Supplementary Material for GateFusion: Hierarchical Gated Cross-Modal Fusion for Active Speaker Detection

Yu Wang Juhyung Ha Frangil M. Ramirez Yuchen Wang David J. Crandall

Indiana University
Bloomington, Indiana, USA
{yw173, juhha, fraramir, wang617, djcran}@iu.edu

A. Pretrained Backbone Configurations

Method	**Pretrain-V**	**Pretrain-A**
ASC (CVPR'20) [1]	ResNet-18 [10]	N/A
TalkNet (MM'21) [32]	N/A	N/A
ASDNet (ICCV'21) [17]	3D-ResNeXt [16]	N/A
MAAS (ICCV'21) [3]	ResNet-18	ResNet-18
EASEE-50 (ECCV'22) [2]	3D-ResNet [9]	ResNet-18
SPELL (ECCV'22) [20, 22]	N/A	N/A
Sync-TalkNet (MLSP'22) [37]	N/A	ResNet-34 [10]
ASD-Transformer (ICASSP'22) [4]	N/A	N/A
LightASD (CVPR'23) [18]	N/A	N/A
STHG (CVPRW'23) [21]	N/A	N/A
TS-TalkNet (Interspeech'23) [13]	N/A	N/A
TalkNCE (ICASSP'24) [14]	N/A	VGGish [11]
BIAS (T-BIOM'24) [30]	N/A	N/A
ASDnB (arXiv'24) [28]	N/A	N/A
LoCoNet (CVPR'24) [35]	N/A	VGGish
LR-ASD (IJCV'25) [19]	N/A	N/A

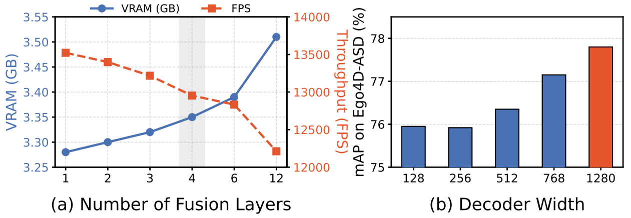
Table 1. Pretrained backbone configurations of state-of-the-art ASD models. **Pretrain-V/A** denote pretrained weights for video/audio encoders (N/A: trained from scratch). For TalkNCE, we use its strongest LoCoNet-based variant. TS-TalkNet trains encoders from scratch but includes a pretrained ECAPA-TDNN [6] for extracting target-speaker embeddings. Though initialized from scratch, ASDnB is pretrained on WASD [29] for AVA-ActiveSpeaker [27], and LoCoNet on AVA-ActiveSpeaker for Ego4D-ASD [8].

Method	**mAP (Baseline)**	**mAP (+ Pretrained Enc.)**
LoCoNet [35]	59.3	60.7
TalkNet [32]	51.7	70.7

Table 2. Performance of representative ASD models before and after adopting our pretrained video and audio encoders on the Ego4D-ASD benchmark. Note that LoCoNet is trained from scratch here.

We summarize in Table 1 the pretrained backbone configurations adopted by the state-of-the-art models. Prior works exhibit variability in their initialization strategies: some rely on pretrained encoders for their visual or audio branches, e.g., a ResNet-18 [10] pretrained on ImageNet [5], while others train all components strictly from scratch. Our model adopts pretrained weights for both video and audio encoders (we retain only the first 12 layers of each pretrained checkpoint to match the depth of our encoders). To verify that our performance improvements do not arise solely from pretrained initialization, we additionally equip two representative baselines, TalkNet [32] and LoCoNet [35], with the same pretrained video and audio encoders used in our framework and evaluate all models under identical conditions on the Ego4D-ASD benchmark [8]. This dataset is particularly challenging and serves as a rigorous testbed for assessing the true utility of pretrained backbones. As reported in Table 2, simply adding our pretrained encoders to existing architectures improves performance but does not match the gains achieved by our full model, thereby demonstrating that our contributions extend beyond encoder initialization.

B. Additional Ablation Studies



(a) Number of Fusion Layers

(b) Decoder Width

Figure 1. Ablations on model hyperparameters. (a) We visualize the trade-off between memory cost (VRAM, blue solid line) and inference throughput (FPS, orange dashed line) as the number of fusion layers increases. The gray band highlights our selected configuration ($N = 4$). (b) Performance across different decoder widths.

We provide additional ablations regarding the number of fusion layers and decoder width on the Ego4D-ASD benchmark [8], as illustrated in Fig. 1. The models are trained with the MAL and OPP objectives.

As discussed in the main paper, increasing N generally improves model capacity. The specific configurations are detailed in Table 3, ranging from single-layer fusion to fully

#Fusion Layers	Layer Index
1	[10]
2	[7, 10]
3	[4, 7, 10]
4	[1, 4, 7, 10]
6	[1, 3, 5, 7, 9, 11]
12	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]

Table 3. Fusion layer configurations for the ablation on the number of fusion layers. **Layer Index** lists the indices (starting from 1) of the Transformer blocks used for fusion.

dense fusion ($N = 12$). However, Fig. 1(a) reveals the associated computational costs: a larger N leads to increased memory footprint and a corresponding drop in inference throughput. We select $N = 4$ as it strikes a balance between efficiency and performance. Fig. 1(b) shows that increasing decoder width steadily improves performance. We adopt 1280 for optimal results.

C. Additional Losses for Model Robustness

Method	AV Prediction (mAP)		
	Audio Noise	Video Blur	Clean
Baseline	60.10	70.80	76.33
Baseline+MAL+OPP	**64.06**	**71.73**	**77.80**

Table 4. Robustness against data corruption. Comparison of AV prediction performance under audio noise and video blur conditions. Our method demonstrates superior robustness compared to the baseline.

In this section, we evaluate the robustness of our proposed method against data quality degradation. We simulate two types of degradation: noisy audio and blurry video, reflecting common real-world scenarios. Specifically, we introduce additive white noise to the audio stream. For the video stream, we implement a heterogeneous blur protocol, where Gaussian blur and Radial blur are applied to the dataset in an equal 50/50 split. We compare our full model (incorporating MAL and OPP) against the baseline without these auxiliary losses.

As shown in Table 4, our proposed framework demonstrates superior stability under sensory corruption. In the Audio Noise setting, the baseline suffers a relative performance drop of 21.3% compared to the clean setting, whereas our full model limits the degradation to 17.7%. Furthermore, under the Video Blur condition, our method consistently maintains a performance advantage over the baseline (71.73% vs. 70.80%). These results indicate that the proposed auxiliary constraints effectively enable the model to recover cues from the corrupted modality by leveraging the context of the clean modality.

We further evaluate the extreme case of complete modality loss, observing a marginal degradation compared to the baseline. This expected trade-off evidences the deeper multimodal interaction learned by our method. Unlike the baseline which processes modalities independently, our approach enforces tight coupling. Consequently, the fusion module relies on learned cross-modal synergy, rendering a missing modality an out-of-distribution input rather than a simple information loss.

D. Frozen Encoder Performance

To assess the impact of encoder fine-tuning, we conduct an experiment where both the video and audio encoders are entirely frozen and only the decoder parameters are trained. In this setting, the model achieves an mAP of 68.16% on the Ego4D-ASD benchmark. Although this is lower than our full fine-tuning setup (76.33% mAP), the performance remains competitive compared to prior works such as LoCoNet (68.4%) and SPELL (60.7%), despite using fixed encoders and no end-to-end optimization.

E. Comparing with Other Fusion Methods

Fusion Method	Fusion Stage	#Layers	mAP (%)
Owens'Fusion [24]	Early	1	37.63
Messenger [38]	Mid	1	73.81
DeepAVFusion [23]	Multi-layer	12	66.06
CATNet [36]	Multi-layer	3	68.15
MLCA [34]	Multi-layer	3	72.01
BiAVIGATE [12]	Multi-layer	4	73.97
HiGate	Multi-layer	4	**76.33**

Table 5. Comparison with other audio-visual fusion strategies. BiAVIGATE denotes bidirectional AVIGATE [12].

Early, mid, and multi-layer fusion strategies are rarely explored in ASD but are well studied in related multimodal works. We implement six representative audio-visual fusion methods on our encoders (12-layer truncated AV-HuBERT [31] and Whisper [26]) without task-specific components or additional losses, adhering closely to the original settings and using official code when available. The compared methods include Owens and Efros' method (Owens'Fusion for simplicity) [24], Messenger [38], DeepAVFusion [23], CATNet [36], MLCA [34], and bidirectional AVIGATE (BiAVIGATE for simplicity) [12]. Results are presented in Table 5.

We implement each fusion method following the configuration described in the original papers. Owens'Fusion [24] combines shallow unimodal features into a single branch. Accordingly, we remove Whisper [26] Transformer blocks, expand the convolutional front end to extract shallow audio features, and process the fused features with the 12-layer AV-HuBERT [31] branch, which is intrinsically designed for audio-visual input. The absence of a strong audio encoder substantially reduces mAP.

Both Messenger [38] and AVIGATE [12] rely on strong pretrained audio and visual encoders. Messenger uses a pretrained ResNet-152 [10] and an 18-layer R(2+1)D [33] model for visual features, together with a pretrained VGGish network [11] for audio features. AVIGATE employs a pretrained CLIP [25] visual encoder (ViT-B) and a pretrained 12-layer AST [7] for audio. Since both methods already rely on powerful encoders, we replace them with our own and re-implement their fusion blocks on top. In addition, because AVIGATE [12] treats audio and video asymmetrically, we extend it to bidirectional fusion to align with our framework.

DeepAVFusion [23] introduces a fusion stream that integrates multimodal features at each Transformer block of both modalities. We follow this setting to adapt their fusion strategy to our 12-layer encoders. CATNet [36] and MLCA [34] both perform multi-stage fusion, but at different depths. For CATNet [36], we split our 12-layer encoders into two stages: the first six layers apply shallow fusion and the remaining six apply middle fusion, with the outputs combined through late fusion. For MLCA [34], we insert fusion blocks after layers 4, 8, and the final layer of our encoders, following the progressive multi-layer fusion design. Each block applies self-attention within each modality, followed by bidirectional cross-modal attention with residual feedback, before returning the updated features to the encoders.

Table 5 shows that the early-fusion baseline (Owens'Fusion [24]) performs worst (37.63% mAP). Dense early fusion across all layers (DeepAVFusion [23], 12 layers) lags at 66.06% mAP, suggesting that aggressive early coupling is suboptimal for ASD. A single mid-layer fusion (Messenger [38]) improves performance (73.81% mAP) but fails to capture stage-specific cues. Multi-layer designs generally help: MLCA [34] reaches 72.01% mAP, CATNet [36] 68.15% mAP, and our bidirectional adaptation of AVIGATE [12] attains 73.97% mAP. Our proposed HiGate achieves the best performance (76.33% mAP), outperforming the next best baseline (BiAVIGATE [12]) by +2.36% mAP, and surpassing Messenger [38] and MLCA [34] by +2.52% and +4.32% mAP, respectively.

F. Efficiency Analysis

We report inference runtime and VRAM consumption in Table 6 using a single NVIDIA L40S GPU. For this, we followed the evaluation protocol reported by LR-ASD [19]. Ordered from most to least efficient, the models rank as follows: LR-ASD, TalkNet [32], our proposed GateFusion, and LoCoNet [35]. Compared to the previous SOTA (LoconNet), GateFusion not only surpasses its performance but also shows greater efficiency in both VRAM usage and runtime. Given 1,000 frames, GateFusion consumes 33% less VRAM with a 75% faster runtime.

Method	Video frames	VRAM (GB)	Time (ms)	FPS
LR-ASD [19]	1000	1.25	38.63	25890
	2000	2.50	78.64	25431
	4000	4.99	156.09	25627
	6000	7.48	236.32	25390
TalkNet [15]	1000	1.89	48.65	20556
	2000	3.52	102.52	19508
	4000	6.96	213.88	18702
	6000	10.32	328.01	18292
LoCoNet [35]	1000	5.02	135.45	7383
	2000	10.19	274.97	7273
	4000	Out of Memory		
GateFusion	1000	3.35	77.20	12953
	2000	4.96	174.37	11470
	4000	8.17	423.37	9448
	6000	11.37	749.56	8005

Table 6. Efficiency analysis using inference VRAM consumption, runtime, and FPS. Comparing with previous SOTA, LoCoNet, our model shows better efficiency in all metrics.

References

[1] Juan León Alcázar, Fabian Caba, Long Mai, Federico Perazzi, Joon-Young Lee, Pablo Arbeláez, and Bernard Ghanem. Active speakers in context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12465–12474, 2020. 1

[2] Juan Leon Alcazar, Moritz Cordes, Chen Zhao, and Bernard Ghanem. End-to-end active speaker detection. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022. 1

[3] Juan León Alcázar, Fabian Caba Heilbron, Ali K. Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 265–274, 2021. 1

[4] Gourav Datta, Tyler Etchart, Vivek Yadav, Varsha Hedau, Pradeep Natarajan, and Shih-Fu Chang. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4568–4572. IEEE, 2022. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proc. Interspeech 2020*, pages 3830–3834, 2020. 1

[7] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio

Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 3

[8] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 1, 3

[12] Boseung Jeong, Jicheol Park, Sungyeon Kim, and Suha Kwak. Learning audio-guided video representation with gated attention for video-text retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26202–26211, 2025. 2, 3

[13] Yidi Jiang, Ruijie Tao, Zexu Pan, and Haizhou Li. Target active speaker detection with audio-visual cues. In *Proc. Interspeech*, 2023. 1

[14] Chaeyoung Jung, Suyeon Lee, Kihyun Nam, Kyeongha Rho, You Jin Kim, Youngjoon Jang, and Joon Son Chung. Talknce: Improving active speaker detection with talk-aware contrastive learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8391–8395. IEEE, 2024. 1

[15] You Jin Kim, Hee Soo Heo, Soyeon Choe, Soo Whan Chung, Yoohwan Kwon, Bong Jin Lee, Youngki Kwon, and Joon Son Chung. Look who's talking: Active speaker detection in the wild. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 4411–4415. International Speech Communication Association, 2021. 3

[16] Okan Kopuklu, Neslihan Kose, Ahmet Gunduz, and Gerhard Rigoll. Resource efficient 3d convolutional neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 1

[17] Okan Köpüklü, Maja Taseska, and Gerhard Rigoll. How to design a three-stage architecture for audio-visual active speaker detection in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1193–1203, 2021. 1

[18] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, and Liangyin Chen. A light weight model for active speaker detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22932–22941, 2023. 1

[19] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, Liangyin Chen, and Yanru Chen. Lr-asd: Lightweight and robust network for active speaker detection. *International Journal of Computer Vision*, 133(7): 4749–4769, 2025. 1, 3

[20] Kyle Min. Intel labs at ego4d challenge 2022: A better baseline for audio-visual diarization. *arXiv preprint arXiv:2210.07764*, 2022. 1

[21] Kyle Min. Sthg: Spatial-temporal heterogeneous graph learning for advanced audio-visual diarization. *arXiv preprint arXiv:2306.10608*, 2023. 1

[22] Kyle Min, Sourya Roy, Subarna Tripathi, Tanaya Guha, and Somdeb Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. In *European conference on computer vision*, pages 371–387. Springer, 2022. 1

[23] Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27186–27196, 2024. 2, 3

[24] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018. 2, 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 2

[27] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1

[28] Tiago Roxo, Joana C Costa, Pedro Inácio, and Hugo Proença. Asdnb: Merging face with body cues for robust active speaker detection. *arXiv preprint arXiv:2412.08594*, 2024. 1

[29] Tiago Roxo, Joana C Costa, Pedro RM Inácio, and Hugo Proença. Wasd: A wilder active speaker detection dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 1

[30] Tiago Roxo, Joana C. Costa, Pedro R. M. Inácio, and Hugo Proença. Bias: A body-based interpretable active speaker approach. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 1

[31] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representa-

tion by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 2, 3

[32] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3927–3935, 2021. 1, 3

[33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[34] He Wang, Pengcheng Guo, Pan Zhou, and Lei Xie. Mlca-avsr: Multi-layer cross attention fusion based audio-visual speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8150–8154. IEEE, 2024. 2, 3

[35] Xizi Wang, Feng Cheng, and Gedas Bertasius. Loconet: Long-short context network for active speaker detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18462–18472, 2024. 1, 3

[36] Xingmei Wang, Jiachen Mi, Boquan Li, Yixu Zhao, and Jiaxiang Meng. Catnet: Cross-modal fusion for audio–visual speech recognition. *Pattern Recognition Letters*, 178:216–222, 2024. 2, 3

[37] Abudukelimu Wuerkaixi, You Zhang, Zhiyao Duan, and Changshui Zhang. Rethinking audio-visual synchronization for active speaker detection. In *2022 IEEE 32nd international workshop on machine learning for signal processing (MLSP)*, pages 01–06. IEEE, 2022. 1

[38] Yating Xu, Conghui Hu, and Gim Hee Lee. Rethink cross-modal fusion in weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5615–5624, 2024. 2, 3