Robust Audio-Visual Target Speaker Extraction with Multimodal Enrollment Fusion

Zhan Jin[1,2], Bang Zeng[1], Peijun Yang[3,4], Jiarong Du[3,4], Juan Liu[3,4], and Ming Li[2,3✉⋆]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] School of Artificial Intelligence, Chinese University of Hong Kong, Shenzhen, China
[3] School of Artificial Intelligence, Wuhan University, Wuhan, China
[4] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
mingli369@cuhk.edu.cn

Abstract. Audio-Visual Target Speaker Extraction (AVTSE) commonly relies on frame-level lip motion, which may be intermittently unavailable because of occlusion, head movement, or video transmission failure. This paper studies multimodal enrollment fusion under temporally contiguous visual occlusion. Each masked segment spans 15–25 frames, while the occlusion ratio and segment positions are randomized at test time. Using TF-GridNet, we compare lip motion with utterance-level face and enrollment-speech cues and frame-level expression cues, and evaluate models trained without occlusion or fine-tuned under fixed or randomly sampled occlusion ratios. A Lip-Face model trained without occlusion decreases from 12.68 to 5.21 dB SI-SNR under 70–90% test occlusion, whereas fixed 80% occlusion fine-tuning retains 12.26 dB. The same trend is observed with SpatialNet, where the corresponding high-occlusion result increases from 4.50 to 9.91 dB. These results indicate that robustness depends more strongly on exposure to temporally missing visual cues during training than on adding modalities alone. Demo samples are available at https://anonymous.4open.science/r/modality-missing-robustness-D9D2.

Keywords: audio-visual target speaker extraction · multimodal enrollment · visual occlusion · robustness

1 Introduction

The "cocktail party effect" describes the human ability to attend to a desired speaker in a complex acoustic environment. Target Speaker Extraction (TSE) addresses a related computational problem by estimating a target speech signal from a mixture with the assistance of speaker-dependent enrollment information. In audio-visual target speaker extraction (AVTSE), visual observations provide information that is complementary to the acoustic mixture and is not directly corrupted by acoustic interference.

⋆ Corresponding author: Ming Li

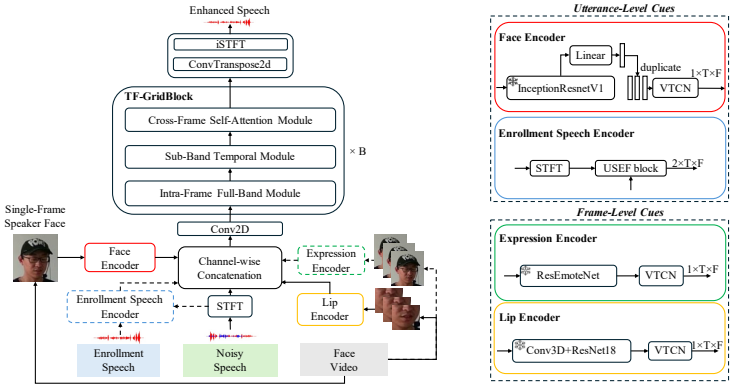


Fig. 1: AVTSE system with multimodal enrollment fusion. Lip motion and facial expression are frame-level cues, whereas the face image and enrollment speech provide utterance-level target-speaker information.

Existing AVTSE systems primarily use frame-level lip motion because its temporal dynamics are correlated with speech production [1–5]. Afouras et al. employed a pretrained lip-reading model to derive visual embeddings [2, 6], while USEV and related work examined the contribution of lip motion and audio-visual synchrony under different overlap conditions [7, 8]. Frame-level visual cues, however, may be unavailable for contiguous periods because of face occlusion, head pose changes, tracking failure, or video transmission interruption. This mismatch between complete visual observations during training and incomplete observations during deployment can cause substantial performance degradation.

Prior studies have considered incomplete or complementary target-speaker observations from several perspectives. ImagineNet combines visual embedding inpainting with target speaker extraction [9]. VisualVoice uses face identity and lip information within a cross-modal learning framework [10]. Afouras et al. investigated audio-visual speech enhancement under visual obstruction and used estimated speech as an additional enrollment source [15]. Gogate et al. further integrated raw enrollment audio with visual observations for personalised audio-visual speech enhancement [34]. These studies establish the relevance of incomplete visual observations and multimodal target-speaker conditioning, but the joint effects of cue type, train–test occlusion mismatch, occlusion-ratio distribution, and separator backbone remain insufficiently characterized within a common experimental protocol.

This study therefore examines intermittent visual-cue loss through multimodal and training-strategy comparisons. Within a TF-GridNet-based AVTSE system, we compare lip motion, a single face image, facial expression, and enrollment speech under five test conditions ranging from complete visual availability to a randomly sampled 70–90% test-time occlusion ratio. We then fine-tune three separate Lip+Face models: one with a fixed 80% training ratio, one with a per-utterance ratio sampled from $\mathcal{U}(30\%, 90\%)$, and one with a per-utterance

ratio sampled from $\mathcal{U}(70\%, 90\%)$. Each of these models is evaluated under all five common test conditions. Accordingly, $\mathcal{U}(30\%, 90\%)$ denotes the training distribution of one fine-tuned model rather than separate models associated with the 30–50%, 50–70%, and 70–90% test intervals. Finally, we repeat the Lip-Face comparison with SpatialNet [35] to determine whether the observed robustness pattern is specific to TF-GridNet. The analysis emphasizes measured behavior under the evaluated conditions rather than treating any individual modality or masking distribution as universally preferable.

2 Related Work

Audio-visual speech processing uses the correspondence between articulatory motion and acoustic speech to support enhancement, separation, and target-speaker extraction. Early visually guided systems demonstrated that mouth-region observations can improve speech reconstruction under acoustic interference [1, 2]. Later methods developed time-domain and time–frequency-domain fusion mechanisms that align visual features with the mixture representation [3, 8, 5]. In target speaker extraction, the visual stream also identifies which source should be recovered. USEV formulates lip motion as a generally applicable target-speaker cue [7], while scenario-aware TF-GridNet combines audio-visual conditioning with a strong time–frequency separator [16]. Most such systems are evaluated with continuously available visual observations, although this assumption may not hold when the face is occluded or tracking fails.

Target-speaker information can be provided by temporally aligned cues, utterance-level identity cues, or both. Lip motion is synchronized with speech production and supplies information about target activity and articulation. A still face image is not synchronized with the acoustic content but can provide relatively stable identity-related information. VisualVoice combines face identity and speech-related visual observations through cross-modal consistency learning [10]. Enrollment speech supplies target identity in the acoustic domain; USEF-TSE directly models the interaction between enrollment and mixture representations without relying on a separately pretrained speaker embedding of fixed dimension [14]. Gogate et al. similarly integrate enrollment audio with visual observations for personalised audio-visual speech enhancement [34]. These cue types differ in their availability under intermittent video loss, which motivates the distinction between frame-level and utterance-level enrollment adopted in this paper.

Incomplete visual input has been investigated through explicit reconstruction and obstruction-aware enhancement. ImagineNet reconstructs intermittent visual embeddings before target speaker extraction [9]. Afouras et al. study audio-visual speech enhancement when the mouth is concealed and use estimated speech to supplement unavailable visual information [15]. The present work instead asks how cue type and masking-aware fine-tuning affect the robustness of an otherwise unchanged AVTSE system. The protocol uses contiguous 15–25-frame masks rather than isolated frame dropout and explicitly separates

the fine-tuning ratio distribution from the test-condition ratio distribution. In particular, a model fine-tuned with ratios sampled from 30–90% is one model evaluated under all test conditions; it is not divided into separate models associated with the 30–50%, 50–70%, and 70–90% test intervals.

3 Methods

3.1 Problem formulation

Let $x \in \mathbb{R}^N$ denote a single-channel mixture containing target speech s and an interference component v:

$$x = s + v. \tag{1}$$

After the STFT, the complex mixture is represented by $X \in \mathbb{C}^{F\times T}$. The available target-speaker observations are written as

$$\mathcal{C} = \{Z^{\text{lip}}, Z^{\text{expr}}, e^{\text{face}}, E^{\text{enr}}\}, \tag{2}$$

where Z^{lip} and Z^{expr} are frame-level sequences, e^{face} is an utterance-level face representation, and E^{enr} is the enrollment-speech representation. A model uses the subset of $\mathcal{C}$ specified by its cue configuration. Each cue is projected to a representation compatible with the mixture feature; utterance-level features are aligned over the mixture time axis, while frame-level features are temporally resampled when necessary. Denoting multimodal projection and fusion by $\Phi(\cdot)$ and the separator by $G_\theta(\cdot)$, the estimate is

$$\widehat{S} = G_\theta(X, \Phi(\mathcal{C})), \qquad \widehat{s} = \text{iSTFT}(\widehat{S}). \tag{3}$$

The robustness question is whether G_θ continues to recover s when a large fraction of the frame-level elements in $\mathcal{C}$ are unavailable. The comparisons therefore keep the backbone and evaluation protocol fixed while changing either the cue subset or the distribution of visual masks used during fine-tuning.

3.2 System overview

The system in Fig. 1 contains a speech encoder, modality-specific target speaker encoders, a fusion module, a separation backbone, and a speech decoder. The speech encoder applies a short-time Fourier transform (STFT) to the mixture. Except for the enrollment-speech pathway, the visual embeddings are projected by modality-specific attractors based on a Visual Temporal Convolutional Network (V-TCN) [16]. The projected embeddings are aligned with the mixture representation, concatenated along the channel dimension, and fused by a two-dimensional convolution before separation.

TF-GridNet [11] is used as the main separation backbone. It contains six TFGridBlock modules combining full-band, sub-band temporal, and cross-frame attention operations. The decoder estimates the target complex spectrogram and reconstructs the waveform through inverse STFT. To assess whether the robustness trend transfers to a different separator, an additional experiment replaces TF-GridNet with SpatialNet [35] while retaining the Lip-Face enrollment configuration and the same test-time occlusion protocol.

3.3 Target speaker encoders

The evaluated cues include two utterance-level cues and two frame-level cues. Enrollment speech and a single face image provide utterance-level target-speaker information, whereas lip motion and facial expression provide temporally aligned frame-level information.

Enrollment speech encoder Both the mixture and enrollment speech are transformed into time–frequency representations. Enrollment features are extracted with the Universal Speaker Embedding Free target speaker extraction block [14], which uses learnable multi-head cross-attention between enrollment and mixture features. This pathway supplies speaker information from the acoustic domain without requiring a separately pretrained fixed-dimensional speaker embedding.

Face encoder A target face image is sampled from the corresponding video and encoded using InceptionResNetV1 [17]. The resulting embedding is repeated along the temporal dimension before projection by the modality attractor. Because this cue is derived from one sampled frame, it is treated as an utterance-level enrollment cue and is not removed by the intermittent frame-level masking operation.

Lip encoder The lip pathway uses the pretrained visual frontend adopted in [4], consisting of a three-dimensional convolution followed by residual convolutional blocks. Linear interpolation aligns the temporal resolution of the lip embeddings with that of the mixture representation before modality fusion.

Expression encoder Facial-expression embeddings are extracted with ResEmoteNet [18]. The classifier is removed and the intermediate frame-level representation is used as the enrollment feature. ResEmoteNet was developed using facial-expression datasets including FER2013, RAF-DB, AffectNet, and ExpW [19–22]. In the present study, this pathway is evaluated as a possible source of temporally varying paralinguistic information; no assumption is made that expression features necessarily encode speaker identity.

3.4 Intermittent occlusion protocol and training regimes

Intermittent visual loss is simulated by setting contiguous frame-level visual embeddings to zero. The same temporal mask is applied to lip and expression embeddings when both are present, whereas the utterance-level face and enrollment-speech cues remain available. Each occluded segment contains 15–25 consecutive frames, and segment positions are sampled randomly. Multiple segments are generated so that the total masked duration follows the specified utterance-level occlusion ratio.

Let ρ_{tr} denote the utterance-level occlusion ratio used during fine-tuning and ρ_{te} denote the ratio used to construct a test condition. These two variables are defined independently. Five test conditions are used. The first contains no occlusion. For each of the remaining conditions, ρ_{te} is independently sampled for every test utterance from $\mathcal{U}(10\%, 30\%)$, $\mathcal{U}(30\%, 50\%)$, $\mathcal{U}(50\%, 70\%)$, or $\mathcal{U}(70\%, 90\%)$. These four conditions are denoted by their interval centers as $20\% \pm 10\%$, $40\% \pm 10\%$, $60\% \pm 10\%$, and $80\% \pm 10\%$. Both ρ_{te} and the segment positions are randomized at test time, and every trained model is evaluated under all five conditions.

The principal TF-GridNet comparison uses a model trained without occlusion and three separately fine-tuned Lip+Face models. For the fixed-ratio model, $\rho_{\mathrm{tr}} = 80\%$ for every fine-tuning utterance, while segment positions remain random. For the other two models, a single ρ_{tr} is sampled independently for each fine-tuning utterance from $\mathcal{U}(30\%, 90\%)$ or $\mathcal{U}(70\%, 90\%)$. Thus, the 30–90% setting is one fine-tuning distribution applied within one model; it is not divided into separate models corresponding to the test-time intervals. The test intervals in the result tables describe only how ρ_{te} is sampled during evaluation. These comparisons separate the effect of exposure to missing visual frames from the effect of the training-ratio distribution.

Let $M = (m_1, \ldots, m_T) \in \{0, 1\}^T$ denote the temporal availability mask, where $m_t = 0$ indicates that the corresponding video observation is unavailable. For a frame-level embedding $Z \in \mathbb{R}^{D \times T}$,

$$\widetilde{Z}_{:,t} = m_t Z_{:,t}, \qquad \rho(M) = 1 - \frac{1}{T} \sum_{t=1}^{T} m_t. \tag{4}$$

When both lip and expression cues are present, the same M is applied to both sequences so that the corruption represents loss of the corresponding video interval. The face image and enrollment speech are not multiplied by M. The mask is composed of randomly positioned contiguous segments of 15–25 frames. Consequently, the fixed-80% regime fixes only the utterance-level value of ρ_{tr}; it does not reuse a fixed temporal pattern.

The training and test distributions describe different experimental axes. A row in Table 3 selects one distribution of ρ_{tr} and therefore one trained model. A column selects one independently defined distribution of ρ_{te} used to evaluate that model. This separation is necessary because the support of the $\mathcal{U}(30\%, 90\%)$ fine-tuning distribution overlaps several test intervals, but the fine-tuned parameters are shared across the entire range.

4 Experiments

4.1 Dataset

The 3rd COG-MHEAR Audio-Visual Speech Enhancement Challenge (AVSEC-3) data are constructed from several public corpora [23]. The mixtures contain

Table 1: TF-GridNet results after training without occlusion. Each cell reports SI-SNR/PESQ/STOI under the indicated test-time visual occlusion condition.

Cue configuration	0%	20% ± 10%	40% ± 10%	60% ± 10%	80% ± 10%
Mix	-4.86/1.176/0.615	-4.86/1.176/0.615	-4.86/1.176/0.615	-4.86/1.176/0.615	-4.86/1.176/0.615
Lip	12.22/2.401/0.874	9.32/2.073/0.817	7.29/1.873/0.767	5.64/1.783/0.719	4.38/1.786/0.704
Lip+Enroll	12.96/2.491/0.886	11.28/2.277/0.847	10.29/2.163/0.818	8.82/2.055/0.782	8.21/2.041/0.773
Lip+Expr	12.56/2.441/0.879	9.79/2.088/0.823	8.12/1.888/0.782	5.69/1.697/0.720	1.54/1.521/0.649
Lip+Face	12.68/2.454/0.882	9.61/2.081/0.818	7.74/1.901/0.764	5.89/1.797/0.711	5.21/1.834/0.699
Lip+Expr+Face	12.63/2.462/0.881	10.55/2.216/0.835	9.36/2.095/0.804	7.11/1.942/0.749	6.15/1.936/0.735
Lip+Enroll+Expr+Face	13.15/2.525/0.888	10.21/2.167/0.828	8.95/2.050/0.791	7.50/1.969/0.755	6.86/2.001/0.752

either target speech with one competing speaker or target speech with non-speech noise. Speech is drawn from LRS3 [25]; noise sources include CEC1 [26], DEMAND [27], the second DNS Challenge corpus [28], MedleyDB [29], and ESC-50 [30]. The robustness experiments use the AVSEC-3 test set. The development and leaderboard evaluation sets are used only for comparison with previously reported challenge systems.

4.2 Implementation details and evaluation metrics

Audio is sampled at 16 kHz. The STFT uses a window length of 128 samples and a hop length of 64 samples, producing 65 frequency bins. After multimodal fusion, the feature dimension is projected to 128 channels. Each TFGridBlock uses bidirectional LSTMs with 240 hidden units and 128 output features, followed by a one-dimensional transposed convolution that maps 256 channels to 128 channels.

The initial model is optimized with a scale-invariant objective. Occlusion fine-tuning uses a hybrid SI-SNR and STOI objective with a STOI weight of 0.5. Dynamic mixing is applied during training. The batch size is 32, and Adam is used with an initial learning rate of 10^{-3}. The learning rate is halved after three consecutive epochs without validation improvement, and training is terminated after ten such epochs. The robustness tables report SI-SNR, wide-band PESQ [32], and STOI [33]. The challenge comparison tables retain the SI-SDR terminology and PESQ configuration used by the corresponding evaluation protocol [31, 23]; therefore, their absolute values should not be directly merged with the robustness tables.

4.3 Results and analysis

Each cell in Tables 1–4 reports SI-SNR/PESQ/STOI. Nonzero test ratios are sampled independently for each utterance, and the temporal positions of all 15–25-frame segments are randomized.

Cue contributions without occlusion-aware training Table 1 shows that the complete four-cue system provides the highest clean-condition scores among

Table 2: TF-GridNet results after fine-tuning with a fixed 80% visual occlusion ratio. Each cell reports SI-SNR/PESQ/STOI.

Cue configuration	0%	20% ± 10%	40% ± 10%	60% ± 10%	80% ± 10%
Lip	12.34/2.427/0.876	12.26/2.417/0.874	12.14/2.404/0.871	12.06/2.397/0.870	11.94/2.384/0.868
Lip+Enroll	13.08/2.511/0.888	12.99/2.497/0.886	12.97/2.494/0.885	12.90/2.487/0.883	12.88/2.484/0.883
Lip+Expr	12.66/2.467/0.881	12.57/2.452/0.878	12.42/2.437/0.875	12.25/2.423/0.872	12.26/2.421/0.872
Lip+Face	12.82/2.479/0.884	12.71/2.467/0.881	12.54/2.453/0.878	12.49/2.450/0.877	12.26/2.430/0.872
Lip+Expr+Face	12.84/2.486/0.885	12.75/2.474/0.882	12.62/2.465/0.879	12.48/2.453/0.876	12.26/2.439/0.872
Lip+Enroll+Expr+Face	13.17/2.543/0.889	13.08/2.533/0.886	13.04/2.528/0.885	12.96/2.518/0.883	12.87/2.510/0.882

the evaluated configurations, reaching 13.15 dB SI-SNR, 2.525 PESQ, and 0.888 STOI. This result is consistent with complementarity under complete observations, but it does not imply robustness to unseen missing frames. Under the 80% ± 10% condition, the four-cue system decreases to 6.86 dB SI-SNR. The Lip+Enroll configuration retains 8.21 dB and is the strongest of the non-occlusion-trained models in SI-SNR at this occlusion level, suggesting that an utterance-level acoustic cue compensates for missing frame-level visual information more effectively than the evaluated face or expression representations.

The expression pathway does not provide consistent robustness when introduced without occlusion-aware training. Lip+Expr decreases from 12.56 to 1.54 dB SI-SNR between the clean and highest-occlusion conditions, which is lower than the Lip-only result of 4.38 dB. Adding both expression and face information improves the high-occlusion result to 6.15 dB, but this remains below Lip+Enroll. These observations indicate that clean-condition gains and missing-modality robustness should be assessed separately.

Effect of fixed high-occlusion fine-tuning Fine-tuning with fixed 80% masking substantially reduces sensitivity to the test occlusion ratio, as shown in Table 2. For Lip+Face, the SI-SNR difference between clean and 80% ± 10% test conditions is 0.56 dB, compared with 7.47 dB for the model trained without occlusion. The corresponding high-occlusion result increases from 5.21 to 12.26 dB. Lip-only also remains above 11.9 dB at the highest test ratio, indicating that the improvement cannot be attributed solely to the availability of an unmasked utterance-level cue.

The relative ranking of multimodal configurations is metric dependent. At 80%±10%, Lip+Enroll attains 12.88 dB SI-SNR and 0.883 STOI, while the four-cue system attains 12.87 dB SI-SNR, 2.510 PESQ, and 0.882 STOI. Because the numerical differences are small and no repeated-run uncertainty is available, these results do not support a statistically grounded claim that one of these two configurations is uniformly superior. They instead show that high-occlusion fine-tuning makes several cue combinations similarly stable across the evaluated test intervals.

Fixed versus per-utterance sampled training distributions Table 3 compares four Lip+Face models, rather than models separately trained for the test-

Table 3: Comparison of four TF-GridNet Lip+Face models under common test-time occlusion conditions. Each row corresponds to one trained model. For the two uniform fine-tuning regimes, one ρ_{tr} is sampled per training utterance from the stated distribution throughout fine-tuning; these rows do not represent multiple models aligned with the test intervals. In each nonzero test column, ρ_{te} is sampled independently per test utterance from the displayed interval. Each cell reports SI-SNR/PESQ/STOI.

Training regime (one model per row)	Test-time occlusion-ratio condition, ρ_{te}				
	0%	$20\% \pm 10\%$	$40\% \pm 10\%$	$60\% \pm 10\%$	$80\% \pm 10\%$
No occlusion training	12.68/2.454/0.882	9.61/2.081/0.818	7.74/1.901/0.764	5.89/1.797/0.711	5.21/1.834/0.699
FT: $\rho_{\text{tr}} = 80\%$	12.82/2.479/0.884	12.71/2.467/0.881	12.54/2.453/0.878	12.49/2.450/0.877	12.26/2.430/0.872
FT: $\rho_{\text{tr}} \sim \mathcal{U}(30\%, 90\%)$	12.76/2.472/0.883	12.70/2.462/0.881	12.52/2.445/0.878	12.28/2.424/0.873	12.09/2.404/0.869
FT: $\rho_{\text{tr}} \sim \mathcal{U}(70\%, 90\%)$	12.78/2.476/0.884	12.68/2.463/0.881	12.52/2.449/0.878	12.33/2.432/0.874	12.18/2.421/0.871

Table 4: SpatialNet results for Lip+Face enrollment with no occlusion training and fixed 80% occlusion fine-tuning. Each cell reports SI-SNR/PESQ/STOI.

Training regime	0%	$20\% \pm 10\%$	$40\% \pm 10\%$	$60\% \pm 10\%$	$80\% \pm 10\%$
No occlusion	10.69/2.123/0.847	8.23/1.857/0.794	6.78/1.743/0.756	5.42/1.678/0.722	4.50/1.662/0.716
Fixed 80%	10.70/2.118/0.847	10.46/2.094/0.843	10.31/2.076/0.840	10.08/2.056/0.835	9.91/2.039/0.832

time ratio intervals. In particular, the $\mathcal{U}(30\%, 90\%)$ row is produced by one model for which each fine-tuning utterance receives a ratio sampled from the full 30–90% range. After fine-tuning, that same model is evaluated under the five test conditions shown in the columns. The $\mathcal{U}(70\%, 90\%)$ row is constructed analogously. By contrast, the fixed-ratio model uses $\rho_{\text{tr}} = 80\%$ for every fine-tuning utterance. The 30–50%, 50–70%, and 70–90% intervals appearing in the test protocol therefore refer exclusively to test-time sampling of ρ_{te} and are not subdivisions of the $\mathcal{U}(30\%, 90\%)$ fine-tuning model.

All three occlusion-aware models are considerably more stable than the model trained without occlusion. Fixed 80% fine-tuning yields the highest SI-SNR at every nonzero test condition. Relative to the single model fine-tuned with $\rho_{\text{tr}} \sim \mathcal{U}(70\%, 90\%)$, its SI-SNR advantages are 0.03, 0.02, 0.16, and 0.08 dB across the four nonzero test conditions, respectively. The model fine-tuned with $\mathcal{U}(70\%, 90\%)$ is 0.05 and 0.09 dB higher than the $\mathcal{U}(30\%, 90\%)$ model at the two most severe test conditions, while the two models are similar at lower occlusion ratios.

These results do not show an advantage for broad training-ratio randomization over fixed high-ratio fine-tuning in the present setting. A plausible interpretation is that fixed 80% masking directly emphasizes severe missing-cue conditions while still exposing the model to random temporal positions and segment combinations. This interpretation remains provisional because the table reports one training run per fine-tuning regime and does not separate the effects of fine-tuning duration, optimization variance, and training-ratio distribution.

Cross-backbone validation with SpatialNet Table 4 shows the same qualitative train–test mismatch for SpatialNet as for TF-GridNet. Without occlusion-aware training, Lip+Face decreases from 10.69 to 4.50 dB SI-SNR as the test condition changes from no occlusion to 80%±10%. Fixed 80% fine-tuning retains 9.91 dB at the highest interval, an absolute improvement of 5.41 dB over the non-occlusion-trained model. Clean-condition SI-SNR is essentially unchanged, although PESQ decreases from 2.123 to 2.118. The transfer of this pattern to a second backbone supports the conclusion that exposure to temporally missing visual cues is not specific to the TF-GridNet separator. It does not, however, establish equivalence across architectures, because the two backbones have different absolute performance levels and were not subjected to repeated-run significance testing.

5 Discussion

The results distinguish multimodal complementarity from robustness to missing synchronized cues. Under complete observations, the four-cue system obtains the strongest clean-condition scores among the evaluated configurations, indicating that the selected modalities provide complementary information. Under severe unseen occlusion, however, Lip+Enroll is stronger than the other models trained without masking. Enrollment speech remains available throughout the utterance and therefore supplies a fallback target-speaker reference when most lip frames are absent. The face cue is also unmasked, but its benefit is smaller in the present implementation. This comparison should not be interpreted as a general ranking of acoustic and facial enrollment, because the two pathways use different encoders and fusion mechanisms.

The expression pathway illustrates why clean-condition improvements should not be equated with robustness. Lip+Expr slightly improves over Lip when no frames are removed, but its SI-SNR falls more sharply at the highest test interval. The result does not establish that expression is generally detrimental. It shows that a frame-level auxiliary representation can increase reliance on continuously available video unless the model is exposed to corresponding absence patterns. After fixed-80% fine-tuning, Lip, Lip+Expr, Lip+Face, and Lip+Expr+Face all remain near 12 dB SI-SNR under the highest test condition, indicating that masking-aware optimization changes how the separator uses frame-level conditioning.

The comparison among fine-tuning distributions does not support the assumption that a wider ratio range is necessarily preferable. Fixed 80% fine-tuning is slightly stronger than both sampled-ratio regimes at every nonzero test condition. Although its total masked proportion is constant, its segment positions and combinations remain random, so the model still observes diverse temporal patterns. The fixed and $\mathcal{U}(70\%, 90\%)$ regimes have the same mean ratio but different variance, whereas $\mathcal{U}(30\%, 90\%)$ changes both the mean and the support. The current experiment therefore cannot isolate whether mean severity or ratio variance is responsible for the small differences. Since only one training run is

Table 5: Performance on the AVSEC-3 development set under the evaluation plan in [23].

	Speech+Noise			Speech+Speech		
System	SI-SDR	PESQ	STOI	SI-SDR	PESQ	STOI
Mix	-4.4	1.15	0.68	-5.0	1.17	0.60
AV-CrossNet [24]	14.3	2.75	0.92	17.3	3.23	0.95
SAV-GridNet [16]	14.2	2.68	0.91	17.5	3.23	0.95
Lip-Face (ours)	14.4	2.57	0.91	17.5	3.08	0.95

Table 6: Performance on the AVSEC-3 leaderboard evaluation set. PESQ follows the challenge evaluation protocol.

System	SI-SDR	PESQ	STOI
ict_avsu_1	8.989	3.087	0.881
Smiip_try2	12.701	3.004	0.886
Lip-Face (ours)	13.259	3.087	0.894

reported for each regime, these differences should be treated as descriptive rather than statistically conclusive.

The SpatialNet experiment reproduces the principal qualitative trend: training only with complete video leads to a marked decline under missing frames, while severe-mask fine-tuning reduces the decline without materially lowering clean-condition SI-SNR. This agreement suggests that the observed train–test mismatch is not unique to TF-GridNet. It does not establish architectural equivalence, because the backbones have different absolute performance and the full set of fine-tuning distributions is not repeated with SpatialNet.

Several limitations remain. The corruption is applied by zeroing visual embeddings and therefore does not reproduce partial spatial occlusion, gradual tracking drift, or errors introduced before the visual encoder. The experiments also do not report variation across random training seeds or multiple independently generated test masks. Future evaluation should include natural visual failures, pre-encoder corruption, repeated training, and confidence intervals. Within the current scope, the evidence supports a practical distinction: additional modalities can improve complete-input performance or provide fallback identity information, whereas robustness to long missing intervals requires explicit exposure to missing cues during optimization.

5.1 Comparison with challenge systems

Tables 5 and 6 provide contextual comparisons under the official AVSEC-3 evaluation settings. These results concern complete-input challenge evaluation and are not substitutes for the controlled occlusion experiments above.

On the development set, Lip-Face has the same speech-plus-speech SI-SDR and STOI as SAV-GridNet, while its PESQ is lower. For speech-plus-noise mixtures, it has a 0.1 dB higher SI-SDR than AV-CrossNet but lower PESQ and STOI. On the leaderboard evaluation set, Lip-Face has the highest SI-SDR and STOI among the three systems listed and ties the reported PESQ of ict_avsu_1. These comparisons indicate competitive complete-input performance within the selected systems, but they do not demonstrate superiority over all published or submitted systems.

6 Conclusion

This study examines multimodal enrollment fusion for AVTSE under temporally contiguous visual occlusion. When models are trained only with complete visual observations, increasing test-time occlusion causes substantial degradation, and the cue combination that performs best without occlusion is not necessarily the most robust. Fine-tuning with severe intermittent masking markedly reduces this train–test mismatch for all evaluated TF-GridNet cue configurations. For Lip+Face, three separate single-model fine-tuning regimes are compared: fixed $\rho_{\mathrm{tr}} = 80\%$, per-utterance $\rho_{\mathrm{tr}} \sim \mathcal{U}(30\%, 90\%)$, and per-utterance $\rho_{\mathrm{tr}} \sim \mathcal{U}(70\%, 90\%)$. Every model is evaluated under the same independently defined test-time intervals. Fixed 80% fine-tuning performs slightly better under the tested nonzero conditions, although the differences among occlusion-aware models are small and require repeated-run validation. SpatialNet exhibits the same qualitative pattern, providing evidence that the benefit is not confined to one separator backbone. Overall, the experiments support treating occlusion-aware training as a separate design requirement from multimodal fusion and evaluating clean-condition accuracy and missing-cue robustness independently. The conclusions are limited to synthetic embedding-level masks and single-run comparisons; repeated training and naturally occurring visual failures are needed for stronger generalization claims.

Acknowledgements This research is funded in part by the National Natural Science Foundation of China (62571223) and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100).

References

1. Gabbay, A., Ephrat, A., Halperin, T., Peleg, S.: Seeing through noise: Visually driven speaker separation and enhancement. In: Proc. of ICASSP, pp. 3051–3055 (2018)
2. Alfouras, T., Chung, J., Zisserman, A.: The conversation: deep audio-visual speech enhancement. In: Proc. of INTERSPEECH, vol. 2018 (2018)
3. Wu, J., Xu, Y., Zhang, S.X., Chen, L.W., Yu, M., Xie, L., Yu, D.: Time domain audio visual speech separation. In: Proc. of ASRU, pp. 667–673 (2019)

4. Li, K., Xie, F., Chen, H., Yuan, K., Hu, X.: An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(10), 6637–6651 (2024)
5. Li, K., Yang, R., Sun, F., Hu, X.: IIANet: an intra-and inter-modality attention network for audio-visual speech separation. In: Proc. of ICML, pp. 29181–29200 (2024)
6. Stafylakis, T., Tzimiropoulos, G.: Combining Residual Networks with LSTMs for Lipreading. Proc. of INTERSPEECH (2017)
7. Pan, Z., Ge, M., Li, H.: USEV: Universal speaker extraction with visual cue. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 3032–3045 (2022)
8. Pan, Z., Tao, R., Xu, C., Li, H.: Selective listening by synchronizing speech with lips. IEEE/ACM Transactions on Audio, Speech, and Language Processing **30**, 1650–1664 (2022)
9. Pan, Z., Wang, W., Borsdorf, M., Li, H.: ImagineNet: Target speaker extraction with intermittent visual cue through embedding inpainting. In: Proc. of ICASSP, pp. 1–5 (2023)
10. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. In: Proc. of CVPR, pp. 15490–15500 (2021)
11. Wang, Z.Q., Cornell, S., Choi, S., Lee, Y., Kim, B.Y., Watanabe, S.: TF-GridNet: Integrating full- and sub-band modeling for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **31**, 3221–3236 (2023)
12. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. In: Proc. of ICASSP, pp. 21–25 (2021)
13. Luo, Y., Chen, Z., Yoshioka, T.: Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In: Proc. of ICASSP, pp. 46–50 (2020)
14. Zeng, B., Li, M.: Usef-tse: Universal speaker embedding free target speaker extraction. IEEE Transactions on Audio, Speech and Language Processing (2025)
15. Afouras, T., Chung, J.S., Zisserman, A.: My lips are concealed: Audio-visual speech enhancement through obstructions. In: Proc. of INTERSPEECH, vol. 2019, pp. 4295–4299 (2019)
16. Pan, Z., Wichern, G., Masuyama, Y., Germain, F.G., Khurana, S., Hori, C., Le Roux, J.: Scenario-aware audio-visual TF-Gridnet for target speech extraction. In: Proc. of ASRU, pp. 1–8 (2023)
17. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proc. of AAAI, vol. 31 (2017)
18. Roy, A.K., Kathania, H.K., Sharma, A., Dey, A., Ansari, M.S.A.: ResEmoteNet: bridging accuracy and loss reduction in facial emotion recognition. IEEE Signal Processing Letters (2024)
19. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., others: Challenges in representation learning: A report on three machine learning contests. In: Proc. of ICONIP, pp. 117–124 (2013)
20. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proc. of CVPR, pp. 2852–2861 (2017)
21. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE transactions on affective computing **10**(1), 18–31 (2017)

22. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: From facial expression recognition to interpersonal relation prediction. International Journal of Computer Vision **126**(5), 550–569 (2018)
23. Blanco, A.L.A., Valentini-Botinhao, C., Klejch, O., Gogate, M., Dashtipour, K., Hussain, A., Bell, P.: AVSE challenge: Audio-visual speech enhancement challenge. In: Proc. of SLT, pp. 465–471 (2023)
24. Kalkhorani, V.A., Yu, C., Kumar, A., Tan, K., Xu, B., Wang, D.: AV-CrossNet: An Audiovisual Complex Spectral Mapping Network for Speech Separation by Leveraging Narrow-and Cross-Band Modeling. IEEE Journal of Selected Topics in Signal Processing (2025)
25. Afouras, T., Chung, J.S., Zisserman, A.: LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018)
26. Graetzer, S., Barker, J., Cox, T.J., Akeroyd, M., Culling, J.F., Naylor, G., Porter, E., Viveros Munoz, R., others: Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing. In: Proc. of INTERSPEECH, vol. 2, pp. 686–690 (2021)
27. Thiemann, J., Ito, N., Vincent, E.: DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments. (No Title) (2013)
28. Dubey, H., Aazami, A., Gopal, V., Naderi, B., Braun, S., Cutler, R., Gamper, H., Golestaneh, M., Aichner, R.: ICASSP 2023 Deep Noise Suppression Challenge. In: Proc. of ICASSP (2023)
29. Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., Bello, J.: Medleydb: A dataset of multitrack audio for music research. In: Proc. of ISMIR (2014)
30. Piczak, K.J.: ESC: Dataset for Environmental Sound Classification. In: Proc. of ACMMM, pp. 1015–1018
31. Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R.: SDR–half-baked or well done?. In: Proc. of ICASSP, pp. 626–630 (2019)
32. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proc. of ICASSP, vol. 2, pp. 749–752 (2001)
33. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on audio, speech, and language processing **19**(7), 2125–2136 (2011)
34. Gogate, M., Dashtipour, K., Hussain, A.: Towards personalised audio visual speech enhancement. In: Proc. of INTERSPEECH, pp. 4853–4857 (2025). https://doi.org/10.21437/Interspeech.2025-1471
35. Quan, C., Li, X.: SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 1310–1323 (2024)