Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparing Multimodal Fusion Strategies for Visual Speech Recognition on LRW-AR

Merieme Yakine

meriemeyakine@gmail.com

University Mohammed First

Ilham Slimani

University Mohammed First

Ilhame El Farissi

University Mohammed First

Research Article

Keywords: Arabic visual speech recognition, lip-reading, LRW-AR dataset, multimodal fusion, facial landmark features

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9610691/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Comparing Multimodal Fusion Strategies for Visual Speech Recognition on LRW-AR

Merieme Yakine[1], Ilham Slimani[2], and Ilhame El Farissi[3]

[1]LIPIO, Faculty of Sciences, University Mohammed First, Oujda, Morocco,
meriemeyakine@gmail.com
[2]LIPIO, Faculty of Sciences, University Mohammed First, Oujda, Morocco,
i.slimani@ump.ac.ma
[3]LIPIO, Faculty of Sciences, University Mohammed First, Oujda, Morocco,
i.elfarissi@ump.ma

Abstract

Visual speech recognition (VSR) and lip-reading systems increasingly rely on fusing multiple modalities—typically appearance-based visual features and geometry-based facial landmark representations—to improve word-level classification. The choice of fusion mechanism, however, is often taken from a single paradigm without systematic comparison. This paper presents a comparative study of five multimodal fusion strategies for Arabic lip-reading on the LRW-AR benchmark: the original bidirectional cross-attention baseline (AttentionFusionOriginal), Feature-wise Linear Modulation (FiLM), gated fusion (GatedFusion), bilateral co-attention (BiCoAttention), and transformer-based fusion (TransformerFusion). All methods operate on the same pre-extracted visual and geometric feature tensors and feed a common linear classifier. Using an 80/20 train/validation split, 200 epochs, and consistent hyperparameters, we report validation top-1 accuracy and macro F1. BiCoAttention achieves the best validation accuracy (28.73%) and F1 (0.278), followed by GatedFusion and the original attention fusion; FiLM performs moderately; TransformerFusion collapses to near chance, suggesting sensitivity to architecture or training. We discuss implications for model design and outline directions for future work.

Keywords: Arabic visual speech recognition, lip-reading, LRW-AR dataset, multimodal fusion, facial landmark features.

1 Introduction

1.1 Research Context

Visual speech recognition (VSR) and automatic lip-reading aim to recognize spoken words or phrases from video of the speaker's face, without access to the acoustic signal [1, 2]. Applications range from assistive technology and silent dictation to robust speech recognition in noisy environments [3, 4]. Recent benchmarks such as LRW [5] and its Arabic variant LRW-AR [6] provide large-scale, in-the-wild datasets for word-level recognition, fostering deep models that combine *visual* features (e.g., from CNNs or video backbones) with *geometric* or landmark-based representations of lip and face dynamics.

Arabic lip-reading poses specific challenges: the script and phonology differ from English, and publicly available large-scale datasets are scarcer [6]. LRW-AR fills this gap by providing word-level annotations aligned with preprocessed video

and facial landmarks. In such pipelines, a critical design choice is *how* to fuse the two streams. Simple concatenation or late fusion often underperforms; instead, attention-based and conditioning mechanisms let one modality modulate or attend to the other [7, 8]. The LRW-AR reference implementation uses a bidirectional cross-attention module. It remains unclear whether this choice is optimal relative to other fusion paradigms—FiLM [8], gated fusion [9], co-attention [10], or transformer encoders [7]—under the same data and training protocol.

1.2 Problem Statement

We consider the setting where pre-extracted feature tensors are available: for each sample, a **visual** feature sequence $\mathbf{V} \in \mathbb{R}^{T \times d_v}$ and a **geometric** (landmark) feature sequence $\mathbf{G} \in \mathbb{R}^{T \times d_g}$, with T the temporal length and d_v, d_g the feature dimensions. The goal is to produce a fixed-size representation that is passed to a linear classifier for word-level classification. The central question is: *Which fusion strategy, among several representative approaches, yields the best accuracy and F1 under a fixed experimental protocol?*

1.3 Research Gap

Existing work on LRW-AR and similar VSR benchmarks typically adopts one fusion design without ablating alternatives. A systematic comparison of multiple fusion families (attention, conditioning, gating, co-attention, transformer) on the same feature extraction pipeline and evaluation protocol is missing. Such a comparison is necessary to guide architecture selection and to understand which inductive biases (e.g., symmetric cross-attention vs. additive conditioning) best suit visual–geometric fusion in lip-reading.

1.4 Contributions

- We compare five fusion strategies (AttentionFusionOriginal, FiLM, GatedFusion, BiCoAttention, TransformerFusion) on pre-extracted LRW-AR visual and geometric features, with an identical classifier and training setup.
- We report validation top-1 accuracy and macro F1 for each method (200 epochs, 80/20 split).
- We show that BiCoAttention yields the best validation performance; the original attention fusion and GatedFusion are competitive; FiLM is weaker; TransformerFusion fails to learn under the current settings, suggesting a need for different depth, normalization, or learning rate.
- We provide a concise mathematical description of each fusion module and discuss practical implications for VSR system design.

2 Related Work

2.1 Visual Speech Recognition and LRW-AR

Lip-reading and VSR have been advanced by large-scale datasets such as LRW [5], LRS2/LRS3 [11], and end-to-end models such as LipNet [12]; surveys [1, 13, 14, 2] review the field. Word-level recognition typically combines convolutional or recurrent encoders with temporal modeling [15]. LRW-AR [6] extends this line to Arabic with word-level annotations and preprocessed video and landmark data. The pipeline extracts a *visual* feature sequence (e.g., from a CNN over cropped lip or face frames) and a *geometric* feature sequence (e.g., from a GNN or MLP over facial landmark coordinates [16]). In this study we use such pre-extracted features and focus on the fusion stage.

2.2 Early vs. Late Fusion and Modality Interaction

Multimodal fusion is commonly categorized into *early*, *late*, and *intermediate* fusion [17, 18]. For VSR, intermediate fusion that allows temporal and cross-modal interaction often outperforms naive concatenation. Our study focuses on intermediate fusion modules that take fixed-dimensional per-timestep features and produce a single fused sequence for classification [19].

2.3 Multimodal Fusion in Deep Learning

Cross-attention [7] allows one modality to attend to the other; the LRW-AR baseline uses bidirectional cross-attention and concatenates the outputs. **FiLM** [8] conditions a feature map via learned scale and shift (gamma/beta) from a conditioning vector; we adapt it so that geometry modulates the visual stream per timestep. **Gated fusion** [9] uses a sigmoid gate to weight one stream before combining with the other. **Co-attention** [10, 20] computes attention in both directions; our BiCoAttention follows this idea with separate multi-head attention blocks for $V \to G$ and $G \to V$ in a common embedding space. **Transformer** fusion applies a transformer encoder to the combined modalities [7]; we use a 2-layer encoder with geometry added to the visual stream. We compare all five in a single framework to isolate the effect of the fusion module.

3 Methodology

3.1 Architecture Overview

The overall pipeline consists of: (1) pre-extracted visual and geometric sequences; (2) a fusion module that maps $(\mathbf{V}, \mathbf{G})$ to a sequence of 1024-dimensional vectors; (3) mean pooling over the time dimension to obtain one vector per sample; (4) a linear classifier mapping $1024 \to C$ classes. All fusion modules take inputs of shape (B, T, d_v) and (B, T, d_g) and output $(B, T, 1024)$. The dimension 1024 is fixed to match the LRW-AR classifier head and to ensure a fair comparison. Variable-length sequences are padded to the same length per batch; the fused sequence is then aggregated by mean pooling over time.

3.2 Data and Feature Pipeline

Each sample is represented by a pre-extracted pair of sequences:

- $\mathbf{V} \in \mathbb{R}^{1\times T\times d_v}$: visual features (e.g., from a CNN backbone over cropped lip or face frames).
- $\mathbf{G} \in \mathbb{R}^{1\times T\times d_g}$: geometric features (e.g., from a GNN or MLP over landmark sequences).

These are produced offline by the LRW-AR pipeline [6]. Sequences are aligned to the same length T within each batch when needed. The number of classes C is given by the dataset.

3.3 Fusion Modules

Each fusion module takes $(\mathbf{V}, \mathbf{G})$ and outputs a tensor of shape $(B, T, 1024)$, where B is the batch size. The classifier then uses mean pooling over time and a linear layer $1024 \to C$. When dimensions differ across modalities, linear projections align them as needed. For attention-based modules, the embedding dimension is chosen divisible by the number of heads (e.g., 8 or 4).

3.3.1 AttentionFusionOriginal (Baseline)

The reference LRW-AR fusion uses two multi-head attention steps:

$$\tilde{\mathbf{V}} = \mathrm{MHA}(\mathbf{V}, \mathbf{G}, \mathbf{G}), \quad \tilde{\mathbf{G}} = \mathrm{MHA}(\mathbf{G}, \mathbf{V}, \mathbf{V}), \tag{1}$$

$$\mathbf{F} = \mathrm{Linear}_{2d_v\to 1024}([\tilde{\mathbf{V}}; \tilde{\mathbf{G}}]). \tag{2}$$

where $\mathbf{G}$ is projected to dimension d_v if $d_g \neq d_v$ via a linear layer. Here, $\mathrm{MHA}(Q, K, V)$ denotes multi-head attention with query Q, key K, value V. The first block lets the visual stream attend to the geometric stream (vision conditioned on landmarks); the second does the reverse. Concatenating both outputs preserves information from both modalities. This yields a symmetric cross-attention representation of length $2d_v$ per timestep, then projected to 1024. The design is relatively parameter-heavy due to two attention blocks and the final projection.

3.3.2 FiLMFusion

Feature-wise Linear Modulation [8] uses geometry to scale and shift the visual stream:

$$\mathbf{F} = \mathrm{Linear}_{d_v\to 1024}\big(\gamma(\mathbf{G}) \odot \mathbf{V} + \beta(\mathbf{G})\big). \tag{3}$$

where γ, β are linear maps from $\mathbf{G}$ to $\mathbb{R}^{d_v}$ and $\odot$ is element-wise multiplication. No explicit attention is used; the geometric stream only provides a per-timestep conditioning that modulates

the magnitude and offset of the visual features. This keeps the model lightweight but may be less expressive when fine-grained alignment between modalities is needed.

3.3.3 GatedFusion

A gate derived from $\mathbf{G}$ modulates $\mathbf{V}$; the result is concatenated with $\mathbf{G}$ and projected:

$$\mathbf{g} = \sigma(\text{Linear}_{d_g \to d_v}(\mathbf{G})), \tag{4}$$

$$\mathbf{F} = \text{Linear}_{d_v + d_g \to 1024}([\mathbf{V} \odot \mathbf{g}; \mathbf{G}]). \tag{5}$$

The gate $\mathbf{g} \in (0,1)^{d_v}$ selectively passes or suppresses elements of $\mathbf{V}$; the geometric stream is then concatenated so that the classifier can use both gated visual and raw geometric information. This is a simple and efficient design with one gate layer and one projection.

3.3.4 BiCoAttention

Two multi-head attention blocks compute vision-to-geometry and geometry-to-vision updates in a common dimension d (chosen so d is divisible by the number of heads h):

$$\mathbf{V}', \mathbf{G}' = \text{Proj}_v(\mathbf{V}), \text{Proj}_g(\mathbf{G}), \tag{6}$$

$$\mathbf{V}_{2G} = \text{MHA}(\mathbf{V}', \mathbf{G}', \mathbf{G}'), \tag{7}$$

$$\mathbf{G}_{2V} = \text{MHA}(\mathbf{G}', \mathbf{V}', \mathbf{V}'), \tag{8}$$

$$\mathbf{F} = \text{Linear}_{2d \to 1024}([\mathbf{V}_{2G}; \mathbf{G}_{2V}]). \tag{9}$$

Both modalities are first projected to the same dimension d; then vision attends to geometry (values from $\mathbf{G}'$) and geometry attends to vision (values from $\mathbf{V}'$). The two updated sequences are concatenated and projected to 1024. Unlike the baseline, BiCoAttention uses a unified embedding dimension and explicit projection of both streams, which can improve gradient flow and symmetry.

3.3.5 TransformerFusion

Geometry is projected to d_v and added to the visual stream; a transformer encoder (2 layers in our experiments) processes the joint sequence:

$$\mathbf{H} = \mathbf{V} + \text{Linear}_{d_g \to d_v}(\mathbf{G}), \tag{10}$$

$$\mathbf{F} = \text{Linear}_{d_v \to 1024}\big(\text{TransformerEncoder}(\mathbf{H})\big). \tag{11}$$

Each layer combines self-attention with a feed-forward block (with residual connections and layer normalization). The modality fusion is purely additive at the input; the transformer then models temporal and cross-position dependencies. No extra positional encoding is added beyond what may be present in the features. This design is the most parameter-heavy and may require careful initialization or learning rate scheduling.

3.4 Classifier and Pooling

The fusion output $\mathbf{F} \in \mathbb{R}^{B \times T \times 1024}$ is aggregated over time by mean pooling: $\bar{\mathbf{f}}_b = \frac{1}{T} \sum_{t=1}^{T} \mathbf{F}_{b,t,:}$, yielding $\bar{\mathbf{F}} \in \mathbb{R}^{B \times 1024}$. A single linear layer maps $\bar{\mathbf{F}}$ to logits in $\mathbb{R}^{B \times C}$. This keeps the classifier identical across all fusion methods and focuses the comparison on the fusion module itself.

3.5 Training and Evaluation Protocol

For each fusion type we: (1) split the data 80% train / 20% validation at random; (2) train the fusion module and a linear classifier ($1024 \to C$) jointly with cross-entropy loss and Adam (learning rate 10^{-3}) for 200 epochs, batch size 8; (3) at the end of each epoch, compute validation top-1 accuracy and macro F1; (4) record the final validation metrics. The same optimizer and hyperparameters are used for all methods. Mean pooling over the time dimension is applied to the fusion output before the classifier. No data augmentation is applied; reported numbers are on the validation set only.

4 Experiments and Implementation

4.1 Dataset

We use the LRW-AR feature set [6]: pre-extracted visual and geometric feature sequences per word sample. The training portion of the dataset is used for both training and validation via an 80/20 random split, so that the comparison focuses on the same data distribution. All

five fusion modules are evaluated on the same feature set; each method is trained and evaluated with an independent random 80/20 split. The number of classes C and the feature dimensions d_v, d_g are determined by the dataset.

4.2 Experimental Setup

- **Train/val split:** 80% / 20% at random, independently per fusion method.
- **Epochs:** 200 (fixed for all methods).
- **Batch size:** 8.
- **Optimizer:** Adam, learning rate 10^{-3}; one optimizer for the fusion module and the classifier jointly.
- **Loss:** Cross-entropy.
- **Sequences:** Padded to the same length per batch when necessary.

4.3 Evaluation Metrics

We report: (1) **validation top-1 accuracy** (% of validation samples correctly classified); (2) **validation macro F1** (macro-averaged F1 over the C classes).

5 Results and Discussion

5.1 Quantitative Results

Table 1 reports the validation performance for each fusion strategy. All methods were evaluated under the same protocol (200 epochs, 80/20 split, same feature set). Results are from the `fusion_output` run.

Table 1: Validation performance for each fusion strategy (200 epochs).

Fusion Method	Acc. (%)	F1
BiCoAttention	**28.73**	**0.2782**
GatedFusion	24.62	0.2369
AttnOriginal	23.68	0.2355
FiLMFusion	20.43	0.1942
Transformer	1.17	0.0020

5.2 Interpretation

BiCoAttention achieves the best validation accuracy (28.73%) and macro F1 (0.278). The bilateral co-attention design allows both modalities to attend to each other and to combine the two views in a structured way, which appears beneficial for this task. The unified embedding dimension and symmetric treatment of both streams may improve gradient flow and representation quality compared to the baseline attention fusion.

GatedFusion and **AttentionFusionOriginal** are close: 24.62% and 23.68% accuracy respectively. The original LRW-AR attention fusion is competitive but slightly behind GatedFusion in this run.

FiLMFusion attains 20.43% accuracy. The additive modulation $\gamma\mathbf{V} + \beta$ may be less expressive than full attention for this dataset, or may require more tuning (e.g., deeper conditioning or residual connections). FiLM is parameter-efficient but does not model query-key-value interactions between modalities.

TransformerFusion collapses to near-random performance (1.17% accuracy, F1 $\approx$ 0.002). This suggests that the current configuration (e.g., 2-layer encoder, no explicit positional encoding, or learning rate) is unsuitable; transformer-based fusion may require different depth, normalization, or curriculum [21] to succeed on this feature setup.

5.3 Limitations

Our comparison is based on validation metrics only. Reported performance may depend on the computational resources available: hardware (e.g., GPU type and memory), batch size, and the capacity to run longer training or larger models can influence both convergence and final accuracy. We do not tune hyperparameters per fusion type; a dedicated search might change the ranking. The fusion modules are evaluated in isolation; combining the best fusion with a stronger backbone or data augmentation could yield higher absolute performance.

6 Conclusion and Future Work

We presented a comparative study of five multimodal fusion strategies for visual speech recognition on LRW-AR, using pre-extracted visual and geometric features and a common training protocol (200 epochs). BiCoAttention achieved the best validation accuracy (28.73%) and macro F1 (0.278), followed by GatedFusion and the original attention fusion; FiLM was weaker, and TransformerFusion failed to learn under the current setup. The results suggest that bilateral co-attention is a promising choice for fusing visual and landmark streams in lip-reading, and that transformer-based fusion requires further tuning (depth, normalization, learning rate, or pretraining) to be effective. We provided detailed mathematical descriptions of each fusion module and discussed limitations.

Future work may include: (1) ablating hyperparameters per fusion type; (2) investigating why TransformerFusion collapses and testing remedies such as layer normalization, reduced depth, or learning rate warmup; (3) combining the best fusion (e.g., BiCoAttention) with stronger backbones or data augmentation; (4) applying the same fusion comparison to other VSR benchmarks (e.g., LRW, LRS2 [5, 11]) to assess generality; and (5) analyzing attention weights to interpret what each fusion strategy learns.

Statements and Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The LRW-AR dataset and accompanying materials analysed in this study are publicly available at `https://github.com/crns-smartvision/lrwar` (accessed 2024).

Authors' contributions All authors contributed to the study conception and design, drafting or critically revising the manuscript, and approved the version submitted.

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgements

The authors thank the maintainers of the LRW-AR benchmark and codebase for providing the dataset and reference implementation.

References

[1] S. Fenghour, D. Chen, K. Guo, B. Li, P. Xiao. Deep Learning-Based Automated Lip-Reading: A Survey. *IEEE Access* 9 (2021) 121184–121205 (Article no. 9522117). `https://doi.org/10.1109/ACCESS.2021.3107946`.

[2] A. Fernandez-Lopez, F.M. Sukno. Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing* 78 (2018) 53–72. `https://doi.org/10.1016/j.imavis.2018.07.002`.

[3] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic. End-to-end audiovisual speech recognition. *Proc. ICASSP*, 2018.

[4] G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE* 91 (9) (2003) 1306–1326.

[5] J.S. Chung, A. Zisserman. Lip reading in the wild. *Proc. ACCV*, 2016.

[6] LRW-AR: Arabic lip-reading dataset and model. `https://github.com/crns-smartvision/lrwar` (accessed 2024).

[7] A. Vaswani et al. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017) 5998–6008.

[8] E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. Courville. FiLM: visual reasoning with a general conditioning layer. *Proc. AAAI*, 2018.

[9] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier. Language modeling with gated convolutional networks. *Proc. ICML*, 2017.

[10] J. Lu, D. Batra, D. Parikh, S. Lee. ViLBERT: pretraining task-agnostic visiolinguistic representations from vision and language. *Proc. NeurIPS*, 2019.

[11] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (12) (2019) 8717–8727. https://doi.org/10.1109/TPAMI.2018.2889052.

[12] Y.M. Assael, B. Shillingford, S. Whiteson, N. de Freitas. LipNet: end-to-end sentence-level lipreading. *Proc. ICLR*, 2017.

[13] M. Hao, M. Mamut, N. Yadikar, A. Aysa, K. Ubul. A Survey of Research on Lipreading Technology. *IEEE Access* 8 (2020) 204078–204099. https://doi.org/10.1109/ACCESS.2020.3036865.

[14] P. Ma, S. Petridis, M. Pantic. Deep learning based automated lip-reading: a survey. *IEEE Transactions on Cognitive and Developmental Systems* 14 (4) (2022) 1138–1152.

[15] T. Stafylakis, G. Tzimiropoulos. Combining residual networks with LSTMs for lipreading. *Proc. INTERSPEECH*, 2017.

[16] S. Yan, Y. Xiong, D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proc. AAAI*, 2018.

[17] T. Baltrusaitis, C. Ahuja, L.-P. Morency. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (2) (2018) 423–443.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng. Multimodal deep learning. *Proc. ICML*, 2011, pp. 689–696.

[19] P.K. Atrey, M.A. Hossain, A. El Saddik, M.S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16 (6) (2010) 345–379.

[20] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian. Deep modular co-attention networks for visual question answering. *Proc. CVPR*, 2019.

[21] R. Xiong et al. On layer normalization in the transformer architecture. *Proc. ICML*, 2020.

[22] S. Jain, B.C. Wallace. Attention is not explanation. *Proc. NAACL-HLT*, 2019, pp. 3543–3556. https://doi.org/10.18653/v1/N19-1357.