 2025a; Yeo et al., 2025b); efficient multimodal token compression in general-purpose models also reinforces this direction (Zhang et al., 2025).

Multilingual speech recognition has advanced along several complementary lines. Whisper demonstrates broad cross-lingual transfer through large-scale weak supervision (Radford et al., 2023). For multilingual audio-visual learning, XLAVS-R extends noise-robust speech recognition and translation to more than 100 languages (Han et al., 2024); AudioVSR studies cross-language-family audio-to-lip transfer (Yang et al., 2024); Zero-AVSR introduces language-agnostic romanized speech representations for zero-shot transfer (Yeo et al., 2025a); and mWhisper-Flamingo jointly trains multilingual AVSR through decoder-side multimodal fusion (Rouditchenko et al., 2025). In parallel, Unified Speech Recognition trains one model for ASR, visual speech recognition (VSR), and AVSR (Haliassos et al., 2024), while recent multimodal LLM systems improve token efficiency, input granularity, and projector scaling (Yeo et al., 2025b; Cappellazzo et al., 2025b,c). These ad-

*First-author contact: 22409015@mail.imu.edu.cn.
†Corresponding author.

vances motivate our study of multilingual adaptation that supports both cross-language sharing and language-specific specialization.

Existing multilingual approaches generally favor either extensive parameter sharing or separate language-dependent modeling. Full sharing promotes transfer but can place conflicting phonological, lexical, and orthographic patterns in the same adaptation space, whereas independent models or adapters preserve specialization at the cost of duplicated capacity and weaker knowledge sharing. Standard low-rank adaptation (LoRA) reduces the trainable parameter count (Hu et al., 2022), but a single shared LoRA has no dedicated language-conditioned capacity, while fully independent LoRAs repeatedly learn patterns shared across languages. A unified multilingual AVSR model therefore benefits from structuring its low-rank capacity around both common and language-dependent components.

Our design builds directly on Omni-AVSR (Cappellazzo et al., 2026), which combines shared and task-specific LoRAs to unify ASR, VSR, and AVSR. Inspired by this task-oriented design, we organize sharing and specialization across *languages within AVSR*, where joint optimization must accommodate different output vocabularies and writing systems. One Shared LoRA receives multilingual gradients, while utterance-level language routing activates one adapter from an L-S LoRA Bank. We further introduce an orthogonality penalty to discourage overlap between their update directions. Building on the shared-plus-specific LoRA pattern, we investigate this language-conditioned formulation through eight-language experiments, budget-matched controls, and adapter swapping.

Our contributions are threefold. First, we propose a unified architecture comprising a Shared LoRA and an L-S LoRA Bank for multilingual AVSR, activating only the Shared LoRA and one language-specific LoRA per utterance. Second, we demonstrate competitive recognition across eight typologically diverse MuAViC languages under clean and severe babble-noise conditions using a 217M trainable adaptation budget, of which 88.6M parameters are active per utterance. Third, we evaluate the design through component ablations, total- and active-budget-matched Shared LoRA baselines, and adapter swapping, complemented by parameter-space analysis of the learned updates.

2 Proposed Method

2.1 Cross-Lingual AVSR Pipeline

As illustrated in Figure 1(a), we build a single multilingual AVSR model from a pre-trained Whisper audio encoder, an AV-HuBERT visual encoder, an audiovisual projector, and a Llama 3.2-3B decoder (Radford et al., 2023; Shi et al., 2022; Grattafiori et al., 2024). We keep the two encoders and the LLM frozen and train the projector and the proposed low-rank adapters. Given an audio waveform a and its synchronized lip-region video v, the encoders produce modality-specific sequences $Z_a = E_a(a)$ and $Z_v = E_v(v)$. The 1024-dimensional audio features are further reduced to 512 dimensions, while visual features retain 1024 dimensions. Temporal downsampling aligns the streams to a common length T_{av}, yielding $\widetilde{Z}_a \in \mathbb{R}^{T_{av}\times 512}$ and $\widetilde{Z}_v \in \mathbb{R}^{T_{av}\times 1024}$. The projector comprises audio, visual, and fusion branches. Each is a two-layer MLP $P_i(X) = \phi(XW_{i,1} + b_{i,1})W_{i,2} + b_{i,2}$, where ϕ is ReLU (Glorot et al., 2011). The fused representation is

$$\begin{aligned} H_a &= P_a(\widetilde{Z}_a), \quad H_v = P_v(\widetilde{Z}_v), \\ Z_{av} &= P_f([H_a \| H_v]). \end{aligned} \tag{1}$$

Here, $[\cdot\|\cdot]$ concatenates features, and the LLM embedding dimension is $d_h = 3072$. For $i \in \{a, v, f\}$, $W_{i,1} \in \mathbb{R}^{d_i \times d_h}$, $W_{i,2} \in \mathbb{R}^{d_h \times d_h}$, and both biases lie in $\mathbb{R}^{d_h}$, with $(d_a, d_v, d_f) = (512, 1024, 6144)$. These three branches contain approximately 51.9M trainable parameters, including biases. The output $Z_{av} \in \mathbb{R}^{T_{av}\times d_h}$ supplies the audiovisual tokens. For language ℓ, an utterance-level identifier $\langle \text{lang}_\ell \rangle$ is embedded and prepended to the audiovisual sequence, giving $Z_{\text{in}} = [\text{Emb}(\langle \text{lang}_\ell \rangle); Z_{av}]$. Here, $[\cdot;\cdot]$ denotes sequence concatenation. The identifier instructs multilingual decoding and selects the corresponding language-specific adapter. The unified decoder generates a transcription $Y = (y_1, \ldots, y_T)$ autoregressively as

$$p(Y \mid Z_{\text{in}}) = \prod_{t=1}^{T} p(y_t \mid Z_{\text{in}}, y_{<t}). \tag{2}$$

In the reported MuAViC (Anwar et al., 2023) experiments, the language identifier is obtained from the dataset metadata.

2.2 XL-LoRA Decoder Adaptation

As illustrated in Figure 1(b), XL-LoRA decomposes decoder adaptation into one Shared LoRA

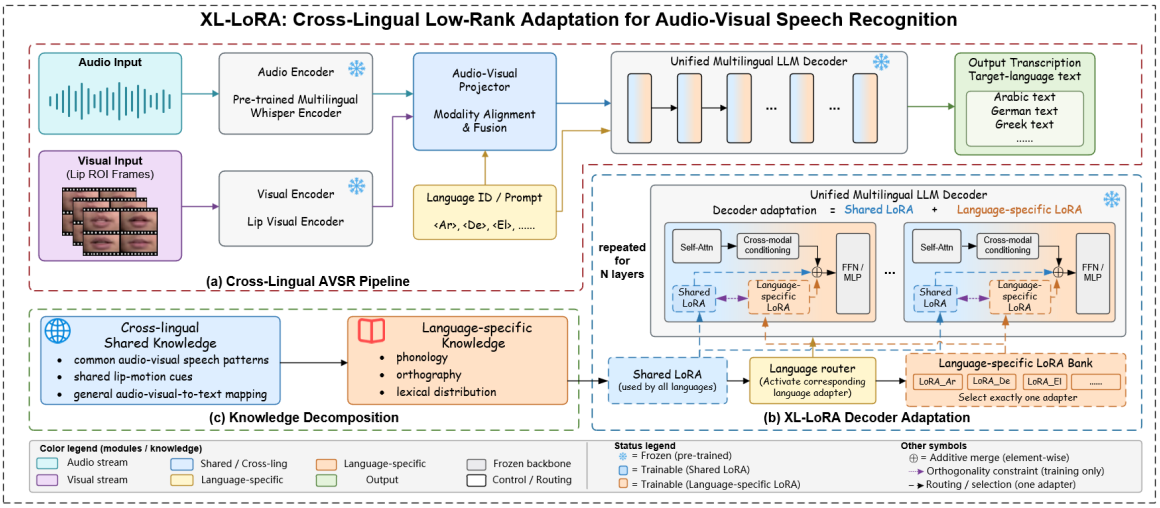


Figure 1: An overview of the proposed XL-LoRA.

and an L-S LoRA Bank. The shared adapter is applied to the query and value projections of every decoder layer and is active for all languages. For each target language $\ell \in \mathcal{L}$, the bank maintains an independent adapter that is activated only for that language. The Shared LoRA is optimized for multilingual audiovisual-to-text mappings, while the selected L-S LoRA provides language-conditioned residual adaptation.

Following standard LoRA (Hu et al., 2022), the two low-rank updates are

$$\Delta W_s = \alpha B_s A_s, \qquad \Delta W_\ell = \alpha B_\ell A_\ell, \quad (3)$$

where $A_s, A_\ell \in \mathbb{R}^{r \times d_{\text{in}}}$, $B_s, B_\ell \in \mathbb{R}^{d_{\text{out}} \times r}$, and α is the effective LoRA scale. Each update corresponds to $m = (j, p)$, with decoder layer j and projection $p \in \{Q, V\}$; this index is omitted below when unambiguous. For a frozen matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the effective matrix is

$$W_{\text{eff}}^{(\ell)} = W_0 + \Delta W_s + \Delta W_\ell. \quad (4)$$

For a column-vector hidden state $h \in \mathbb{R}^{d_{\text{in}}}$, the transformation is

$$h' = W_0 h + \Delta W_s h + \Delta W_\ell h. \quad (5)$$

Thus, every utterance activates the Shared LoRA and exactly one member of the L-S LoRA Bank, without replicating the encoders, projector, or LLM backbone.

2.3 Knowledge Decomposition and Training

The additive parameterization and language-conditioned routing create two complementary optimization pathways. The Shared LoRA receives gradients from the complete multilingual distribution and learns mappings useful for jointly trained languages. In contrast, ΔW_ℓ receives transcription-loss gradients only from language ℓ and specializes the shared mapping to its language-dependent output patterns. These pathways distinguish shared from language-conditioned adaptation, while the objective below further regularizes alignment between their updates.

For language ℓ, the autoregressive transcription loss is

$$\mathcal{L}_\ell = -\sum_{t=1}^{T} \log p(y_t \mid Z_{\text{in}}, y_{<t}), \quad (6)$$

and multilingual training minimizes

$$\mathcal{L}_{\text{ASR}} = \sum_{\ell \in \mathcal{L}} \lambda_\ell \mathcal{L}_\ell, \quad (7)$$

where λ_ℓ controls the contribution of each language. During both training and inference, the language identifier routes each utterance through the Shared LoRA and its matched adapter in the L-S LoRA Bank; all unmatched adapters remain inactive in the forward computation. We use the Romance/non-Romance groups only to report performance, not to guide training or routing.

To discourage overlap in update directions, we normalize $\overline{\Delta W} = \Delta W / (\|\Delta W\|_F + \epsilon)$, where $\epsilon > 0$ prevents division by zero, and minimize

$$\mathcal{L}_{\text{orth}} = \frac{1}{|\mathcal{L}||\mathcal{M}|} \sum_{\ell \in \mathcal{L}} \sum_{m \in \mathcal{M}} \langle \overline{\Delta W}_{s,m}, \overline{\Delta W}_{\ell,m} \rangle_F^2, \quad (8)$$

$$\mathcal{L} = \mathcal{L}_{\text{ASR}} + \beta \mathcal{L}_{\text{orth}}, \quad (9)$$

where $\mathcal{M} = \{(j, p) : j = 1, \ldots, N,\ p \in \{Q, V\}\}$ indexes the adapted projections in all N decoder layers and $\beta = 10^{-3}$. This orthogonality constraint explicitly discourages parameter-space overlap between the Shared LoRA and L-S LoRA Bank while preserving the original language-conditioned routing structure.

3 Experiments

3.1 Experimental Setup

We use MuAViC, a 1141-hour, nine-language corpus based on LRS3 and mTEDx (Anwar et al., 2023; Afouras et al., 2018; Salesky et al., 2021). We adopt eight target languages: Arabic (Ar, 16 h), German (De, 10 h), Greek (El, 25 h), Spanish (Es, 178 h), French (Fr, 176 h), Italian (It, 101 h), Portuguese (Pt, 153 h), and Russian (Ru, 49 h). All languages follow official MuAViC splits and identical preprocessing, transcript normalization, noise augmentation, and decoding. For analysis, *Romance-4* (Rom) comprises Es, Fr, It, and Pt; *Non-Romance-4* (NonRom) comprises Ar, De, El, and Ru. NonRom spans Semitic, Germanic, Hellenic, and Slavic languages and Arabic, Latin, Greek, and Cyrillic scripts. Training includes all eight languages; the groups assess multilingual coverage rather than unseen-language transfer. WER is macro-averaged over languages. The equally sized groups give $\mathrm{Avg8} = (\mathrm{Rom} + \mathrm{NonRom})/2$, the unweighted eight-language mean. Audio is resampled to 16 kHz and video to 25 fps; 96×96 mouth regions are extracted using RetinaFace (Deng et al., 2020). Babble noise from MUSAN is mixed with probability 0.75, and noisy test results use 0-dB SNR (Snyder et al., 2015). The model uses frozen Whisper-medium and AV-HuBERT Large encoders, a trainable audiovisual projector, and a frozen Llama 3.2-3B decoder (Radford et al., 2023; Shi et al., 2022; Grattafiori et al., 2024). Decoder query and value projections use rank-64 LoRA with scaling factor 16. We train for 10 epochs using AdamW, learning rate 10^{-3}, weight decay 0.1, and cosine annealing; decoding uses beam width 5 (Loshchilov and Hutter, 2019).

3.2 Comparison with Multilingual AVSR Systems

Table 1 compares XL-LoRA with multilingual AVSR systems on eight MuAViC languages. XL-LoRA obtains 20.3% clean Avg8 WER, comparable to mWhisper-Flamingo's 20.4%. We interpret the published noisy results in light of their different evaluation protocols: XL-LoRA uses MUSAN babble, mWhisper-Flamingo uses babble from 30 LRS3 speakers, and XLAVS-R uses MuAViC-derived babble (Rouditchenko et al., 2025; Han et al., 2024). We omit Fast Conformer's noisy results because they use -5 dB. Related LLM-based systems also differ in supervision: Omni-AVSR and MMS-LLaMA evaluate task unification and token compression on LRS2/LRS3 and LRS3, respectively (Cappellazzo et al., 2026; Yeo et al., 2025b); Zero-AVSR uses romanized supervision and a broader multilingual corpus (Yeo et al., 2025a). We therefore compare adaptation designs under a common backbone and protocol in Table 2.

3.3 Ablation on Knowledge Decomposition

We examine the contributions of individual components and the adaptation budget in Table 2. Removing the orthogonality constraint, the Shared LoRA, or the L-S LoRA Bank consistently degrades both clean and noisy Avg8 WER. The orthogonality ablation shows that adding the update-alignment penalty improves recognition under the evaluated conditions. The larger degradation without the L-S LoRA Bank highlights the role of language-conditioned residual adaptation in limiting cross-lingual interference. The Shared LoRA baselines yield clean/noisy Avg8 WERs of 21.8%/43.9% and 22.9%/44.5% under total-budget and active-budget matching, respectively, compared with XL-LoRA's 20.3%/41.3%. Without the orthogonality constraint, the shared–specific model achieves 20.7%/41.7%, also below the active-matched baseline. These comparisons support the effectiveness of combining shared and language-specific adaptation under both budget-matching criteria.

3.4 Routing and Parameter-Space Analysis

For a closer examination of the adapters, we focus on Es, Fr, It, and Pt, each with over 100 h of training data. These analyses complement the eight-language results in Tables 1–2.

We study multilingual AVSR with known utterance-level language identities from MuAViC metadata and use hard routing to select the matched adapter. For language ℓ, the indicator $r_k(\ell) = \mathbb{I}[k = \ell]$ selects the L-S LoRA:

$$\Delta W^{(\ell)}_{\mathrm{route},m} = \Delta W_{s,m} + \sum_{k \in \mathcal{L}} r_k(\ell) \Delta W_{k,m}. \quad (10)$$

Here, $\mathbb{I}[\cdot]$ denotes the indicator function, which

Method	Reported Params (M)	WER ↓ Ar	De	El	Es	Fr	It	Pt	Ru	Avg8
Clean										
AV-HuBERT (Shi et al., 2022)	477	89.4	52.0	46.2	17.4	20.3	20.8	22.1	44.7	39.1
Intuitive Multilingual (Hong et al., 2023b)	477	90.9	48.1	42.4	18.0	21.5	19.7	20.8	54.1	39.4
XLAVS-R 2B (Han et al., 2024)	2,150	79.3	44.4	19.0	9.1	12.3	10.6	11.2	25.0	26.4
Fast Conformer (Burchi et al., 2024)	197	–	46.8	–	9.0	11.4	11.6	11.8	–	–
mWhisper-Flamingo (Rouditchenko et al., 2025)	1,390	**70.2**	22.0	14.0	8.1	**11.3**	**10.2**	10.8	16.7	20.4
XL-LoRA	217†	70.5	**21.6**	**13.8**	**8.0**	11.4	10.3	**10.6**	**16.5**	**20.3**
Noisy (0-dB babble)										
AV-HuBERT	477	104.7	74.2	70.4	43.1	43.0	48.2	47.5	67.4	62.3
Intuitive Multilingual	477	99.7	65.9	61.9	38.9	37.4	42.5	42.1	69.0	57.2
XLAVS-R 2B	2,150	**93.5**	58.5	38.6	**23.9**	**23.5**	24.6	26.1	41.0	**41.2**
mWhisper-Flamingo	1,390	95.3	**49.4**	41.8	28.0	27.5	35.2	36.0	**36.1**	43.7
XL-LoRA	217†	95.1	54.6	**38.4**	26.3	**23.5**	**24.2**	**25.6**	42.5	41.3

Table 1: WER (%) on eight MuAViC languages; Avg8 is their unweighted mean. External results and model sizes are source-reported; † marks XL-LoRA's trainable adaptation parameters, excluding frozen backbones. Noisy results use source-specific 0-dB babble protocols. Fast Conformer reports five clean-language results.

Setting	Params T/A (M)	Rom-C	NonRom-C	Avg8-C	Rom-N	NonRom-N	Avg8-N
XL-LoRA	217/88.6	**10.1**	**30.6**	**20.3**	**24.9**	**57.7**	**41.3**
w/o Orth.	217/88.6	10.4	30.9	20.7	25.3	58.0	41.7
w/o Shared LoRA	198.7/70.3	10.8	31.5	21.2	26.7	59.6	43.2
w/o L-S LoRA Bank	70.3/70.3	13.3	33.8	23.6	29.1	61.9	45.5
w/o Both	51.9/51.9	18.9	39.6	29.3	35.9	68.7	52.3
Shared LoRA (total-matched)	217/217	11.5	32.1	21.8	27.4	60.3	43.9
Shared LoRA (active-matched)	≈ 88.6/88.6	12.6	33.2	22.9	28.2	60.8	44.5

Table 2: Ablation on knowledge decomposition under clean (C) and 0-dB babble-noise (N) conditions. Params T/A denotes total trainable adaptation parameters and active adaptation parameters per utterance. Shared LoRA baselines remove the L-S LoRA Bank and adjust the shared rank to match XL-LoRA's total or per-utterance active adaptation budget, respectively, including the projector.

equals 1 when its condition is satisfied and 0 otherwise; therefore, exactly one language-specific adapter is activated for each utterance. To examine the role of language-conditioned routing, we replace the matched L-S LoRA with another language's adapter while keeping the audiovisual input and language identifier unchanged. The diagonal cells in Figure 2 represent correct routing, whereas the off-diagonal cells represent controlled routing errors.

Correct routing yields the lowest WER for each analyzed language in both conditions, while mismatched adapters degrade recognition, demonstrating their functional specialization. Removing the L-S LoRA Bank also increases eight-language Avg8 WER from 20.3%/41.3% to 23.6%/45.5% under clean/noisy conditions. Together, these results support the role of language-conditioned specialization in multilingual recognition.

For nonzero LoRA update matrices $\mathbf{W}_A$ and $\mathbf{W}_B$, we compute cosine similarity and normalized

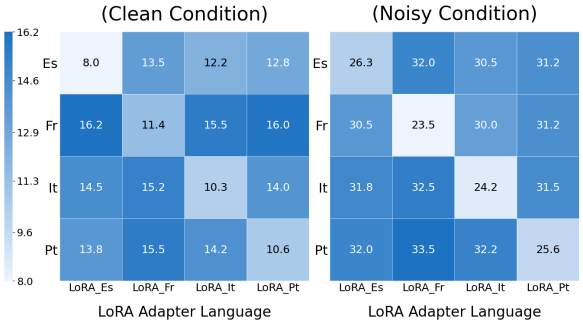


Figure 2: Cross-lingual adapter swapping with the final XL-LoRA model under clean and 0-dB babble-noise conditions. Rows denote target languages and columns denote activated adapters. Color ranges are 8.0–16.2 for clean speech and 23.5–33.5 for noisy speech.

Frobenius distance:

$$\begin{aligned} \mathrm{sim}(\mathbf{W}_A, \mathbf{W}_B) &= \frac{\langle \mathbf{W}_A, \mathbf{W}_B \rangle_F}{\|\mathbf{W}_A\|_F \|\mathbf{W}_B\|_F}, \\ \widetilde{D}_F(\mathbf{W}_A, \mathbf{W}_B) &= \frac{\|\mathbf{W}_A - \mathbf{W}_B\|_F}{\|\mathbf{W}_A\|_F + \|\mathbf{W}_B\|_F}. \end{aligned} \tag{11}$$

Here, $\langle \cdot, \cdot \rangle_F$ and $\|\cdot\|_F$ denote the Frobenius inner product and norm; flattening the updates gives the equivalent vector cosine. Both metrics are computed separately for corresponding Q/V updates

in each decoder layer and then averaged with the same weights across layers and projections. Cosine values retain their signs before averaging.

Figure 3 compares the shared and language-specific updates in parameter space. Shared–language pairs yield mean signed cosine similarities of 0.06–0.09, compared with 0.15–0.20 among language-specific adapters. Their mean normalized Frobenius distances are also slightly larger, ranging from 0.47 to 0.52 versus 0.42 to 0.46. Together, these measurements reveal different aggregate update patterns for the shared and language-specific components.

With cosine signs retained, we use the reported means to summarize directional relationships across projections, allowing individual updates to exhibit positive, negative, or near-orthogonal alignment. This parameter-space perspective complements the ablation and adapter-swapping results, which demonstrate the benefits of combining shared adaptation with language-conditioned specialization.

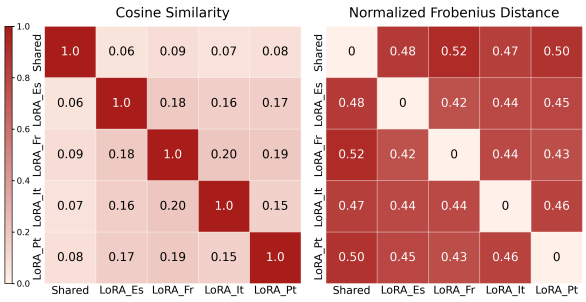


Figure 3: Layer-averaged signed cosine similarity and normalized Frobenius distance for the Shared LoRA and four language adapters. Color ranges are 0–1 (cosine) and 0–0.56 (distance).

3.5 Parameter Efficiency and Practical Implications

XL-LoRA has 217M trainable adaptation parameters, including the projector and all LoRAs. Each utterance uses the projector, Shared LoRA, and one L-S LoRA: 88.6M active adaptation parameters. We report both counts separately from the frozen backbones, which remain part of the complete inference model. At fixed rank, each added language requires one additional stored adapter while keeping the per-utterance active adaptation count unchanged. We assess efficiency here through adapter storage and adaptation budgets. The two Shared LoRA controls in Table 2 separately match the total and active adaptation budgets. XL-LoRA achieves lower WER in both comparisons, supporting the allocation of adaptation parameters between shared and language-specific components.

4 Conclusion

In this work, we present XL-LoRA, a unified framework for multilingual audio-visual speech recognition that combines a Shared LoRA and an L-S LoRA Bank within a frozen LLM backbone. Joint multilingual optimization and language-conditioned routing support shared adaptation alongside residual specialization. On eight MuAViC languages, XL-LoRA achieves average WERs of 20.3% in clean conditions and 41.3% under 0-dB MUSAN babble noise, with 217M trainable adaptation parameters and 88.6M active adaptation parameters per utterance. Component ablations and the total-budget-matched control support the value of shared–specific adaptation beyond aggregate parameter count. Adapter swapping demonstrates functional specialization in the four analyzed languages, and parameter-space analysis summarizes their aggregate update relationships. These findings establish XL-LoRA as an effective adaptation strategy for jointly trained multilingual AVSR.

Limitations

This study focuses on joint adaptation across eight MuAViC languages, with detailed adapter analyses on four resource-rich Romance languages. Extending these analyses to low-resource Ar and De, and evaluating held-out-language and few-shot adaptation, are natural next steps. Routing currently uses metadata language IDs; automatic language identification and intra-utterance code-switching offer further directions.

The evaluation uses one backbone combination and fixed LoRA ranks. Exploring other configurations would clarify the design's broader applicability. Multi-seed evaluation and confidence intervals would help assess the stability of small WER differences, which we do not interpret as statistically significant. Targeted linguistic probes could complement the parameter-space analysis by examining the content associated with the shared and language-specific updates. Since external systems differ in training data, architectures, normalization, and noise protocols, shared-backbone comparisons and full-model memory, FLOPs, and latency measurements would strengthen practical characteriza-

tion. Additional datasets, noise types, and SNRs would extend the present MuAViC and 0-dB MUSAN babble evaluation.

References

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

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. MuAViC: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. In *Interspeech 2023*, pages 4064–4068.

Maxime Burchi, Krishna C Puvvada, Jagadeesh Balam, Boris Ginsburg, and Radu Timofte. 2024. Multilingual audio-visual speech recognition with hybrid CTC/RNN-T fast conformer. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10211–10215. IEEE.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. 2025a. Large language models are strong audio-visual speech recognition learners. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Umberto Cappellazzo, Minsu Kim, and Stavros Petridis. 2025b. Adaptive audio-visual speech recognition via matryoshka-based multimodal LLMs. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Umberto Cappellazzo, Minsu Kim, Stavros Petridis, Daniele Falavigna, and Alessio Brutti. 2025c. Scaling and enhancing LLM-based AVSR: A sparse mixture of projectors approach. In *Interspeech 2025*, pages 1823–1827.

Umberto Cappellazzo, Xubo Liu, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2026. Omni-AVSR: Towards unified multimodal speech recognition with large language models. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17772–17776. IEEE.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proc. CVPR*, pages 5202–5211. IEEE/CVF.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323. JMLR Workshop and Conference Proceedings.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alexandros Haliassos, Rodrigo Mira, Honglie Chen, Zoe Landgraf, Stavros Petridis, and Maja Pantic. 2024. Unified Speech Recognition: A single model for auditory, visual, and audiovisual inputs. In *Advances in Neural Information Processing Systems*, volume 37, pages 139673–139699. Neural Information Processing Systems Foundation.

HyoJung Han, Mohamed Anwar, Juan Pino, Wei-Ning Hsu, Marine Carpuat, Bowen Shi, and Changhan Wang. 2024. XLAVS-R: Cross-lingual audio-visual speech representation learning for noise-robust speech perception. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12896–12911. Association for Computational Linguistics.

Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 2023a. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *Proc. CVPR*, pages 18783–18794. IEEE/CVF.

Joanna Hong, Se Park, and Yong Ro. 2023b. Intuitive multilingual audio-visual speech recognition with a single-trained model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4886–4890. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. 2019. Recurrent neural network transducer for audio-visual speech recognition. In *Proc. ASRU*, pages 905–912. IEEE.

Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *Proc. ICASSP*, pages 6548–6552. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Andrew Rouditchenko, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. mWhisper-Flamingo for multilingual audio-visual noise-robust speech recognition. *IEEE Signal Processing Letters*, 32:2144–2148.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Matt Post. 2021. The multilingual TEDx corpus for speech recognition and translation. In *Interspeech 2021*, pages 3655–3659.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Satoshi Tamura, Hiroshi Ninomiya, Norihide Kitaoka, Shin Osuga, Yurie Iribe, Kazuya Takeda, and Satoru Hayamizu. 2015. Audio-visual speech recognition using deep bottleneck features and high-performance lipreading. In *Proc. APSIPA*, pages 575–582. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Xiaoda Yang, Xize Cheng, Jiaqi Duan, Hongshun Qiu, Minjie Hong, Minghui Fang, Shengpeng Ji, Jialong Zuo, Zhiqing Hong, Zhimeng Zhang, and Tao Jin. 2024. AudioVSR: Enhancing video speech recognition with audio data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15352–15361. Association for Computational Linguistics.

Jeong Hun Yeo, Minsu Kim, Chae Won Kim, Stavros Petridis, and Yong Man Ro. 2025a. Zero-AVSR: Zero-shot audio-visual speech recognition with LLMs by learning language-agnostic speech representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6693–6703.

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. 2025b. MMS-LLaMA: Efficient LLM-based audio-visual speech recognition with minimal multimodal speech tokens. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20724–20735. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. LLaVA-Mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*.