SPATIAL RESIDUAL ADAPTATION OF FROZEN AUDIO-VISUAL SPEAKER EXTRACTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-channel audio-visual target-speaker extractors leave the spatial information of microphone arrays unused. We propose D-CRPF, a 70.9K-parameter dual-direction complex residual post-filter that adds spatial evidence to a frozen extractor without accessing its hidden features or training graph. Two analytic beamformers condition complex reference correction and signed beam-disagreement correction around the frozen output. We characterize identity initialization, input-dependent correction bounds, and reference-preserving channel-permutation invariance. On simulated six-microphone mixtures covering all 1,243 full-duration official LRS2 test utterances, D-CRPF improves matched AV-CrossNet by 1.254 dB SI-SDR versus 0.045 dB from an equal-parameter single-channel control, reaching 14.433 dB and 8.14% WER. Under a fixed-reference circle-to-line change, the complete system loses 0.806 dB versus over 4 dB for direct multichannel references; its gain over Stage 1 remains positive but contracts from 1.246 to 0.441 dB. The residual-only intermediate variant, CRPF-R, isolates the first correction direction and provides complementary transfer evidence: it improves three released front ends by 0.904–1.502 dB SI-SDR and matched IIANet by 1.561 dB. Missing-video tests expose failure boundaries.

1 INTRODUCTION

Audio-visual target-speaker extraction uses visual speech to recover a designated speaker from competing speech and reverberation (Afouras et al., 2018; Gu et al., 2020; Gao & Grauman, 2021; Pan et al., 2022; Pegg et al., 2024). We ask how additional microphones can improve an existing extractor when its internal representations and training graph are unavailable. Our constraint is stronger than freezing original weights: the deployed front-end mapping itself remains unchanged, and only its estimated signal is exposed to adaptation. Direct multichannel separation (Quan & Li, 2024) instead learns a replacement mapping over the array.

A post-filter offers another route, but adding capacity alone does not establish a spatial benefit. Replacing the frozen output by a beam is also risky: imperfect target estimates can yield misleading spatial statistics, especially when source transfer functions are similar. We therefore preserve the AV estimate as an output anchor and use the beams as evidence for a learned correction, rather than assuming that beamforming already solves target extraction.

Our final method, D-CRPF, implements *spatial residual adaptation* in three steps (Figure 1): select the target, construct spatial evidence, and refine the estimate. A proxy from the frozen estimate guides two complementary MVDR views. Together with the reference spectrum and proxy, these views condition two correction directions: complex reference correction and signed beam-disagreement correction. Zero-initialized output heads recover Stage 1 before learning; the extractor remains fixed. The residual-only intermediate variant CRPF-R omits the signed branch to isolate its contribution.

This construction admits zero correction and bounds nonzero updates relative to the observations; it does not guarantee improvement after training. Beamforming reduces M channels to two complex spectra (four real maps). Together with the complex reference spectrum and proxy, they form seven maps independently of M. Reference-preserving channel relabeling leaves this interface unchanged, whereas microphone removal and physical array changes alter the observations and require empirical evaluation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

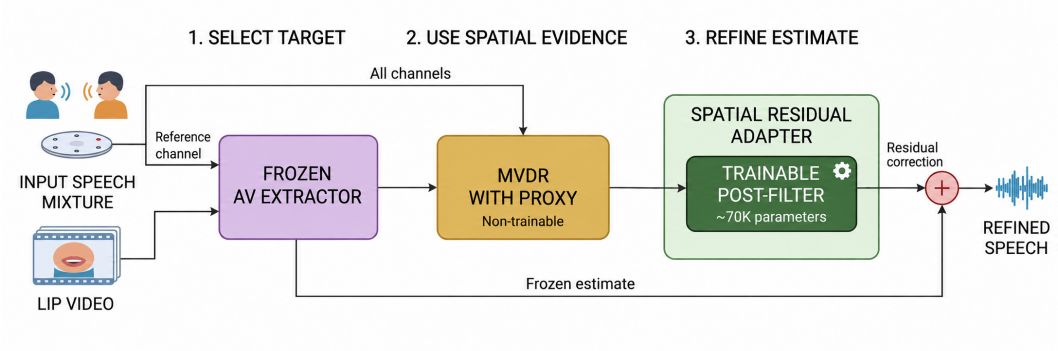

Figure 1: The proposed D-CRPF framework. A frozen AV extractor selects the target using one reference microphone and lip video. Proxy-guided MVDR supplies complementary spatial views; the compact adapter predicts two corrections added to the frozen estimate. Only the adapter is trained. Figures 2 and 3 detail the spatial interface and residual reconstruction.

We first test the shared spatial interface through intermediate CRPF-R on released Dolphin, AV-MossFormer2, and IIANet. We then develop final D-CRPF with matched AV-CrossNet and compare recent matched architectures, followed by capacity/decoder controls and array changes. Public-front-end and matched-IIANet results use separately trained CRPF-R, not the complete dual-direction decoder. Earlier AV-DPCRN studies provide mechanism and failure analyses. Main-table quality comparisons cover all 1,243 official LRS2 test utterances without 2-s truncation.

Our contributions are:

- D-CRPF, a 70.9K-parameter output-level adapter combining complex reference correction and signed spatial correction, without front-end hidden features or backbone updates;
- an analysis of identity initialization, input-dependent correction bounds, and reference-preserving channel-permutation invariance, with explicit limits on what these properties guarantee; and
- full-duration evaluation of final D-CRPF with capacity/decoder controls and array comparisons, supplemented by cross-front-end transfer tests of its residual-only intermediate variant.

2 Related Work

Audio-visual target extraction. Visual cues support speaker-independent separation (Ephrat et al., 2018), enhancement in unconstrained video (Afouras et al., 2018), time-domain separation (Wu et al., 2019), cross-modal consistency (Gao & Grauman, 2021), and universal extraction (Pan et al., 2022). Recent time–frequency approaches include RTFS-Net (Pegg et al., 2024), scenario-aware AV-GridNet (Pan et al., 2023), and AV-CrossNet (Kalkhorani et al., 2025); Dolphin combines lip semantics with global–local modeling (Li et al., 2026). RAVSS uses interactions among speakers to improve separation with missing visual cues (Pan et al., 2024). We instead adapt a deployed extractor to array information through its output; this does not by itself resolve missing-video target selection.

Beamforming and learned spatial separation. Mask-based covariance estimation supports neural beamforming (Souden et al., 2010; Heymann et al., 2017; Yoshioka et al., 2018; Tu et al., 2019). Audio-only SpatialNet learns spatial information within the separator (Quan & Li, 2024), whereas multimodal multichannel extraction also conditions target selection on visual cues (Gu et al., 2020). Transform–average–concatenate (TAC) learns channel pooling that accommodates changing channel order and count (Luo et al., 2020). We use analytic spatial compression before the trainable module. Neither beamforming nor channel symmetry is new; the contribution is their output-level interface to a frozen AV estimate and its residual correction.

Beam-guided neural refinement. Combining separation, beamforming, and learned post-filtering predates this work. Multi-microphone complex spectral mapping uses MVDR outputs as evidence for a second network (Wang et al., 2021); Beam-Guided TasNet iteratively refines separation with beamforming guidance (Chen et al., 2022). Multichannel TF-GridNet likewise places an analytic filter between two networks (Wang et al., 2023). Our distinction is not the cascade itself, but a 70.9K dual-direction correction around a frozen AV output, using complementary spatial views and an interface whose dimension does not depend on microphone count.

Output-level versus internal adaptation. Adapters and LoRA introduce trainable modules or low-rank updates inside a frozen backbone (Houlsby et al., 2019; Hu et al., 2022). Side-Tuning instead adds a lightweight network alongside an unchanged base network (Zhang et al., 2020). Thus additive adaptation itself is not new. D-CRPF specializes output-level adaptation to complex speech spectra and analytic spatial evidence, without front-end hidden activations or backpropagation. Each front end receives separately trained adapter weights. We do not claim superiority over matched-information internal PEFT, nor does small added parameter count imply lower complete-system latency or memory.

3 METHOD

3.1 FRAMEWORK

Given an M-microphone mixture and a synchronized mouth sequence $\mathbf{V}$, we estimate target speech s at reference microphone r under the evaluation reference convention. Let $\mathbf{Y}_{ft} \in \mathbb{C}^M$ denote the mixture short-time Fourier transform (STFT), indexed by frequency f and frame t; $y^{(r)}$ and $Y^{(r)}$ denote the reference waveform and spectrum.

The framework separates target selection from spatial refinement (Figure 1). A frozen single-channel AV extractor first identifies the target using $y^{(r)}$ and $\mathbf{V}$. Its estimate then guides two analytic beamformers, producing target- and interference-oriented spatial views. Finally, the proposed dual-direction adapter, D-CRPF, predicts complex reference and signed beam-disagreement corrections added to the frozen estimate, which remains an explicit output skip.

Only the adapter is trained. Figures 2 and 3 illustrate spatial evidence construction and residual adaptation, respectively. Visual conditioning reaches the spatial module through the frozen estimate and its proxy; the adapter has no separate video input.

3.2 FROZEN AV EXTRACTOR

Given the reference waveform and target video, an existing AV extractor E_θ produces

$$\widehat{s}^{(1)} = E_\theta(y^{(r)}, \mathbf{V}), \qquad \widehat{S}^{(1)} = \mathrm{STFT}(\widehat{s}^{(1)}). \tag{1}$$

Here $\widehat{s}^{(1)}$ is the deployed, length- and scale-aligned estimate. The paired unadapted baseline uses the same alignment, preventing calibration changes from being counted as adaptation gains. The extractor receives no other array channels.

Throughout adaptation, θ and normalization statistics remain fixed, the extractor stays in evaluation mode, and no gradient passes through its audio or visual network. The deployed extractor is thus the fixed operating point.

Using the estimated signal rather than hidden embeddings lets waveform and complex-spectrum extractors share the downstream interface after alignment to the common STFT. Each front end has separately trained adapter weights; a shared architecture does not imply shared weights. The estimate conditions beamforming and anchors final reconstruction.

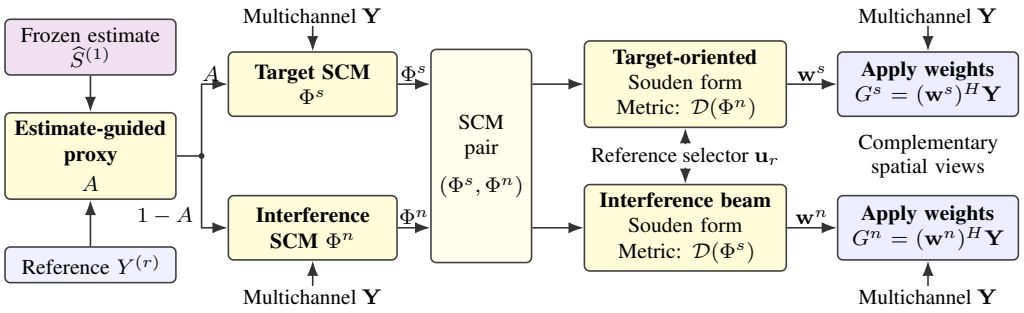

Figure 2: Frozen-estimate-guided spatial evidence. The proxy A and its complement weight the same multichannel observations to form two spatial covariance matrices. Both MVDR branches use the covariance pair, with exchanged roles and a common reference selector. Applying their weights yields complementary complex views G^s and G^n, without trainable parameters.

3.3 Multi-Channel Beamforming

Estimate-guided proxy masks. As illustrated in Figure 2, we convert the frozen estimate into a deterministic Wiener-style proxy:

$$A_{ft} = \mathrm{clip}\left(\frac{|\widehat{S}^{(1)}_{ft}|^2}{|\widehat{S}^{(1)}_{ft}|^2 + |Y^{(r)}_{ft} - \widehat{S}^{(1)}_{ft}|^2 + \epsilon_A}, 10^{-3}, 1 - 10^{-3}\right). \tag{2}$$

The legacy mask-predicting AV-DPCRN instead supplies A directly, with $\widehat{S}^{(1)} = AY^{(r)}$; Appendix B records the protocol for each front end. The residual $Y^{(r)} - \widehat{S}^{(1)}$ can contain both interference and incorrectly removed target speech. Thus A is a proxy for constructing spatial evidence, not an oracle speech-presence probability or a calibrated confidence score. Its complement $1 - A$ defines the interference-oriented view.

Complementary spatial views. The two masks weight the same multichannel observations to estimate spatial covariance matrices:

$$\Phi^k_f = \frac{\sum_t a^k_{ft}\mathbf{Y}_{ft}\mathbf{Y}^H_{ft}}{\max(\sum_t a^k_{ft}, \epsilon_C)}, \qquad a^s = A, \quad a^n = 1 - A. \tag{3}$$

With the reference selector $\mathbf{u}_r$, a regularized Souden MVDR construction (Souden et al., 2010) gives

$$K^s_f = \mathcal{D}(\Phi^n_f)^{-1}\Phi^s_f, \qquad \mathbf{w}^s_f = \frac{K^s_f\mathbf{u}_r}{h_\epsilon(\mathrm{tr}\, K^s_f)}, \qquad G^s_{ft} = (\mathbf{w}^s_f)^H\mathbf{Y}_{ft}. \tag{4}$$

Here $\mathcal{D}$ adds trace-scaled diagonal loading, and h_ϵ safeguards near-zero denominators. We solve the linear system without explicitly forming an inverse. Swapping Φ^s and Φ^n, with the same reference selector, yields the complementary beam G^n. Appendix B specifies the stabilization constants and covariance symmetrization.

With multirank reverberant sample covariances, this trace-normalized construction does not generally enforce a distortionless response to a particular steering vector. The complex beams retain magnitude and phase but are not guaranteed clean, mutually orthogonal, or better than Stage 1. They serve as correction evidence, not final outputs. This analytic module has no trainable parameters.

Array interface and channel relabeling. Beamforming compresses M channels into two spectra, keeping the neural interface independent of M. For a microphone permutation P satisfying $P\mathbf{u}_r = \mathbf{u}_r$, the reference signal, frozen estimate, and masks remain unchanged. Covariances transform as $\Phi' = P\Phi P^\top$; the regularized solve yields $\mathbf{w}' = P\mathbf{w}$, hence $G' = G$. Both views and the deterministic adapter output are unchanged in exact arithmetic (proof in Appendix A). This is invariance to channel *relabeling*, not microphone removal or physical array changes. Spatial computation still grows with M; geometry robustness must be measured.

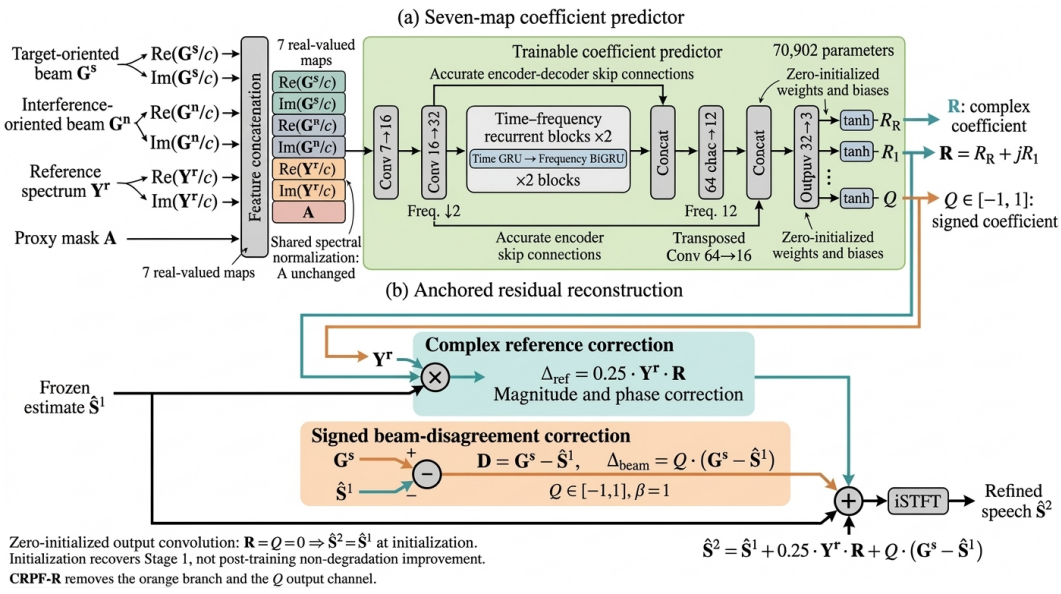

Figure 3: The final D-CRPF spatial residual adapter. Seven real-valued maps condition a three-output coefficient predictor. The frozen estimate remains the output anchor, augmented by both complex reference and signed beam-disagreement corrections. The 70.9K parameters count the complete adapter, including its zero-initialized output convolution; zero correction recovers Stage 1 at initialization.

3.4 Spatial Residual Adapter

Fixed-dimensional spectral input. Figure 3 summarizes the adapter and reconstruction paths. Let F and T be the numbers of frequency bins and frames, and let $c = \max((FT)^{-1/2}\|Y^{(r)}\|_F, 10^{-6})$ be the reference spectral RMS. We construct seven real-valued input maps:

$$\mathbf{Z} = \left[\Re(G^s/c), \Im(G^s/c), \Re(G^n/c), \Im(G^n/c), \Re(Y^{(r)}/c), \Im(Y^{(r)}/c), A\right]. \tag{5}$$

The six spectral maps share scale c; the dimensionless mask A is not rescaled. In this default design, Stage 1 conditions the predictor through A and the beams, with its complex spectrum explicit in output reconstruction. Table 2 separately tests adding its real/imaginary input maps. Neither interface uses microphone coordinates, front-end embeddings, or separate video features.

Compact encoder–decoder. The final D-CRPF predictor P_ϕ uses two convolutional encoders with 16 and 32 channels, two dual-path recurrent blocks, and skip-connected transposed-convolution decoding. A single output convolution jointly predicts three components for the two correction directions: real and imaginary reference-residual coefficients and a signed real beam-disagreement coefficient. Writing the corresponding logits as $P_\phi^{\Re}$, $P_\phi^{\Im}$, and P_ϕ^{q}, we obtain

$$R = \tanh(P_\phi^{\Re}(\mathbf{Z})) + j\tanh(P_\phi^{\Im}(\mathbf{Z})), \qquad Q = \tanh(P_\phi^{q}(\mathbf{Z})). \tag{6}$$

The complete D-CRPF adapter has 70.9K parameters; detailed layer settings are in Appendix B. Its intermediate residual-only variant, CRPF-R, removes the signed output head (0.481K parameters) and is trained independently with the same seven-map interface, totaling 70.4K parameters with $Q = 0$ by construction. This variant is not an inference-time head switch on a trained D-CRPF checkpoint.

Two correction directions. D-CRPF reconstructs the final spectrum around the frozen estimate using both correction directions:

$$\widehat{S}_{ft}^{(2)} = \widehat{S}_{ft}^{(1)} + \underbrace{\alpha Y_{ft}^{(r)} R_{ft}}_{\text{complex reference correction}} + \underbrace{\beta Q_{ft}(G_{ft}^s - \widehat{S}_{ft}^{(1)})}_{\text{beam-disagreement correction}}, \quad \alpha = 0.25, \quad \beta = 1. \tag{7}$$

The first path can correct magnitude and phase relative to the reference observation, consistent with the motivation for complex ratio masking (Williamson et al., 2016). Unlike a real-only multiplier, complex R also permits a quadrature correction; its scale is tied to $|Y^{(r)}|$. Here it predicts an additive correction, not the complete target mask. The second path follows the disagreement between the target beam and Stage 1. It supplies a spatial correction direction even where cancellation makes the reference spectrum weak but the target beam remains informative.

The coefficient $Q \in [-1, 1]$ is signed, not calibrated confidence. Without the first correction, the output is $(1 - \beta Q)\widehat{S}^{(1)} + \beta Q G^s$; negative Q permits extrapolation, so the decoder is not restricted to convex averaging. The coefficients R and Q are dimensionless. Equation 7 uses physical spectral units; implementation divides spectra by c before synthesis and restores that scale afterwards. The output waveform is obtained with inverse STFT at the original length.

Learning around the frozen operating point. All active output-head weights and biases are initialized to zero. Hence $R = Q = 0$ and $\widehat{S}^{(2)} = \widehat{S}^{(1)}$ initially; waveform equality also requires consistent analysis/synthesis, length, and inverse scaling, up to floating-point error. Zero correction remains in the adapter's function class, but this does not guarantee convergence to a better solution or non-degradation on every test example.

We optimize only ϕ using

$$\mathcal{L} = -\,\text{SI-SDR}(\hat{s}^{(2)}, s) + 0.1\mathcal{L}_{\text{SNR}} + 0.1\mathcal{L}_{\text{spec}}. \tag{8}$$

The SI-SDR term evaluates extraction without fixing gain, while the negative-SNR loss and log-magnitude spectral loss penalize scale-sensitive and spectral discrepancies. The spectral loss compares $\log(1+|\widehat{S}^{(2)}/c|)$ with $\log(1+|S/c|)$, where $S = \text{STFT}(s)$. Training details are given in Appendix B.

Correction bounds and interpretation. Componentwise $\tanh$ implies $|R| \le \sqrt{2}$ and $|Q| \le 1$. For $\Delta = \widehat{S}^{(2)} - \widehat{S}^{(1)}$, the triangle inequality gives

$$|\Delta_{ft}| \le \sqrt{2}|\alpha|\,|Y^{(r)}_{ft}| + |\beta|\,|G^s_{ft} - \widehat{S}^{(1)}_{ft}|. \tag{9}$$

For the final D-CRPF, the bound depends on both the reference magnitude and beam disagreement: an inaccurate beam can still induce a large correction. This is an input-dependent bound, not a uniform stability guarantee. For the intermediate CRPF-R variant, it reduces to $0.354|Y^{(r)}_{ft}|$.

To interpret the role of spatial evidence, define the baseline error $e = s - \hat{s}^{(1)}$, correction $\delta = \hat{s}^{(2)} - \hat{s}^{(1)}$, and squared-error risks $\mathcal{R}_i = \mathbb{E}\|s - \hat{s}^{(i)}\|_2^2$. Under finite second moments,

$$\mathcal{R}_2 - \mathcal{R}_1 = \mathbb{E}\|\delta\|_2^2 - 2\mathbb{E}\langle e, \delta\rangle. \tag{10}$$

A useful correction must align with the original error strongly enough to offset its own energy. Spatial views are intended to supply information about that error; extra parameters alone do not establish such information. This motivates the separately trained single-channel capacity control. The identity is explanatory, not a guarantee for the mixed training objective, SI-SDR, perceptual scores, or WER. Proofs and assumptions are collected in Appendix A.

Incorrect visual target selection can corrupt the anchor and proxy, while similar source transfer functions can make spatial views insufficiently discriminative. Neither identity initialization nor numerical stabilization resolves these failures; missing video and adverse geometry are evaluated explicitly.

4 Experimental Setup

Data and complete-utterance evaluation. We simulate six-microphone reverberant mixtures from LRS2 (Chung et al., 2017). The matched training pool has 19,624 two-speaker scenes (10.90 h), represented as 39,248 target-speaker views; validation has 476 scenes and 952 views. Training and the primary test use the same circular six-microphone configuration and scene-generation conventions. Main-table adapter runs use the full training loader without a 500-batch cap. The test uses all 1,243 unique official target utterances at 16 kHz, lasting 0.640–5.824 s (0.577 h total). Each is

Table 1: Intermediate CRPF-R improves released front ends on 1,243 full-duration simulated LRS2 targets; only the 70.4K adapter is trained. The AV-MossFormer2 citation dates its audio backbone, not the AV release. Arrows indicate preferred metric directions.

Front end	Spatial adaptation	SI-SDR↑ (dB)	PESQ↑	STOI↑	WER↓ (%)
Dolphin (Li et al., 2026)	frozen	11.219	2.257	.883	17.55
Dolphin	frozen + CRPF-R	12.123	2.633	.904	13.19
AV-MossFormer2 (Zhao et al., 2024)	frozen	12.868	2.487	.914	11.61
AV-MossFormer2	frozen + CRPF-R	14.238	2.930	.937	8.35
IIANet (Li et al., 2024)	frozen	11.771	2.252	.892	14.56
IIANet	frozen + CRPF-R	13.274	2.783	.926	9.27

paired with different-speaker interference covering its complete duration and SIR sampled from $\{-10, -5, 0, 5, 10\}$ dB. The first 50 ms of the target RIR defines the reference. Inference windowing covers the full utterance; it is not truncated-duration scoring.

Two comparison regimes. *Frozen public transfer* trains intermediate CRPF-R adapters on released Dolphin, AV-MossFormer2, and IIANet without changing their weights. *Matched comparison* develops final D-CRPF on AV-CrossNet trained on the simulation, with the frozen Stage 1 and separately trained single-channel and CRPF-R controls. Matched IIANet uses intermediate CRPF-R after front-end fine-tuning; AV-SpatialNet is trained directly on multichannel input. Adapter weights are not shared across front ends. All adapters use validation-SI-SDR selection and the same frozen checkpoint as their paired baseline. We report one run per configuration, not seed averages. Earlier fixed-budget AV-DPCRN (Le et al., 2021) experiments remain a separate mechanism study.

Metrics and provenance. We report SI-SDR (Le Roux et al., 2019), wide-band PESQ (Rix et al., 2001), STOI (Taal et al., 2011), and corpus WER from Whisper large-v3 (Radford et al., 2023). These objective scores do not establish human-perceived quality. Completed main-table quality rows cover all 1,243 outputs and 6,950 normalized reference words. Differences use unrounded scores or word-error counts; table entries are rounded independently. Clean-target WER is 4.98%, an ASR reference rather than a guaranteed lower bound. Appendix B specifies inference, selection, and scoring. Main target-extraction comparisons use no oracle assignment.

5 Results

5.1 Spatial adaptation of released AV extractors

Table 1 tests intermediate CRPF-R on released Dolphin (Li et al., 2026), AV-MossFormer2 (Zhao et al., 2024), and IIANet (Li et al., 2024) without front-end retraining. Training only a 70.4K adapter per model yields 0.904–1.502 dB SI-SDR gains and 3.27–5.29-point WER reductions. PESQ and STOI also improve. These results support transfer of the shared spatial interface, not cross-front-end validation of full D-CRPF or uniform perceptual improvement.

5.2 From residual-only adaptation to D-CRPF

To improve the residual-only design, we develop final D-CRPF on AV-CrossNet (Kalkhorani et al., 2025), whose public implementation makes matched-data retraining and controlled comparisons practical. Table 2 freezes one matched checkpoint across independently trained variants. The 0.481K-parameter signed beam-disagreement head adds 0.225 dB SI-SDR and reduces WER by 0.50 points relative to CRPF-R. This single-run increment does not establish seed-stable superiority. Section 5.3 analyzes capacity and input controls.

Matched comparison with recent architectures. Table 3 compares final D-CRPF with matched IIANet (Li et al., 2024) and AV-SpatialNet, our visual extension of SpatialNet (Quan & Li, 2024). D-CRPF improves AV-CrossNet by 1.254 dB SI-SDR and 2.71 WER points; intermediate CRPF-R

Table 2: From CRPF-R to final D-CRPF on AV-CrossNet (Kalkhorani et al., 2025), with capacity and input controls. All rows use 1,243 full-duration targets, frozen Stage 1 epoch 29, full-loader 30-epoch adapter training, and validation selection. Nine maps add the Stage 1 complex spectrum. Parameters exclude the frozen front end; K denotes 10^3.

Adapter	Added params↓ (K)	SI-SDR↑ (dB)	PESQ↑	STOI↑	WER↓ (%)
None	0	13.179	2.661	.924	10.85
Single-channel control	70.9	13.224	2.674	.923	10.42
CRPF-R	70.4	14.208	3.015	.941	8.65
D-CRPF (7 maps)	70.9	14.433	3.091	.944	8.14
D-CRPF (9 maps)	71.4	14.527	3.069	.944	7.91

Table 3: Matched comparison of final D-CRPF and recent architectures, with intermediate CRPF-R on matched IIANet. AV-SpatialNet is our visual extension of the cited SpatialNet backbone. All scores use full-duration targets without oracle assignment; each adapted row shares its paired Stage 1 checkpoint.

System	Input / adaptation	SI-SDR↑ (dB)	PESQ↑	STOI↑	WER↓ (%)
AV-CrossNet (Kalkhorani et al., 2025)	reference channel	13.179	2.661	.924	10.85
AV-CrossNet + D-CRPF	frozen Stage 1, six channels	**14.433**	**3.091**	**.944**	**8.14**
AV-SpatialNet (Quan & Li, 2024)	direct six-channel	13.790	3.048	.917	14.39
IIANet (Li et al., 2024)	reference channel	12.686	2.333	.909	12.95
IIANet + CRPF-R	frozen Stage 1, six channels	14.247	2.898	.940	8.33

improves matched IIANet by 1.561 dB and 4.62 WER points. Recent architectures thus retain spatial adaptation gains after matched training, with the two variants' evidence kept distinct.

AV-CrossNet+D-CRPF gains 0.643 dB SI-SDR and 6.24 WER points over AV-SpatialNet. This margin is checkpoint-sensitive: historical epoch 19 reaches 14.464 dB SI-SDR and 3.092 PESQ, essentially matching 14.433 dB and 3.091 after adaptation, while WER is 10.79% versus 8.14%. We retain validation rather than retrospective test selection. Within-front-end gains are the primary evidence, not universal multichannel superiority.

5.3 Ablation Study

Spatial evidence versus added capacity. Table 2 holds frozen AV-CrossNet epoch 29 and the full-loader budget fixed. The single-channel control retains 70.9K parameters and three active heads, replacing the beams by $AY^{(r)}$ and $(1-A)Y^{(r)}$ and the signed direction by $AY^{(r)} - \widehat{S}^{(1)}$. It never accesses other channels. It gains 0.045 dB over Stage 1, versus 1.029 and 1.254 dB for spatial CRPF-R and D-CRPF. The 1.209-dB margin over this equal-parameter control supports a spatial benefit beyond added capacity. Legacy AV-DPCRN (Le et al., 2021) controls on intermediate CRPF-R test beam removal, real-only residuals, and random initialization (Appendix Table 6); duration strata appear in Appendix C.8.

Exposing the frozen complex estimate. Appending $\Re(\widehat{S}^{(1)}/c)$ and $\Im(\widehat{S}^{(1)}/c)$ adds 0.480K parameters. With unchanged training and decoding, nine maps improve SI-SDR by 0.095 dB and WER by 0.23 points over seven maps, but reduce PESQ by 0.022; STOI is nearly unchanged. This is a single-run trade-off, not uniform or statistically established superiority. Seven maps remain the default (Appendix B.3).

Array dependence. Permutation invariance does not imply physical array robustness. Our four-array suite fixes checkpoints, speech, source gains, and non-array scene parameters (Table 4). Circular and linear 8 cm also share the exact reference waveform. This topology change reduces SI-SDR by

Table 4: Controlled array comparison (SI-SDR, dB) on 1,243 full-duration targets per cell, with shared non-array scene factors and fixed checkpoints. These jointly re-rendered mixtures differ from Table 3; sizes denote circle diameter or line length. Stage 1 and adapted outputs are from the same export. Full metrics: Appendix Table 12.

	SI-SDR (dB)↑			
System	Circular 4 cm	Circular 8 cm	Circular 12 cm	Linear 8 cm
SpatialNet (oracle assignment)	**14.347**	**14.778**	**14.864**	9.954
AV-SpatialNet	13.367	13.787	13.905	9.724
AV-CrossNet Stage 1	13.058	13.171	13.317	13.171
AV-CrossNet + D-CRPF	14.088	14.418	14.746	**13.612**

0.806 dB for AV-CrossNet+D-CRPF, versus 4.063 dB for AV-SpatialNet and 4.824 dB for oracle-assigned SpatialNet. On the line, the proposed system reaches 13.612 dB and 8.85% WER, versus 9.954 dB and 14.99% for that audio-only oracle reference. No geometry-specific retraining is used.

Stage 1 outputs are sample-identical for circular and linear 8 cm, scoring 13.171 dB. Adaptation gain thus falls from 1.246 to 0.441 dB: positive on the line, but only 35.4% of the circular-array improvement measured in dB. The fixed front end anchors the system without making correction geometry-invariant or guaranteeing a performance floor. SpatialNet leads circular-array SI-SDR; aperture changes also move the reference sensor (Appendix C.6).

Failure boundaries. Legacy D-CRPF is 0.314 dB worse than Stage 1 at exact collinearity, but 0.892 dB better at 2°. Paired visual tests retain 2.470–2.634 dB gains with normal, low-resolution, or partially missing video. With complete visual loss, Stage 1 falls to -3.061 dB and adaptation further lowers it to -3.404 dB. In SI-SDR, failure is therefore inherited and amplified. Appendix C separates these legacy boundaries from strong-front-end results; no recovery policy is validated.

6 Conclusion

We proposed D-CRPF, a 70.9K-parameter dual-direction spatial residual adapter for frozen AV extractors. Full-duration matched AV-CrossNet experiments support paired gains, spatial benefit beyond equal-capacity single-channel processing, and positive but attenuated gains under array shift. The residual-only intermediate CRPF-R supplies complementary transfer evidence across released and matched front ends. Together, these results support output-level array adaptation while preserving the deployed extractor, with explicit limits under array and visual changes.

Reproducibility Statement

Section 3 and Appendix B specify the method, numerical conventions, inference, selection, and metrics. Paired comparisons share frozen checkpoints. Audits verify unique targets, full-length outputs, metric coverage, and artifact hashes, including all sixteen corrected-array rows. Configurations, manifests, per-utterance scores, and permitted examples will accompany the anonymous reproducibility package.

AI Use Statement

In preparing this manuscript, the authors used a large language model (LLM) solely to assist with polishing the wording, improving the clarity of English expression and finding related works. The LLM did not contribute to the research design, data collection, data analysis, or the generation of scientific ideas.

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Interspeech*, pp. 3244–3248, 2018.

Hangting Chen, Yi Yang, Feng Dang, and Pengyuan Zhang. Beam-Guided TasNet: An iterative speech separation framework with multi-channel output. In *Interspeech*, pp. 866–870, 2022. doi: 10.21437/Interspeech.2022-230.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, pp. 3444–3453, 2017.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):112:1–112:11, 2018. doi: 10.1145/3197517.3201357.

Ruohan Gao and Kristen Grauman. VisualVoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, pp. 15495–15505, 2021.

Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Lianwu Chen, Yuexian Zou, and Dong Yu. Multi-modal multi-channel target speech separation. *IEEE Journal of Selected Topics in Signal Processing*, 14 (3):530–541, 2020.

Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach. Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system. In *ICASSP*, pp. 5325–5329, 2017.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, pp. 2790–2799, 2019.

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

Vahid Ahmadi Kalkhorani, Cheng Yu, Anurag Kumar, Ke Tan, Buye Xu, and DeLiang Wang. 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*, 19(4): 685–694, 2025. doi: 10.1109/JSTSP.2025.3567838.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Xiaohuai Le, Hongsheng Chen, Kai Chen, and Jing Lu. DPCRN: Dual-path convolution recurrent network for single channel speech enhancement. In *Interspeech*, pp. 2811–2815, 2021.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR–half-baked or well done? In *ICASSP*, pp. 626–630, 2019.

Kai Li, Runxuan Yang, Fuchun Sun, and Xiaolin Hu. IIANet: An intra- and inter-modality attention network for audio-visual speech separation. In *International Conference on Machine Learning*, pp. 29181–29200, 2024.

Kai Li, Kejun Gao, and Xiaolin Hu. Efficient audio-visual speech separation with discrete lip semantics and multi-scale global-local attention. In *International Conference on Learning Representations*, 2026.

Yi Luo, Zhuo Chen, Nima Mesgarani, and Takuya Yoshioka. End-to-end microphone permutation and number invariant multi-channel speech separation. In *ICASSP*, 2020.

Tianrui Pan, Jie Liu, Bohan Wang, Jie Tang, and Gangshan Wu. RAVSS: Robust audio-visual speech separation in multi-speaker scenarios with missing visual cues. In *ACM Multimedia*, pp. 4748–4756, 2024. doi: 10.1145/3664647.3681261.

Zexu Pan, Meng Ge, and Haizhou Li. USEV: Universal speaker extraction with visual cue. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 30, pp. 3032–3045, 2022.

Zexu Pan, Gordon Wichern, Yoshiki Masuyama, François G. Germain, Sameer Khurana, Chiori Hori, and Jonathan Le Roux. Scenario-aware audio-visual TF-GridNet for target speech extraction. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2023. doi: 10.1109/ASRU57964.2023.10389618.

Samuel Pegg, Kai Li, and Xiaolin Hu. RTFS-Net: Recurrent time-frequency modelling for efficient audio-visual speech separation. In *ICLR*, 2024.

Changsheng Quan and Xiaofei Li. SpatialNet: Extensively learning spatial information for multi-channel joint speech separation, denoising and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1310–1323, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, pp. 28492–28518, 2023.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, volume 2, pp. 749–752, 2001.

Mehrez Souden, Jacob Benesty, and Sofiène Affes. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18 (2):260–276, 2010. doi: 10.1109/TASL.2009.2025790.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

Yan-Hui Tu, Jun Du, Lei Sun, Feng Ma, Hai-Kun Wang, Jing-Dong Chen, and Chin-Hui Lee. An iterative mask estimation approach to deep learning based multi-channel speech recognition. In *Speech Communication*, volume 106, pp. 31–43, 2019.

Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2001–2014, 2021. doi: 10.1109/TASLP.2021.3083405.

Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. TF-GridNet: Integrating full- and sub-band modeling for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3221–3236, 2023. doi: 10.1109/TASLP.2023.3304482.

Donald S. Williamson, Yuxuan Wang, and DeLiang Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3): 483–492, 2016. doi: 10.1109/TASLP.2015.2512042.

Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu. Time domain audio visual speech separation. In *ASRU*, pp. 667–673, 2019.

Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva. Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks. In *Interspeech*, pp. 3038–3042, 2018.

Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: A baseline for network adaptation via additive side networks. In *ECCV*, pp. 698–714, 2020. doi: 10.1007/978-3-030-58580-8_41.

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. MossFormer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP*, 2024.

A STRUCTURAL PROPERTIES AND PROOFS

These are algebraic properties of the implemented adapter, not new beamforming optimality results. We assume finite inputs, deterministic evaluation with a fixed front end, and well-defined stabilized spatial operations. The features use a common reference spectral-RMS scale. All spectra in the following proofs are in physical units, as in Eq. 7. Let $\mathcal{B}_N$ denote the real-linear iSTFT operator with the specified window, hop, boundary convention, and output length N. It includes no amplitude rescaling. The implementation's normalized synthesis is equivalent because $c\,\mathcal{B}_N(S/c) = \mathcal{B}_N(S)$ for the positive, fixed input scale c.

A.1 IDENTITY INITIALIZATION AND BASELINE INCLUSION

Write the output-head logits as $U = W_o * H_\psi(\mathbf{Z}) + b_o$, where H_ψ contains the preceding adapter layers. Setting $W_o = b_o = 0$ gives $U = 0$ for every finite hidden representation. Since $\tanh(0) = 0$, the two complex-residual components vanish, as does the third, gate component when present. Substitution into Eq. 7 gives

$$\widehat{S}^{(2)} = \widehat{S}^{(1)}, \qquad \mathcal{B}_N(\widehat{S}^{(2)}) = \mathcal{B}_N(\widehat{S}^{(1)}).$$

Thus the adapter hypothesis class contains the deployed baseline. This baseline includes its specified amplitude calibration and synthesis convention; equality to an original waveform additionally requires a consistent STFT/iSTFT pair and the same length. Floating-point reconstruction need not be bitwise identical.

Consequently, for any well-defined output-based risk $\mathcal{R}$, the zero-head parameter setting ϕ_0 satisfies $\inf_\phi \mathcal{R}(\phi) \le \mathcal{R}(\phi_0) = \mathcal{R}(\text{baseline})$; the same statement holds for empirical risk. This is a containment statement, not a guarantee about the trained parameters, optimization convergence, or test performance. Initialization from an already trained, nonzero adapter does not have the zero-head identity property.

A.2 INPUT-DEPENDENT CORRECTION BOUND AND ERROR DECOMPOSITION

Let $\Delta = \widehat{S}^{(2)} - \widehat{S}^{(1)}$. For the tanh parameterization, $|R_{ft}|^2 = (\Re R_{ft})^2 + (\Im R_{ft})^2 \le 2$ and $|Q_{ft}| \le 1$. The triangle inequality therefore proves Eq. 9:

$$|\Delta_{ft}| \le \sqrt{2}|\alpha|\,|Y^{(r)}_{ft}| + |\beta|\,|G^s_{ft} - \widehat{S}^{(1)}_{ft}|.$$

With $\delta = \mathcal{B}_N(\Delta)$ and $C_N = \|\mathcal{B}_N\|_{\text{op}}$, linearity additionally yields

$$\|\delta\|_2 \le C_N\Big(\sqrt{2}|\alpha|\,\|Y^{(r)}\|_F + |\beta|\,\|G^s - \widehat{S}^{(1)}\|_F\Big).$$

The gate term is absent for the residual-only configuration. For a fixed length and synthesis convention, C_N depends on the operator, not the input amplitude. The bound controls correction magnitude, not target error, and still depends on the input spectrum and beam discrepancy. The signed gate is neither a convex interpolation nor a calibrated confidence score.

For a real target waveform s, define $e = s - \mathcal{B}_N(\widehat{S}^{(1)})$. Assuming finite second moments, expanding $\|e - \delta\|_2^2$ gives the squared-error risk identity

$$\mathbb{E}\|e - \delta\|_2^2 - \mathbb{E}\|e\|_2^2 = \mathbb{E}\|\delta\|_2^2 - 2\mathbb{E}\langle e, \delta\rangle,$$

as in Eq. 10. Improvement requires sufficient alignment with the baseline error to offset correction energy. This explains the purpose of residual evidence; it does not establish improvement for the mixed training loss, SI-SDR, perceptual scores, or recognition error.

A.3 REFERENCE-PRESERVING MICROPHONE PERMUTATIONS

Let P be a microphone permutation matrix satisfying $P\mathbf{u}_r = \mathbf{u}_r$, and replace every observation by $\mathbf{Y}'_{ft} = P\mathbf{Y}_{ft}$. The reference waveform, frozen estimate, proxy mask, and normalization scalar are unchanged. Hence each masked covariance, including its scalar denominator safeguard, transforms

as $\Phi' = P\Phi P^\top$. Trace-scaled loading has the form $\mathcal{D}(\Phi) = \Phi + \lambda(\Re\,\mathrm{tr}(\Phi)/M)I$, with $\lambda(z) = 10^{-3}\max(z, 10^{-6}) + 10^{-6}$ before and after permutation. Trace invariance and $PIP^\top = I$ imply

$$\mathcal{D}(P\Phi P^\top) = P\mathcal{D}(\Phi)P^\top, \qquad K = \mathcal{D}(\Phi^n)^{-1}\Phi^s,$$
$$K' = \mathcal{D}(\Phi^{n\prime})^{-1}\Phi^{s\prime} = PKP^\top.$$

The trace, retained in complex dtype by the implementation, and its scalar safeguard $h_\epsilon(\mathrm{tr}\,K)$ are unchanged: $h_\epsilon(z) = z$ when $|z| \geq \epsilon$, and $h_\epsilon(z) = \epsilon$ otherwise. Since $P^\top\mathbf{u}_r = \mathbf{u}_r$, also $K'\mathbf{u}_r = PK\mathbf{u}_r$. Thus $\mathbf{w}' = P\mathbf{w}$ and $(\mathbf{w}')^H\mathbf{Y}' = \mathbf{w}^H\mathbf{Y}$. Swapping target and noise covariances gives the same result for the complementary beam. All seven feature maps, and therefore the adapter output, are unchanged in exact arithmetic; finite-precision reductions can introduce small numerical differences.

The seven-map neural interface also has a parameter count independent of M. This does not make predictions invariant to adding or removing microphones: the covariances change, and their construction and dense solves cost $O(FTM^2)$ and $O(FM^3)$, respectively. Changing physical geometry is likewise not a permutation of unchanged recordings. Array robustness remains an empirical question, not a consequence of this symmetry.

B Implementation, evaluation, and provenance

B.1 Numerical spatial interface and adapter architecture

The adapter uses 16-kHz waveforms, a 512-point periodic Hann window, a 512-point FFT, and a 128-sample hop. Its STFT is centered with reflection padding, uses no FFT normalization, and retains the 257-bin one-sided complex spectrum; the matching iSTFT is given the original waveform length. Front ends retain their own analysis/synthesis conventions; in particular, the direct spatial baseline uses a 256-point FFT with a 128-sample hop. The proxy in Eq. 2 uses $\epsilon_A = 10^{-8}$. Covariance denominators use $\epsilon_C = 10^{-6}$, and estimated matrices are Hermitian-symmetrized. The diagonal-loading operator in Eq. 4 is

$$\mathcal{D}(C) = C + \left[10^{-3}\max\left(\frac{\Re\,\mathrm{tr}\,C}{M}, 10^{-6}\right) + 10^{-6}\right] I. \tag{11}$$

Only the matrix being inverted in each beam is loaded. The trace safeguard is $h_\epsilon(z) = \epsilon$ for $|z| < \epsilon$ and z otherwise, with $\epsilon = 10^{-6}$. For positive-definite $H = \mathcal{D}(\Phi^n)$ and positive-semidefinite Φ^s, exact arithmetic gives $\mathrm{tr}(H^{-1}\Phi^s) = \mathrm{tr}(H^{-1/2}\Phi^s H^{-1/2}) \geq 0$, a real number. The implementation retains the computed complex trace, including possible roundoff-level imaginary components. Both beams use the same reference selector. These safeguards improve numerical conditioning but do not correct an erroneous target mask.

Table 5 specifies final D-CRPF and its residual-only intermediate CRPF-R. Tensors have axes $[B, C, T, F]$; convolution kernels and strides are ordered (T, F). Both encoder layers and the transposed convolution use batch normalization and a scalar PReLU. Each recurrent block applies a single-layer, unidirectional time GRU with hidden size 32 to $[BF', T, 32]$, then a single-layer, bidirectional frequency GRU with hidden size 32 per direction to $[BT, F', 32]$, where $F' = 129$. Linear projections map their outputs from 32 and 64 channels to 32, respectively; each path is added residually. Before each GRU, cumulative layer normalization is applied to $[BF', 32, T]$, accumulating over channels and time up to the current frame; the frequency path then reshapes the normalized tensor. The recurrent blocks do not segment time into DPRNN chunks.

All convolutions use padding $(1, 2)$; decoder tensors are cropped to the encoder skip's spatial size before concatenation. External encoder–decoder skips concatenate channels, whereas recurrent paths use addition. The extra signed output channel costs $(32 \cdot 3 \cdot 5 + 1)/10^3 = 0.481$K parameters. Seven maps are the default, not a claim that omitting the Stage 1 complex spectrum is optimal. The evaluated nine-map D-CRPF appends its normalized real and imaginary parts, increasing only Encoder 1 by $(2 \cdot 16 \cdot 3 \cdot 5)/10^3 = 0.480$K parameters (1.729K to 2.209K), for 71.4K total. No nine-map CRPF-R result is reported.

Front-end mask protocol. Released Dolphin, AV-MossFormer2, and IIANet, as well as matched IIANet and AV-CrossNet, use Eq. 2 computed from their deployed waveform outputs. Legacy AV-DPCRN instead uses its predicted real mask, clipped to $[10^{-3}, 1 - 10^{-3}]$, both for spatial covariance

Table 5: Trainable adapter architecture. Counts are exact in K (10^3), including biases and normalization affine parameters but excluding batch-normalization running statistics. k and s denote kernel and stride. Output heads use componentwise $\tanh$ and zero-initialized weights and biases. The nine-map ablation changes only Encoder 1.

Block	Setting	Parameters (K)↓
Encoder 1	Conv $7 \to 16$, $k = (3, 5)$, $s = (1, 1)$; BN, PReLU	1.729
Encoder 2	Conv $16 \to 32$, $k = (3, 5)$, $s = (1, 2)$; BN, PReLU	7.777
Time paths ($\times 2$)	CumLN(32); GRU $32 \to 32$; linear $32 \to 32$	14.912
Frequency paths ($\times 2$)	CumLN(32); BiGRU $32 \to 32$ per direction; linear $64 \to 32$	29.632
Decoder	Concatenate encoder-2 and recurrent outputs; transpose conv $64 \to 16$, $k = (3, 5)$, $s = (1, 2)$; BN, PReLU	15.409
CRPF-R output	Concatenate encoder-1 and decoder outputs; conv $32 \to 2$, $k = (3, 5)$	0.962
D-CRPF output	Same concatenation; conv $32 \to 3$, $k = (3, 5)$	1.443
CRPF-R total	Shared backbone plus two-component output	70.421
D-CRPF total	Shared backbone plus three-component output	70.902
D-CRPF (9 maps)	Same three-component output; Encoder 1 uses Conv $9 \to 16$	71.382

estimation and for $\widehat{S}^{(1)} = AY^{(r)}$. This native mask is not replaced by a second proxy computed from $AY^{(r)}$; the two constructions are generally unequal. In both cases, A also supplies the seventh adapter input map.

All complex spatial features share the reference spectral-RMS divisor c from Section 3.4; the proxy mask is dimensionless and is not divided by c. Dolphin additionally aligns its Stage 1 waveform RMS to the reference waveform RMS before proxy construction. The paired Dolphin baseline uses that same deployed alignment. It must not be confused with the spectral-RMS normalization internal to the adapter, or with retraining the released model.

B.2 Objective, training budget, and complete inference coverage

For zero-mean target $\bar{s}$ and estimate $\bar{\hat{s}}$, the SI-SDR projection is

$$s_{\text{proj}} = \frac{\langle \bar{\hat{s}}, \bar{s} \rangle}{\|\bar{s}\|_2^2} \bar{s}, \qquad \text{SI-SDR} = 10 \log_{10} \frac{\|s_{\text{proj}}\|_2^2}{\|\bar{\hat{s}} - s_{\text{proj}}\|_2^2}. \tag{12}$$

The SNR loss is negative SNR: $\mathcal{L}_{\text{SNR}} = -10 \log_{10}(\|s\|_2^2 / \|s - \hat{s}\|_2^2)$. The spectral loss is the mean absolute difference between $\log(1 + |\widehat{S}^{(2)}/c|)$ and $\log(1 + |S/c|)$, not a complex-spectrum regression objective. Small denominator and energy floors are used in numerical evaluation. The common amplitude normalization is also applied to the waveform loss inputs. The coefficients in Eq. 8 are 1, 0.1, and 0.1. Main-table residual configurations fix $\alpha = 0.25$ and $\beta = 1$; no systematic scale-sensitivity sweep is reported.

Main adapters use Adam (Kingma & Ba, 2015) and validation-SI-SDR checkpoint selection, for up to 30 epochs. Matched AV-CrossNet's seven- and nine-map D-CRPF, single-channel control, and CRPF-R use learning rate 3×10^{-4}, no weight decay, gradient clipping at 5, learning-rate reduction by a factor of 0.5 after three non-improving validation epochs, and early-stop patience six. They use seed 20260825 and three GPUs with batch size four per GPU, synchronized batch normalization, and BF16 autocast. The complete loader has 3,270 steps per epoch; standard distributed sampling and last-batch dropping retain 39,240 of 39,248 target views per epoch. All four configurations complete 30 epochs. "Full loader" means no additional 500-batch cap, not that an incomplete terminal batch is padded into every epoch. This budget is distinct from the older fixed-budget AV-DPCRN controls.

The strong-front-end single-channel control in Table 2 retains all 70.9K parameters and three active output heads. It replaces the two spatial views by $AY^{(r)}$ and $(1 - A)Y^{(r)}$ and uses $AY^{(r)} - \widehat{S}^{(1)}$ as the signed-correction direction. Only the reference channel reaches STFT, normalization, feature construction, or decoding; the other recorded channels supply no evidence. This is not the legacy A1 beam-zeroing control in Table 6. The strong CRPF-R control instead retains both six-channel MVDR views and removes the signed-correction head, leaving 70.4K parameters. Both new controls

start from zero-initialized output heads without loading an earlier adapter. They share the original training and validation manifests, 2-s training/validation crops, objective, and full-loader budget with D-CRPF; matching the training seed does not imply identical sample-level augmentation draws across differently sized networks.

Dolphin, IIANet, and matched AV-CrossNet cover long utterances with 2-s windows and nominal 0.5-s overlap (1.5-s hop). Each prediction is weighted by a nonperiodic Hann window floored at 10^{-3}, accumulated, and divided samplewise by the accumulated window weights (denominator floor 10^{-6}). This weighted overlap-add does not assume a constant-overlap-add Hann identity. A final full window ending at the last sample is appended if needed, so its overlap can be larger. Utterances of at most 2 s are passed directly to the front end without export-time window padding. Chunk predictions are length-aligned; the Dolphin and AV-CrossNet exporters also zero-pad a returned short prediction before accumulation. Mouth frames are selected proportionally to each audio window's boundaries.

AV-MossFormer2 instead uses 3-s windows with 1.8-s hop and retained-center stitching: internal windows discard 0.6 s at each end, while boundary windows retain the utterance edges. The last window is aligned to the utterance end; utterances of at most 3 s use a single direct call. This is not Hann overlap-add. For all these front ends, proxy construction, covariance estimation, and adapter inference follow reconstruction of the complete Stage 1 waveform. Training/validation covariances use the 2-s crops, whereas test covariances use the full utterance; duration-dependent estimation effects have not been separately isolated. The export audit checks final sample counts against the target reference, rather than inferring full duration from an option. Direct AV-SpatialNet uses its full-duration target decoder. None of these procedures establishes causal streaming behavior.

The official test has 1,243 unique target utterances, totaling 33,223,680 samples at 16 kHz. The RIR pool has 3,241 held-out geometries. One exceptionally short interferer is periodically extended; other interferers supply a crop covering the target. Natural angular bins contain 340/549/354 examples for $\leq 5^\circ$, 5–20°, and $> 20^\circ$. Physical-separation bins contain 338/577/328 examples for ≤ 0.15 m, 0.15–0.35 m, and > 0.35 m. The joint difficult subset contains 233 examples. These observational bins are not controlled geometric interventions.

B.3 Nine-map input ablation

The expanded input is $\mathbf{Z}_9 = [\mathbf{Z}, \Re(\widehat{S}^{(1)}/c), \Im(\widehat{S}^{(1)}/c)]$, with the same reference spectral-RMS divisor c as Eq. 5. Both MVDR views, the proxy, three output heads, $\alpha = 0.25$, $\beta = 1$, objective, and frozen front end are unchanged. Only the first convolution expands. The adapter is trained afresh with zero-initialized output heads, not warm-started from the seven-map checkpoint. Training/validation manifests, 2-s crops, and distributed validation aggregation are unchanged; the same seed policy does not ensure identical augmentation draws after changing network size.

Table 2 compares complete test outputs under the same Stage 1 window reconstruction and full-utterance spatial processing. Relative to seven maps, SI-SDR rises by 0.094547 dB and WER falls from 566 to 550 errors over 6,950 words (0.230216 percentage points): deletions stay at 47, insertions decrease from 289 to 280, and substitutions from 230 to 223. PESQ decreases by 0.021683, and STOI changes from 0.944182 to 0.944078. These are point estimates from one training run per configuration, not a significance claim. The 0.480K-parameter increase (0.68%) is disclosed rather than described as exact capacity matching. Seven maps remain the default; the nine-map model has not been evaluated on the array-shift suite. Existing array and duration-stratified results for AV-CrossNet+D-CRPF continue to use its seven-map configuration.

B.4 Metrics, audits, and statistical scope

Intrusive scores use the designated s_1 reference and exact full-length alignment: PESQ is wide-band and STOI is the standard, non-extended version. Whisper large-v3 uses the same English transcript normalization and decoding protocol for paired comparisons. Full-test corpus WER is the sum of substitutions, deletions, and insertions divided by 6,950 reference words, rather than an unweighted average of utterance WERs. Subset WER uses that subset's reference-word total. For the selected AV-SpatialNet export, 431 substitutions, 87 deletions, and 482 insertions give 14.38849% WER. No subjective listening study is reported; objective score gains should not be interpreted as established listener preference.

Audits check target IDs, sampling rates, exact waveform lengths, and complete metric rows before aggregation. They also distinguish the simulation's full duration from internal inference chunk size. Paired utterance-bootstrap intervals in the legacy analyses use 10,000 replicates; the seed is 20260818 for decoder/ablation analyses and 20260820 for geometry, visual, channel, and array analyses. These intervals quantify variability over test utterances, not variability from repeated training. Main tables report one training run per configuration. In particular, the 0.225-dB SI-SDR difference between matched AV-CrossNet+D-CRPF and +CRPF-R is an observed point-estimate gain in these runs, not evidence of a statistically established or universal advantage of the signed branch.

C Extended controlled experiments

Throughout this appendix, SI-SDR is in dB and WER is in percent. Subset sizes are explicitly indicated and use their own corpus-WER denominators. Differences are computed before rounding, so subtracting displayed entries can differ by 0.01 from a reported difference.

The legacy experiments in this appendix use the earlier AV-DPCRN instantiation with training seed 20260814 and a common, fixed training budget. They study mechanisms and failure boundaries; they are not additional released-checkpoint transfer results, and their training budget is not substituted for the full-loader budget of the main comparisons. Every separately trained legacy ablation uses the same controlled data budget and seed. Evaluation retains complete utterances, with 1,243 examples per full-test condition. Bootstrap intervals describe test-utterance variability, not variability across training seeds. The corrected strong-model array suite in Appendix C.6 and the later independent full-data AV-DPCRN run in Appendix C.7 use different checkpoints and are explicitly separated from this legacy study.

C.1 Full mechanism and decoder ablations

Table 6 analyzes the components leading to final D-CRPF, with CRPF-R as the residual-only intermediate variant. A1–A4 ablate this two-output intermediate decoder: A1 zeros both spatial beams, A2 zeros only the interference beam, A3 replaces zero identity initialization by ordinary random initialization, and A4 forces the imaginary residual to zero. The second block varies the decoder with approximately matched capacity. These are separately trained variants, not inference-time switches applied to a single trained model. The parameter column counts additional adapter parameters; the frozen Stage 1 itself does not have zero total parameters.

Table 6: Complete AV-DPCRN mechanism and decoder ablations on the official 1,243-utterance, full-duration test set under the earlier fixed training budget. Parameters count the adapter only. Arrows indicate preferred directions.

Variant	Added params↓	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
Stage 1	0	6.57	1.734	.827	21.94
Residual-branch mechanism					
A1: single-channel control	70.4K	6.90	1.767	.833	22.24
A2: no interference beam	70.4K	8.14	2.051	.868	14.63
A3: no identity initialization	70.4K	8.18	2.062	.868	14.40
A4: real-only residual	70.4K	8.01	2.028	.865	14.63
CRPF-R	70.4K	8.34	2.073	.872	14.19
Output parameterization					
Target-beam complex residual	70.4K	8.00	2.063	.871	**11.55**
Target-beam real mask	69.9K	8.57	2.205	.881	11.73
Target-beam gate only	69.9K	8.46	2.156	.872	13.32
D-CRPF	70.9K	**9.21**	**2.258**	**.883**	12.04

The parameter-matched single-channel adapter recovers only 0.32 dB over Stage 1, whereas CRPF-R exceeds A1 by 1.447 dB (paired 95% CI $[1.381, 1.512]$). Thus, 82% of the 1.768 dB CRPF-R gain remains after controlling for capacity in this experiment. The interference beam contributes 0.206 dB ($[0.177, 0.234]$), the imaginary residual contributes 0.327 dB ($[0.302, 0.352]$), and identity initialization contributes 0.162 dB ($[0.140, 0.184]$). The latter's 0.22-point WER change is unresolved

([−0.63, 1.06]), so the result does not establish a uniform per-metric improvement from identity initialization.

Among the decoder choices in this controlled study, D-CRPF gives the strongest SI-SDR, PESQ, and STOI. Its 0.866 dB gain over CRPF-R accompanies a 0.481K-parameter increase. The target-beam complex-residual variant attains the lowest lightweight WER (11.55%); the dual-direction decoder retains a 2.14-point WER reduction over CRPF-R but does not attain the lowest WER. These paired intervals use 10,000 utterance-bootstrap replicates and seed 20260818. They quantify the legacy comparison, not uncertainty for another front end.

The separate strong-front-end test is complete in Table 2: full-loader matched AV-CrossNet reaches 13.224 dB with the single-channel capacity control, 14.208 dB with CRPF-R, and 14.433 dB with default seven-map D-CRPF. Relative to the capacity control, the two spatial variants gain 0.984 and 1.209 dB; the dual-direction correction adds 0.225 dB over CRPF-R and reduces WER from 8.65% to 8.14%. The additional nine-map input ablation reaches 14.527 dB and 7.91% WER but lowers PESQ (Appendix B.3). All four adapters share Stage 1 epoch 29 and the complete 30-epoch training budget; checkpoints are selected by validation SI-SDR. Their single-channel control is defined in Appendix B, not by the legacy A1 beam-zeroing rule.

C.2 Natural geometry strata and the joint difficult subset

The natural test RIR sample contains 340/549/354 mixtures with angular difference $\leq 5°$, 5–20°, and $> 20°$, respectively. Physical source-separation bins ≤ 0.15 m, 0.15–0.35 m, and > 0.35 m contain 338/577/328 mixtures. Their joint difficult subset, with angle $\leq 5°$ and separation ≤ 0.15 m, contains 233 mixtures. These are observational strata of the original test set, not controlled interventions.

Table 7: Geometry-stratified observational results for the controlled AV-DPCRN instantiation. SI-SDR is in dB and WER in percent. R and D denote Stage 1+CRPF-R and Stage 1+D-CRPF.

Axis	Bin	N	SI-SDR (dB)↑				WER (%)↓			
			Stage 1	MVDR	R	D	Stage 1	MVDR	R	D
Angle	$\leq 5°$	340	6.38	2.16	7.52	**7.85**	24.14	26.61	18.51	**16.42**
	5–20°	549	6.35	4.08	8.19	**8.98**	22.93	16.77	12.63	**10.82**
	$> 20°$	354	7.11	6.98	9.36	**10.87**	18.37	**9.61**	12.56	9.86
Distance	≤ 0.15 m	338	6.52	2.25	7.69	**8.07**	22.87	25.71	17.49	**14.54**
	0.15–0.35 m	577	6.38	4.21	8.24	**9.06**	22.89	16.02	13.07	**11.71**
	> 0.35 m	328	6.98	6.87	9.20	**10.65**	19.44	11.20	12.73	**10.09**

At angular separation $\leq 5°$, MVDR reaches only 2.16 dB and 26.61% WER, while final D-CRPF reaches 7.85 dB and 16.42%. At $> 20°$, MVDR retains a slight WER advantage (9.61% versus 9.86%), whereas D-CRPF reaches 10.87 versus 6.98 dB SI-SDR. Physical separation exhibits the same broad trend, although D-CRPF has the best WER in all three distance bins. The result therefore does not imply that the adapter always gives lower WER than beamforming.

Table 8: Joint spatially difficult subset for the controlled AV-DPCRN instantiation: angle $\leq 5°$ and source separation ≤ 0.15 m ($N = 233$).

System	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
Mixture	0.07	1.249	0.660	61.88
Stage 1	6.45	1.733	0.823	24.55
MVDR	1.90	1.463	0.762	27.31
Stage 1 + CRPF-R	7.44	1.930	0.859	19.57
Stage 1 + D-CRPF	**7.71**	**2.000**	**0.863**	**16.18**

On the joint difficult subset, D-CRPF improves Stage 1 by 1.26 dB and 8.37 WER points. Relative to MVDR, it gains 5.81 dB and 11.13 WER points. Because utterance content and acoustic conditions can vary between the natural bins, these differences alone do not identify a causal effect of source angle or distance.

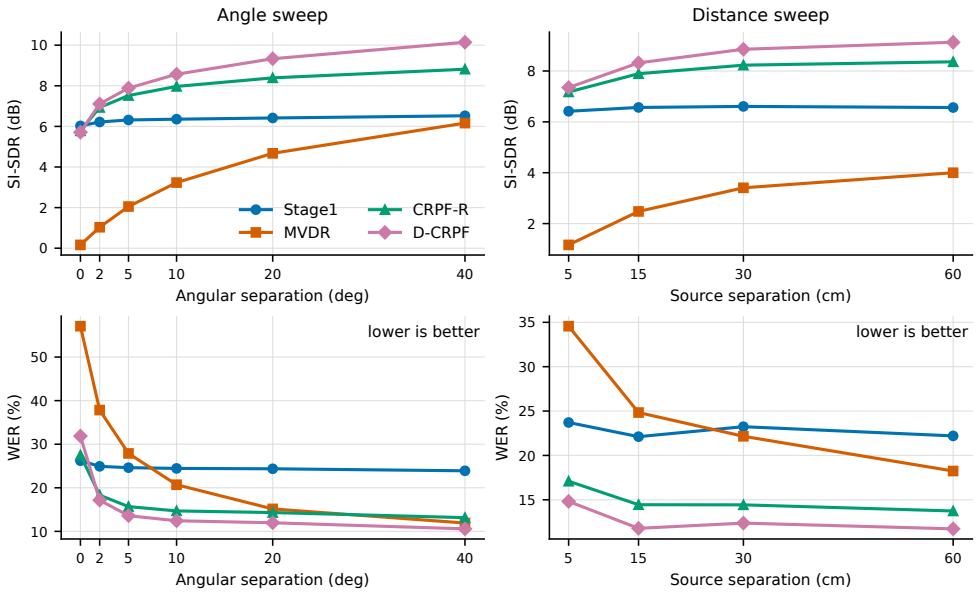

Figure 4: Paired controlled geometry sweeps for Stage 1, MVDR, CRPF-R, and final D-CRPF with the controlled AV-DPCRN instantiation on all 1,243 full-duration utterances. Only the indicated interferer coordinate changes; speech, SIR, room, RT60, target and array are fixed per utterance.

C.3 Paired angle and distance interventions

Figure 4 reports a separate controlled sweep, not the historical array-mismatch re-rendering discussed in Appendix C.5. All ten conditions contain the same 1,243 full-duration utterances, speech pairs, SIRs, target levels, rooms, RT60s, target positions, and 8-cm circular six-microphone array. Rooms are selected from a common feasible set so no condition expands a room. The angle sweep rotates only the interferer at fixed horizontal range and height $(0, 2, 5, 10, 20, 40°)$, alternating rotation direction across utterances. The distance sweep keeps both sources collinear and moves only the interferer radially outward (5, 15, 30, 60 cm). Its sample-level audit verifies all 37,290 rendered waveforms and every fixed field. The later finding of condition-dependent room repairs in the separate array-mismatch generator does not describe this independently controlled angle/distance suite.

Exact collinearity is a sharp failure boundary: D-CRPF reaches 5.71 dB SI-SDR and 31.87% WER, 0.314 dB below Stage 1 (95% CI $[-0.380, -0.247]$) with 5.65 more WER points ($[1.49, 12.66]$), although it still exceeds MVDR by 5.55 dB. At only 2°, the result rises to 7.11 dB/17.17% WER; relative to Stage 1, this is $+0.892$ dB ($[0.812, 0.972]$) and 7.76 fewer WER points ($[6.22, 9.33]$). The dual-direction decoder is already useful there, exceeding CRPF-R by 0.177 dB ($[0.132, 0.222]$) and 1.14 WER points ($[0.07, 2.24]$).

At 40°, D-CRPF reaches 10.14 dB/10.58% and exceeds Stage 1 by 3.62 dB and 13.32 WER points, MVDR by 3.98 dB and 1.34 points, and CRPF-R by 1.33 dB and 2.58 points; every interval excludes zero. Increasing physical separation from 5 to 60 cm improves D-CRPF by 1.782 dB ($[1.619, 1.947]$) and 3.09 WER points ($[1.83, 4.35]$). From 30 to 60 cm, the 0.276-dB gain remains resolved, but the WER change is unresolved, indicating partial saturation. These paired interventions isolate an effect of the indicated interferer coordinate within this simulation and backbone, while exposing exact directional coincidence as a concrete failure mode.

C.4 Visual reliability and failure

We degrade the same 1,243 examples without changing audio. For low resolution, each mouth ROI is downsampled to 16×16 and restored before the visual encoder. For partial loss, the middle 40% of the visual embedding sequence is zeroed. Full loss zeros the entire visual sequence. Table 9 pairs final D-CRPF with the unadapted controlled AV-DPCRN front end under each identical

corruption. Saved sample identifiers, target/interferer identities, mixture and target paths, durations, SIRs, RT60s, angles, and distances agree across each pair. Low resolution costs the adapted system 0.701 dB SI-SDR relative to normal video (95% paired CI $[-0.979, -0.432]$) and raises WER by 2.68 points ($[1.45, 3.97]$); partial loss costs 2.005 dB ($[-2.222, -1.808]$) and raises WER by 3.91 points ($[2.66, 5.21]$).

Table 9: Paired visual stress tests with the controlled AV-DPCRN front end on all 1,243 full-duration mixtures. D denotes Stage 1+D-CRPF; Δ is D minus the same-condition Stage 1. WER uses 6,950 reference words.

Visual condition	System	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
Normal	Stage 1	6.574	1.734	.827	21.94
	D	9.208	2.258	.883	12.04
	Δ	+2.634	+.524	+.056	−9.90
16×16 ROI	Stage 1	5.926	1.685	.811	24.88
	D	8.508	2.187	.868	14.72
	Δ	+2.582	+.501	+.057	−10.16
Middle 40% missing	Stage 1	4.733	1.548	.768	29.22
	D	7.203	2.000	.837	15.96
	Δ	+2.470	+.452	+.069	−13.27
All video missing	Stage 1	−3.061	1.180	.526	78.23
	D	−3.404	1.308	.548	65.22
	Δ	−.344	+.128	+.022	−13.01

Table 10: The same paired visual comparison on the joint spatially difficult subset ($N = 233$, angle $\leq 5^\circ$ and source separation ≤ 0.15 m). D and Δ are defined as in Table 9; subset WER uses 1,267 reference words.

Visual condition	System	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
Normal	Stage 1	6.454	1.733	.823	24.55
	D	7.711	2.000	.863	16.18
	Δ	+1.256	+.267	+.040	−8.37
16×16 ROI	Stage 1	5.483	1.671	.792	30.78
	D	6.646	1.925	.835	21.94
	Δ	+1.162	+.254	+.042	−8.84
Middle 40% missing	Stage 1	4.760	1.565	.761	30.54
	D	5.940	1.797	.811	20.21
	Δ	+1.180	+.232	+.050	−10.34
All video missing	Stage 1	−2.827	1.183	.529	74.66
	D	−3.211	1.251	.546	67.09
	Δ	−.384	+.068	+.018	−7.58

The paired baselines distinguish inherited visual failure from additional adapter error. Under low resolution and partial loss, adaptation retains 2.582 and 2.470 dB SI-SDR gains and reduces WER by 10.16 and 13.27 points; the difficult-subset gains are 1.162/1.180 dB and 8.84/10.34 WER points. Under complete visual loss, Stage 1 is already at -3.061 dB and adaptation further reduces SI-SDR by 0.344 dB (0.384 dB on the subset). However, PESQ and STOI improve, and WER falls from 78.23% to 65.22% overall and from 74.66% to 67.09% on the subset. Thus the remaining severe failure is metric-dependent, not a uniform deterioration from adaptation. These experiments neither validate a fallback policy nor establish identical visual sensitivity for every released front end. Modality dropout and reliability-aware bypass remain unevaluated.

C.5 Channel subsets and historical array diagnostics

The channel-subset experiment in the first block of Table 11 retains the original 8-cm-diameter six-microphone circular recordings and changes only which recorded channels are passed at inference. The subsets preserve the reference microphone: one channel, adjacent and opposite two-microphone controls, uniformly spaced three microphones, a balanced four-microphone subset, and all six microphones. No model is retrained. These results use the controlled AV-DPCRN checkpoints.

Qualification of the historical re-rendered array suite. The second block retains the previous `rir-generator` 0.3.0 results as historical diagnostics, not as a strictly array-only intervention. An audit on 2026-09-10 found that, relative to circular 8 cm, circular 4 cm, circular 12 cm, and linear 8 cm had respectively 73, 84, and 60 examples with additional changes to the resolved room, coordinate translation, or source coordinates. The legacy generator repaired each geometry separately. Fixed utterance identities and full waveform lengths do not eliminate this acoustic-scene confounding; using the same RIR backend also does not eliminate it. Consequently, the numerical differences and bootstrap intervals below cannot identify a causal effect of microphone geometry alone. This qualification concerns the re-rendered array block, not the channel subsets or the independent angle/distance sweep in Appendix C.3.

Table 11: Controlled AV-DPCRN channel subsets and historical array re-rendering for final D-CRPF. The two blocks use separate test suites. The lower block is confounded by condition-dependent scene repairs and is retained only as a historical diagnostic, not evidence of strictly array-only robustness. Higher is better except WER.

Condition	M	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
Channel subsets of the original 8-cm circular array					
Reference only	1	4.12	1.412	.790	30.62
Adjacent pair	2	8.02	1.855	.863	17.51
Opposite pair	2	7.26	1.754	.851	18.26
Uniform subset	3	9.05	2.133	.882	13.02
Balanced subset	4	9.09	2.198	.883	12.36
All microphones	6	**9.21**	**2.258**	**.883**	**12.04**
Historical six-microphone re-rendering: scene-confounded					
Circular, 4 cm	6	8.96	2.180	.881	12.14
Circular, 8 cm	6	9.19	2.253	**.883**	11.84
Circular, 12 cm	6	**9.24**	**2.271**	.882	**11.67**
Linear, 8 cm	6	7.77	1.977	.865	15.80

Three and four microphones are 0.163 ($[-0.191, -0.134]$) and 0.119 dB ($[-0.139, -0.099]$) below all six. Their WER increases are 0.98 points ($[0.18, 1.80]$) and 0.32 points ($[-0.33, 0.98]$), respectively. Reference-only inference loses 5.083 dB ($[-5.329, -4.839]$) and increases WER by 18.58 points ($[16.43, 20.86]$). Count alone remains insufficient: the opposite pair is 0.751 dB worse than the adjacent pair ($[-0.819, -0.683]$), although their 0.75-point WER difference is unresolved. The severe reference-only loss shows that the trained beam-gating path cannot be assumed reliable when its spatial alternatives are removed.

In particular, reference-only adaptation is 2.449 dB below the unadapted Stage 1 (6.574 dB), with 8.68 more WER percentage points (30.62% versus 21.94%). Thus the identity skip is not a post-training performance floor. A deployment could explicitly bypass the adapter when only one functioning microphone is available, recovering Stage 1 by construction; this policy is not enabled in the reported results and has not been evaluated for intermittent channel faults. Covariance conditioning or proxy-reliability triggers would additionally require calibration and validation rather than being implied by the correction bound.

For completeness, the historical re-rendering comparison gave a 0.236-dB reduction for circular 4 cm relative to circular 8 cm ($[-0.320, -0.151]$), with an unresolved 0.30-point WER increase. Circular 12 cm differed by $+0.050$ dB ($[-0.029, 0.124]$) and -0.17 WER points ($[-0.92, 0.58]$), neither resolved. Linear 8 cm differed by -1.425 dB ($[-1.520, -1.329]$) and $+3.96$ WER points ($[2.84, 5.11]$). All intervals in this subsection use paired utterance bootstraps with 10,000 replicates and seed 20260820. In the re-rendering block, they quantify uncertainty in the observed joint geometry-and-scene change; bootstrapping does not repair the confounding.

The corrected strong-model experiment below uses a separate jointly planned rendering; no score from the confounded historical block is substituted into it.

C.6 CORRECTED STRONG-MODEL ARRAY COMPARISON

The corrected suite retains the same 1,243 target/interferer pairs, complete target durations, interferer crop starts and looping choices, and original RIR-scene identities across four six-microphone conditions. Circular 4/8/12 cm denotes the array *diameter*; linear 8 cm denotes the end-to-end length. The array center and orientation are shared. For each scene, the union of all four microphone geometries and both source positions is resolved once, producing one common room-axis order, room size, coordinate translation, effective RT60, and RIR length. Thus any required scene repair is identical across the four conditions, unlike the historical generator.

Both source gains are calibrated on the newly rendered circular-8-cm early images and then shared across all conditions. A single additional safety gain, determined jointly over the four renderings, prevents clipping without condition-specific rescaling. The prescribed SIR therefore denotes calibration on the canonical circular-8-cm condition, not identical received SIR at every microphone in every array. Audio, plan, and manifest hashes preserve the fixed speech content and common-scene provenance.

The reference selector remains channel zero in every condition. Circular 8 cm and linear 8 cm have exactly the same physical reference position and sample-identical reference-channel mixture and source waveforms. Circular 4 cm and circular 12 cm move the reference microphone with the array diameter; they therefore need not preserve the Stage 1 input or reference waveform. Each system is evaluated using the input and designated s_1 reference of its own condition. This separates a fixed-reference topology comparison from aperture changes that also move the reference sensor. The corrected circular-8-cm rendering is the within-suite comparison anchor, not the original official-test waveform set used in Tables 2 and 3; score changes must not be computed by mixing those suites.

Table 12 reports SI-SDR, PESQ, STOI, and WER for the sixteen system–condition pairs, including unadapted Stage 1. Every condition uses fixed checkpoints: AV-CrossNet Stage 1 epoch 29 with D-CRPF epoch 27, AV-SpatialNet epoch 20, and audio-only SpatialNet epoch 24 with clean-target oracle output assignment. No model is retrained for the changed geometry. Oracle assignment uses clean references and is not a deployable target-selection rule.

Independent audits verify all sixteen system–condition rows: 1,243 unique targets, complete metric coverage, 33,223,680 samples at 16 kHz, finite full-length audio, and exact s_1 reference-channel targets. Recomputed word-error counts and metric means match the summaries; checkpoint, manifest, and result hashes bind each row to the shared-scene suite.

The fixed-reference circle-to-line change costs AV-CrossNet+D-CRPF 0.806 dB SI-SDR, versus 4.063 dB for AV-SpatialNet and 4.824 dB for oracle SpatialNet. WER increases by 0.82, 3.06, and 6.81 percentage points, respectively. On the line, the adapted system reaches 13.612 dB and 8.85% WER and leads both references on the four reported metrics. This supports retention under the tested topology change, not arbitrary-array invariance or robustness of every frozen front end.

Final-output robustness versus retained adaptation gain. The retained Stage 1 outputs are scored against the same full-length source-1 channel-zero targets. Circular and linear 8 cm have sample-identical Stage 1 and target PCM and identical transcripts for all 1,243 IDs; this verified equality permits reuse of Stage 1 WER. Their baseline is 13.171 dB. Circular 4/12 cm move the reference microphone and have baselines of 13.058/13.317 dB. None uses the 13.179-dB original-test baseline.

Let b_g and a_g denote the same-suite Stage 1 and adapted SI-SDR for geometry g, with $\Delta_g = a_g - b_g$. Circular 4/8/12 cm yield gains of 1.031/1.246/1.429 dB, and linear 8 cm yields 0.441 dB. Thus $\Delta_{\mathrm{linear}}/\Delta_{\mathrm{circular8}} = 35.4\%$: a descriptive ratio of improvements measured in dB, not an acoustic-energy retention rate. PESQ and STOI gains also remain positive on the line ($+0.216$ and $+0.011$). The smaller complete-system drop therefore coexists with substantial attenuation of the spatial-adaptation gain. An unchanged frozen anchor helps distinguish these effects, but the identity-initialized skip is not a post-training performance floor.

Adaptation improves SI-SDR, PESQ, STOI, and WER in all four conditions. WER falls by 3.15/2.89/3.37/2.07 percentage points for circular 4/8/12 cm and linear 8 cm, respectively. These reported metrics do not establish a general improvement in perceived quality.

Table 12: Corrected jointly rendered array suite: each score covers 1,243 full-duration targets. Every array has six microphones; circular sizes are diameters and linear size is total length. WER uses 6,950 words. Stage 1 shares the original export with its adapted row. No original-test or historical confounded results are substituted.

System	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
Circular, 4 cm				
AV-CrossNet Stage 1	13.058	2.654	.923	11.71
SpatialNet (oracle)	14.347	2.932	.945	8.12
AV-SpatialNet	13.367	2.989	.914	14.75
AV-CrossNet + D-CRPF	14.088	3.025	.941	8.56
Circular, 8 cm (within-suite anchor)				
AV-CrossNet Stage 1	13.171	2.663	.923	10.92
SpatialNet (oracle)	14.778	2.995	.948	8.19
AV-SpatialNet	13.787	3.053	.917	14.49
AV-CrossNet + D-CRPF	14.418	3.093	.944	8.03
Circular, 12 cm				
AV-CrossNet Stage 1	13.317	2.664	.928	10.94
SpatialNet (oracle)	14.864	2.999	.948	7.87
AV-SpatialNet	13.905	3.070	.916	13.32
AV-CrossNet + D-CRPF	14.746	3.119	.950	7.57
Linear, 8 cm (same reference position as circular 8 cm)				
AV-CrossNet Stage 1	13.171	2.663	.923	10.92
SpatialNet (oracle)	9.954	2.246	.888	14.99
AV-SpatialNet	9.724	2.381	.873	17.55
AV-CrossNet + D-CRPF	13.612	2.879	.935	8.85

Reducing circular diameter from 8 to 4 cm costs the adapted system 0.329 dB, versus 0.421 and 0.431 dB for AV-SpatialNet and SpatialNet; WER changes are $+0.53$, $+0.26$, and -0.07 points. At 12 cm, the adapted system gains 0.328 dB and reduces WER by 0.46 points. SpatialNet leads circular-array SI-SDR throughout, so neither uniformly smaller sensitivity nor universal superiority follows.

C.7 Residual-only reproduction and mask-guided MVDR diagnostic

Table 13 reports a separate full-data AV-DPCRN adaptation run (adapter seed 20260826) on all 1,243 full-duration targets; it does not replace the earlier mechanism or robustness studies. Its epoch-28 checkpoint implements CRPF-R with two residual output channels and no signed beam-disagreement head. Raw MVDR lowers WER relative to Stage 1 despite poorer SI-SDR, PESQ, and STOI. Applying the frozen AV mask, $M_{\mathrm{AV}} \odot Y_{\mathrm{MVDR}}$, improves the four reported metrics over raw MVDR, but learned CRPF-R gains another 4.026 dB and reduces WER by 5.31 points over that fixed-mask control.

Table 13: Independent full-data reproduction of intermediate CRPF-R with AV-DPCRN and a fixed mask-guided MVDR diagnostic on all 1,243 full-duration mixtures. This run is separate from the earlier fixed-budget mechanism study. Higher is better except WER.

System	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
Stage 1	6.706	1.754	.826	21.83
MVDR	4.433	1.719	.821	17.61
MVDR + frozen AV mask	5.359	1.884	.828	16.59
Stage 1 + CRPF-R	**9.385**	**2.270**	**.891**	**11.28**

C.8 DURATION-STRATIFIED PAIRED EVALUATION

We stratify the existing matched AV-CrossNet epoch 29 and D-CRPF epoch 27 outputs by their full utterance duration, without selecting new checkpoints or rerunning inference. All 1,243 identifiers and original-test scene records agree with the source export. The three disjoint bins contain 132, 835, and 276 utterances and 519, 4,095, and 2,336 normalized reference words, respectively. Table 14 reports paired means. WER is reconstructed from saved normalized ASR strings with the original edit-count routine and aggregated by each bin's word count; it is not the mean of per-utterance WER.

Table 14: Duration-stratified matched AV-CrossNet evaluation on the original full-duration test set. D denotes Stage 1+D-CRPF; Δ is D minus Stage 1. These are observational test subsets, not separate training runs or a controlled duration intervention.

Duration	N	System	SI-SDR (dB)↑	PESQ↑	STOI↑	WER (%)↓
≤ 1 s	132	Stage 1	10.642	2.387	.883	25.63
		D	11.513	2.742	.906	19.85
		Δ	+.871	+.354	+.023	-5.78
$(1, 2]$ s	835	Stage 1	13.227	2.660	.926	12.14
		D	14.469	3.089	.946	9.01
		Δ	+1.242	+.428	+.020	-3.13
> 2 s	276	Stage 1	14.246	2.792	.936	5.31
		D	15.720	3.264	.956	4.02
		Δ	+1.473	+.472	+.020	-1.28

The SI-SDR gain remains positive in all three bins, including utterances longer than the 2-s training/validation crop. This does not establish that longer context causes a larger adaptation benefit: duration co-varies with speech content, and the short-utterance bin has markedly higher baseline WER. Nor does this analysis resolve the validation-to-test horizon mismatch or single-run checkpoint sensitivity of AV-SpatialNet, which would require a separate validation analysis. It only checks that the reported paired adaptation improvement is not confined to the shortest test utterances.