WHERE SHOULD MULTIMODAL TEST-TIME ADAPTATION CAPACITY LIVE? A CONTROLLED STUDY OF PROMPTS AND RESIDUAL ADAPTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study where to place a fixed online adaptation budget in a frozen audio-visual transformer. Multimodal test-time adaptation must respond to audio-only, video-only, and paired shifts, yet prompt capacity and residual adapter capacity impose different update structures. We isolate this choice under a fixed BriMPR objective and a near-matched trainable-parameter budget, comparing Prompt78 with identity-initialized Prompt+Adapter capacity. The adapter is exactly identity at initialization, while a two-step integrity test verifies the intended gradient flow. Under a validated batch-32 severity-5 common-batch protocol (48 cases: 2 datasets $\times$ 3 corrupted modalities $\times$ 8 methods, three target-stream seeds), Prompt+Adapter attains the highest mean accuracy in all six dataset–modality settings, improving over Prompt78 by 0.04–2.41 percentage points and over the unadapted source model by 8.28 points on average. Per-domain permutation tests show the advantage is decisive on paired corruptions and on VGGSound-C, while the two configurations are within seed-level parity on single-modality KS50-C shifts. On paired corruptions the ranking is stable across severities 1, 3, and 5 on KS50-C, and placement and width pilots show that visual-side adapters match dual-side adapters while wider capacity helps both families. An earlier batch-64 protocol exhibits one dataset–modality inversion and is reported separately in the appendix. We deliberately do not make claims about universal superiority, causal mechanisms, or deployment efficiency.

1 INTRODUCTION

Audio-visual models can encounter shifts in either modality or in both modalities at deployment. Test-time adaptation (TTA) can respond to these shifts without source-data retraining, but multimodal TTA introduces a design choice that is usually entangled with the adaptation objective: where should the online trainable capacity be placed? Entropy-based TTA updates normalization parameters (Wang et al., 2021), while later methods control forgetting, instability, or unreliable samples (Niu et al., 2022; 2023; Lee et al., 2024). Multimodal methods additionally address reliability imbalance and cross-modal alignment (Yang et al., 2024; Li & Feng, 2025). These methods do not isolate the placement of a near-matched adaptation budget while holding the backbone, objective, stream, optimizer, and evaluation accounting fixed.

Figure 1 summarizes this controlled comparison.

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.

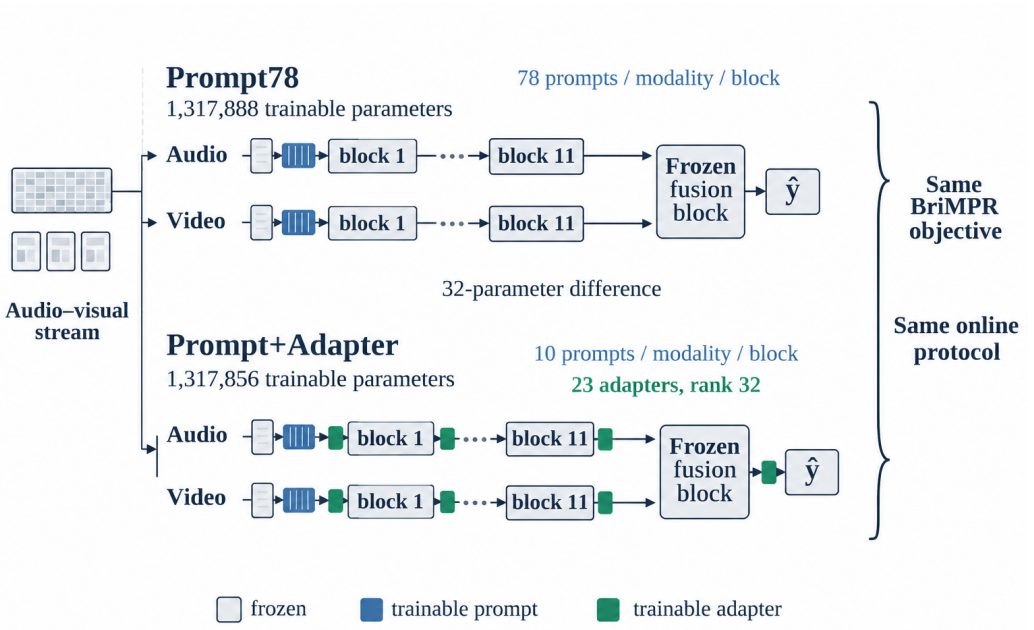

Figure 1: Controlled capacity allocation. Prompt78 and Prompt+Adapter use the same frozen audio–visual backbone, BriMPR objective, and online protocol. Prompt+Adapter retains trainable prompts and adds identity initialized residual adapters; the two configurations differ by 32 trainable parameters.

We study this controlled capacity-allocation question in a frozen CAV-MAE backbone (Gong et al., 2023). We compare prompt capacity with distributed residual adapters under the BriMPR objective. Prompt78 allocates 78 trainable tokens to each modality following the deep prompt variant, inserting tokens into all 11 modality-specific transformer blocks per modality ($78 \times 2 \times 768 \times 11 = 1{,}317{,}888$ trainable parameters). Prompt+Adapter retains 10 deep prompts per modality and adds identity-initialized bottleneck adapters, totalling 1,317,856 trainable parameters — a 32-parameter difference in trainable capacity. This design separates capacity placement from the adaptation objective and from large changes in trainable parameter count. A standalone BriMPR baseline tunes the 10 deep prompt tokens per modality of the original method (168,960 trainable parameters, no adapters).

Our validated evidence supports three scoped conclusions. First, the asymmetric adapter initialization is exactly identity at initialization and permits gradient flow across successive updates. Second, under the validated batch-32 severity-5 common-batch protocol, Prompt+Adapter attains the highest mean accuracy in all six dataset–modality settings, ahead of Prompt78, BriMPR, and five standard TTA baselines evaluated under identical accounting. Third, the advantage is decisive on paired corruptions and grows with corruption severity, while single-modality KS50-C shifts exhibit seed-level parity. An earlier batch-64 protocol shows one inversion (VGGSound-C video), which we report separately in Appendix A as evidence of batch-size sensitivity rather than as the main result.

Our contributions are threefold. First, we formulate controlled capacity allocation for multimodal TTA by comparing prompts and residual adapters under one objective and a near-matched trainable budget. Second, we provide an identity-initialized adapter construction together with an executable two-step gradient-flow test, which makes the initialization and optimizer wiring auditable. Third, under one accounting protocol we compare eight methods across 48 cases and show that Prompt+Adapter leads all six dataset–modality settings; severity, placement, and width pilots then characterize where the advantage concentrates. We keep broader mechanism and efficiency statements outside the scope of the evidence reported here.

2 Related Work

2.1 Test-time and multimodal adaptation

Entropy minimization updates normalization parameters online in Tent (Wang et al., 2021). EATA controls redundancy and forgetting (Niu et al., 2022), SAR targets unstable updates (Niu et al., 2023), and DeYO filters samples using factors beyond entropy (Lee et al., 2024). We evaluate Tent, EATA, and SAR; DeYO provides related context and is not in our benchmark. For multimodal adaptation, READ addresses modality reliability bias (Yang et al., 2024), while BriMPR progressively re-aligns features through source-statistics matching and cross-modal interaction (Li & Feng, 2025). We retain the latter's objective, backbone, and stream accounting to compare where a near-matched trainable budget acts. The study concerns the capacity parameterization rather than a new adaptation loss.

2.2 Parameter-efficient capacity

Prompt tuning learns input tokens (Lester et al., 2021), including deep prompts for visual transformers (Jia et al., 2022). Residual adapters learn bottleneck transformations (Houlsby et al., 2019); unified views emphasize their different structural constraints at similar parameter counts (He et al., 2022). Low-rank weight adaptation (Hu et al., 2022) offers a further placement outside this study. In our comparison, prompts modify a frozen block's attention context, whereas adapters modify its output representation. Equal trainable counts do not imply equal sequence length, activation memory, or latency. We examine these interfaces under online, unlabeled audio-visual shifts.

3 Controlled Capacity Allocation

We separate the adaptation objective from the parameterization that expresses its update. Following BriMPR (Li & Feng, 2025), the source model comprises audio and visual encoders, a fusion module, and a classifier. All pretrained weights θ are frozen; only prompts and, when present, adapters form the online parameters ϕ. Figure 2 locates these interfaces in CAV-MAE (Gong et al., 2023).

3.1 Online setting and capacity interfaces

At time t, the model receives an unlabeled batch $\mathcal{B}_t = \{(x_j^a, x_j^v)\}_{j=1}^B$. Predictions are recorded using ϕ_t before any update on $\mathcal{B}_t$. Three Adam steps on the shared loss then produce ϕ_{t+1}. Each corruption domain and target-stream seed starts from the same checkpoint and initial trainable state, with optimizer moments and AMP scaler restored.

Let $E_{\ell-1}^u \in \mathbb{R}^{n_u \times d}$ be the non-prompt sequence for modality $u \in \{a, v\}$, $P_\ell^u \in \mathbb{R}^{p \times d}$ the layer-specific prompts, and T_ℓ^u a frozen transformer block. Writing R_p for removal of the p prompt positions, the forward recurrence is

$$H_\ell^u = T_\ell^u([P_\ell^u; E_{\ell-1}^u]), \qquad E_\ell^u = R_p(A_\ell^u(H_\ell^u)), \quad \ell = 1, \ldots, L. \tag{1}$$

Concatenation is along the token axis. Prompts are replaced at each layer, not accumulated; $L = 11$ and $d = 768$. Prompt78 sets $p = 78$ and $A = I$. Prompt+Adapter sets $p = 10$ and applies a token-wise residual adapter after each complete block. The fusion block receives $[E_L^a; E_L^v]$ without additional prompts and is followed by one more adapter. Frozen normalization, pooling, and the prediction head complete the model.

For a column token $h \in \mathbb{R}^d$, the adapter is

$$A(h) = h + U\,\sigma(Dh + b_d) + b_u, \qquad D \in \mathbb{R}^{r \times d}, \quad U \in \mathbb{R}^{d \times r}, \tag{2}$$

where σ is GELU and $r = 32$. This residual construction follows bottleneck adapters (Houlsby et al., 2019); our comparison reallocates online capacity while retaining the BriMPR loss and a small deep-prompt base. Including biases, the trainable counts are

$$\begin{aligned} N_{\text{prompt}}(p) &= 2Lpd, \\ N_{\text{P+A}}(p, r) &= 2Lpd + (2L + 1)(2dr + r + d). \end{aligned} \tag{3}$$

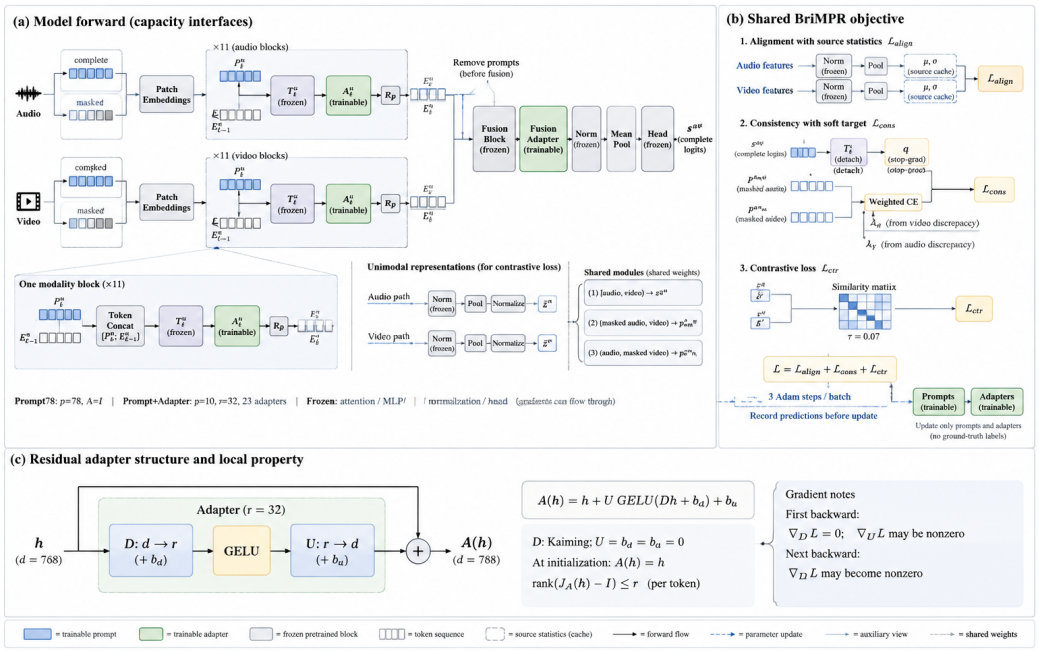

Figure 2: Method overview. (a) The shared frozen audio–visual backbone exposes deep prompts before each modality-specific block and residual adapters after the blocks and fusion block. (b) Both configurations optimize the fixed BriMPR objective with masked and complete views; predictions are recorded before the online update. (c) The adapter is identity initialized, and its local residual Jacobian has rank at most r per token.

Thus $N_{\text{prompt}}(78) = 1{,}317{,}888$ and $N_{\text{P+A}}(10, 32) = 1{,}317{,}856$, a 32-parameter difference. The ten-prompt BriMPR reference has 168,960 parameters. Matching this count controls capacity size, not computation or memory.

3.2 Shared Progressive Re-alignment Objective

We retain BriMPR's three loss components (Li & Feng, 2025). For each layer, let $z^u_{\ell,j}$ be the pooled non-prompt feature after the frozen normalization used by the implementation, and let $(\hat{\mu}^u_{s,\ell}, \hat{\sigma}^u_{s,\ell})$ be cached source statistics. The batch mean and sample standard deviation define

$$\hat{\mu}^u_{t,\ell} = \frac{1}{B}\sum_j z^u_{\ell,j}, \qquad \hat{\sigma}^u_{t,\ell} = \left[\frac{1}{B-1}\sum_j (z^u_{\ell,j} - \hat{\mu}^u_{t,\ell})^{\odot 2}\right]^{1/2},$$
$$\Delta_u = \frac{1}{L}\sum_{\ell=1}^{L}\left(\|\hat{\mu}^u_{s,\ell} - \hat{\mu}^u_{t,\ell}\|_2 + \|\hat{\sigma}^u_{s,\ell} - \hat{\sigma}^u_{t,\ell}\|_2\right), \quad \mathcal{L}_{\text{align}} = \Delta_a + \Delta_v. \tag{4}$$

The squares and square roots are elementwise. This matches means and standard deviations, rather than introducing a covariance-matching objective. The code retains the numerical offset of its Euclidean-distance primitive.

For cross-modal consistency, let s_j^{av} denote complete-input logits and $p_j^{a_m v}, p_j^{a v_m}$ the softmax predictions obtained by masking one modality's patches. With sg denoting stop-gradient,

$$T_t = 1 + \frac{\tau_0}{1 + \exp(D_0 - \mathrm{sg}(\Delta_f))}, \qquad q_j = \mathrm{sg}\big[\mathrm{softmax}(s_j^{av}/T_t)\big],$$
$$\lambda_a = \frac{\mathrm{sg}(\Delta_v)}{\max(\mathrm{sg}(\Delta_a + \Delta_v), 10^{-12})}, \qquad \lambda_v = \frac{\mathrm{sg}(\Delta_a)}{\max(\mathrm{sg}(\Delta_a + \Delta_v), 10^{-12})}, \tag{5}$$
$$\mathcal{L}_{\mathrm{cons}} = -\frac{1}{B}\sum_{j,k} q_{j,k}\left(\lambda_a \log p_{j,k}^{a_m v} + \lambda_v \log p_{j,k}^{a v_m}\right).$$

Here Δ_f is the analogous discrepancy for the pooled fusion feature; τ_0, D_0 are BriMPR temperature parameters, and k indexes classes. The target, temperature, and modality weights are detached from differentiation. Masked and complete views share all parameters.

Let $\bar{z}_j^u$ be the normalized unimodal representation after the shared fusion module. Bidirectional instance contrast and the final loss are

$$\mathcal{L}_{\mathrm{ctr}} = -\frac{1}{2B}\sum_{j=1}^{B}\sum_{u\neq v}\log\frac{\exp(\langle \bar{z}_j^u, \bar{z}_j^v\rangle/\tau)}{\sum_{i=1}^{B}\exp(\langle \bar{z}_j^u, \bar{z}_i^v\rangle/\tau)}, \tag{6}$$
$$\mathcal{L}_{\mathrm{BriMPR}} = \mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{cons}} + \mathcal{L}_{\mathrm{ctr}}, \qquad \tau = 0.07.$$

The ordered modality pairs are (a, v) and (v, a). Both capacity configurations use one joint backward pass per update, Adam with learning rate 10^{-4}, and AMP initial scale 1024. There is no additional teacher, distillation term, or placement-dependent loss in this comparison.

3.3 Identity initialization and local update structure

The initialization must preserve the residual interface without trapping both projections at zero. Set D to its Kaiming initialization and $U = b_d = b_u = 0$. Then $A(h) = h$ for every finite input. This equality compares the same prompt configuration with and without adapters; it does not equate the initial predictions of 10- and 78-prompt models.

To derive the gradient schedule, write $a = \sigma(Dh + b_d)$, $g = \nabla_{A(h)}\mathcal{L}$, and $S = \mathrm{diag}(\sigma'(Dh + b_d))$. The chain rule gives

$$\nabla_U \mathcal{L} = g a^\top, \quad \nabla_{b_u}\mathcal{L} = g, \quad \nabla_D \mathcal{L} = (SU^\top g)h^\top, \quad \nabla_{b_d}\mathcal{L} = SU^\top g. \tag{7}$$

For shared token-wise parameters these terms are summed over tokens and examples. Initially, both down-projection gradients are exactly zero; up-projection gradients can be nonzero because D is nonzero. After an update changes U, gradients can reach D on the next backward pass. Nonzero gradients still require a nondegenerate loss signal and non-cancelling sums. The two-step integrity test checks this behavior on its test inputs; it is not a guarantee for every target batch.

The same construction characterizes the local representation change:

$$J_A(h) - I = USD, \qquad \mathrm{rank}(J_A(h) - I) \le r, \qquad \|J_A(h) - I\|_2 \le \|U\|_2\|S\|_2\|D\|_2. \tag{8}$$

This is a per-token, single-adapter statement. It distinguishes a local bottleneck correction from prompt-mediated changes inside frozen attention. For a frozen downstream map G differentiable at h, the first-order effect of $\delta h = A(h) - h$ is $J_G(h)\delta h$; its remainder is $o(\|\delta h\|)$ as $\delta h \to 0$. Consequently, placement changes the downstream transformation of a correction even at the same parameter budget. Neither the local rank bound nor identity initialization orders the two families' accuracy or proves a causal explanation for the placement pilots. Appendix C gives the proof and its scope.

4 Experiments

This section evaluates the capacity choices under a common protocol and then examines how their behavior changes with severity, placement, and stream conditions.

Table 1: Mean test accuracy (%) under the validated batch-32 severity-5 common-batch protocol (48 cases: 2 datasets × 3 corrupted modalities × 8 methods; seeds 111/222/333; KS50-C: 15 video, 6 audio, 90 paired domains; VGGSound-C: 15 video, 6 audio, 15 paired domains). Best per row in bold. Source performs no adaptation and has no seed variance. Margins quoted in the text and figures are computed from unrounded means.

Dataset	Modality	Source	Tent	EATA	SAR	READ	BriMPR	Prompt78	Prompt+Adapter
KS50-C	Video	59.79	47.97±0.37	61.24±0.04	51.37±0.24	63.51±0.02	65.50±0.09	66.54±0.04	**66.94±0.12**
KS50-C	Audio	69.30	26.01±2.00	44.39±1.67	43.90±1.75	71.49±0.01	72.31±0.18	72.68±0.10	**72.72±0.14**
KS50-C	Paired	30.69	9.80±0.03	19.65±0.01	19.54±0.05	37.64±0.06	40.70±0.03	42.54±0.03	**44.77±0.10**
VGGSound-C	Video	55.97	37.42±1.19	56.04±0.02	42.87±0.37	56.22±0.06	56.12±0.07	55.75±0.05	**57.46±0.04**
VGGSound-C	Audio	24.92	1.15±0.39	9.20±2.01	8.69±0.43	21.83±0.46	35.75±0.14	34.67±0.07	**35.88±0.05**
VGGSound-C	Paired	8.75	0.53±0.00	4.23±0.02	2.92±0.00	12.22±0.08	19.72±0.04	18.90±0.10	**21.32±0.10**

VGGSound-C EATA estimates Fisher information using clean test-split media instead of unavailable training media. This introduces test-set information leakage; those cells are not a strictly source-only EATA comparison.

4.1 Protocol and evaluation

We use Kinetics50-C (KS50-C) and VGGSound-C, derived from Kinetics and VGGSound (Kay et al., 2017; Chen et al., 2020), with common visual corruptions following established robustness benchmarks (Hendrycks & Dietterich, 2019). The main protocol evaluates, at severity 5 and batch size 32, eight methods on a common batch layout: Source (no adaptation), Tent, EATA, SAR, READ, BriMPR, Prompt78, and Prompt+Adapter. For each dataset we evaluate video-only, audio-only, and paired corruption. KS50-C uses 15 video, 6 audio, and 90 paired domains; VGGSound-C uses 15 video, 6 audio, and 15 paired domains (each video corruption paired with one fixed audio corruption). Target-stream seeds are 111, 222, and 333. This yields 48 cases in total ($2 \times 3 \times 8$), all of which completed and passed validation. Every reported result has finite losses, exact prediction counts, per-sample predictions, and immutable provenance in the registry. OOM, NaN, incomplete, smoke, pre-fix, and CPU fallback outputs are excluded from analysis.

Ablations reuse the same protocol. Severity is studied on KS50-C at severities 1, 3, and 5 for the three capacity configurations. Placement and width pilots use six paired domains at severity 5; the 145-prompt width case ran at batch 16 for memory reasons and is labeled accordingly. An earlier batch-64 comparison of the three capacity configurations is retained in Appendix A and is not mixed with the batch-32 evidence.

4.2 Common-batch comparison

We first compare all methods on the validated batch-32 protocol. Table 1 reports the complete validated batch-32 evidence. Prompt+Adapter attains the highest mean accuracy in all six dataset–modality settings. Against Prompt78, the margins are $+0.41$ (KS50-C video), $+0.04$ (KS50-C audio), $+2.23$ (KS50-C paired), $+1.71$ (VGGSound-C video), $+1.22$ (VGGSound-C audio), and $+2.41$ (VGGSound-C paired) percentage points, computed from unrounded means. Against the unadapted source model, Prompt+Adapter improves by 8.28 points on average across the six settings (minimum $+1.49$, maximum $+14.09$). Tent, EATA, and SAR degrade below the source model in most settings under this accounting, while READ is competitive on single-modality shifts but behind the capacity configurations on paired shifts. A sign-flip permutation test over per-domain mean differences (exact for at most 20 domains; 2×10^5 Monte-Carlo draws for the 90-domain paired setting) shows the margin is decisive on paired corruptions (KS50-C: 85/90 domains won, $p < 10^{-5}$; VGGSound-C: 15/15, $p = 6 \times 10^{-5}$) and on VGGSound-C single-modality shifts (video: 14/15, $p = 3 \times 10^{-4}$; audio: 6/6, $p = 0.031$), whereas KS50-C single-modality shifts are at seed-level parity (video: 8/15 domains, $p = 0.16$; audio: 3/6, $p = 0.94$). We therefore read the advantage as concentrated on compound paired shifts and on VGGSound-C, with single-modality KS50-C at parity.

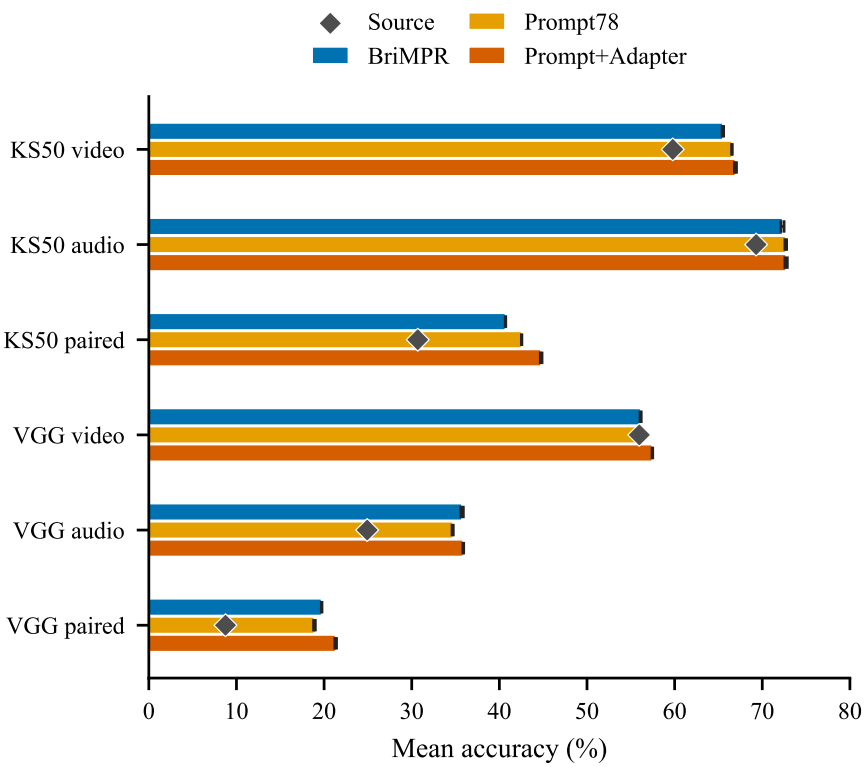

Figure 3: Mean accuracy for the three capacity configurations across the six dataset-modality settings under the batch-32 common-batch protocol. Error bars show the sample standard deviation over the three registered target-stream seeds; diamonds mark the unadapted source model.

4.3 Stress tests and placement analysis

We next test sensitivity to corruption severity and the location and width of the trainable capacity. Table 2 reports every cell with the sample standard deviation over the three target-stream seeds. On paired corruptions the ordering among the three capacity configurations is stable as severity changes, and the margin of Prompt+Adapter over Prompt78 grows from $+0.71$ points at severity 1 to $+1.54$ at severity 3 and $+2.23$ at severity 5, indicating that the advantage concentrates where the shift is strongest. Single-modality margins are small at all severities, and two severity-3 cells show inversions within one standard error (video: Prompt78 74.00 vs. Prompt+Adapter 73.97; audio: Prompt78 74.23 vs. Prompt+Adapter 74.19).

Table 3 reports the placement and width pilots on paired corruptions. All configurations retain the 10 deep prompt tokens per modality of Prompt+Adapter; adapter ranks apply to every adapted branch, and prompt counts are per modality. Placing adapters on the visual branch alone (38.53) matches dual-branch placement (38.43), while audio-only placement is lower (34.45) and fusion-only placement is the weakest (31.23). The fusion-only row concentrates the full adapter budget in the single fusion block, which forces a near-full-rank bottleneck (r=747 against a 768-dimensional hidden state) and removes the low-rank regularization present in the other rows; the fusion drop therefore conflates placement with rank expansion, and we do not read it as evidence against fusion-block adaptation in general. Widening capacity helps both families monotonically over the probed range: adapter bottleneck r=8/16/64 yields 36.44/37.35/38.61, and 28/45/145 prompts yield 33.97/34.42/35.50 (the 145-prompt case ran at batch 16 for memory reasons, so its improvement conflates width with batch size and is labeled accordingly). These are pilot results on six paired domains and characterize trends rather than exhaustive rankings.

Table 2: KS50-C accuracy (%) across severities 1/3/5 (batch 32), with the sample standard deviation over the three target-stream seeds. Severity-5 rows are the common-batch results of Table 1; severities 1 and 3 reuse the same protocol per severity.

Modality	Method	Sev. 1	Sev. 3	Sev. 5
Video	BriMPR	78.67±0.03	73.28±0.04	65.50±0.09
	Prompt78	78.69±0.05	74.00±0.03	66.54±0.04
	Prompt+Adapter	78.87±0.07	73.97±0.08	66.94±0.12
Audio	BriMPR	74.75±0.14	73.77±0.16	72.31±0.18
	Prompt78	75.23±0.01	74.23±0.05	72.68±0.10
	Prompt+Adapter	75.43±0.05	74.19±0.29	72.72±0.14
Paired	BriMPR	68.34±0.02	56.61±0.03	40.70±0.03
	Prompt78	69.19±0.01	58.57±0.03	42.54±0.03
	Prompt+Adapter	69.90±0.02	60.11±0.03	44.77±0.10

Table 3: Placement and width pilots on KS50-C paired corruptions (severity 5, batch 32, 6 paired domains). †The 145-prompt case ran at batch 16 due to memory limits and is not strictly comparable with the batch-32 rows.

Configuration	Accuracy (%)
Adapter placement	
audio-only (r=68)	34.45
visual-only (r=68)	38.53
audio+visual (r=33 each)	38.43
fusion-only (r=747)	31.23
Capacity width	
adapter r=8	36.44
adapter r=16	37.35
adapter r=64	38.61
28 prompts	33.97
45 prompts	34.42
145 prompts†	35.50

We additionally profiled the three capacity configurations on a 25-batch paired-corruption stream (gaussian noise with traffic noise, severity 5, batch 32) to sanity-check early-stream behavior; mean accuracies over the profiled prefix are 10.13 (BriMPR), 11.46 (Prompt78), and 12.25 (Prompt+Adapter). This profile covers only a short stream prefix and is not a full-stream accuracy measurement, nor a systems-level deployment benchmark; we report it as a behavioral check and draw no efficiency conclusions from it.

Finally, we revisit batch-size sensitivity. An earlier batch-64 protocol (Appendix A) exhibits one inversion: on VGGSound-C video, Prompt78 (57.18) is slightly above Prompt+Adapter (56.82). This inversion does not persist under the batch-32 common-batch protocol, where Prompt+Adapter leads by $+1.71$ points in the same setting. Doubling the batch size also halves the number of online updates over the same finite stream, which interacts with the two-step gradient warm-up of identity-initialized adapters, and learning-rate rescaling across batch sizes remains untested; the inversion therefore cannot be attributed to batch-size statistics alone. We treat it as evidence that the allocation ranking can be sensitive to batch size, and we keep the two protocols strictly separated.

5 Limitations and Conclusion

The evidence covers two datasets, one CAV-MAE backbone, synthetic corruptions, and episodic streams. The main comparison fixes severity 5 and batch 32; severity generalization covers KS50-C only. Placement and width results use six paired domains, and the 145-prompt case changes batch size to 16. The fusion-only pilot also changes bottleneck rank. These confounds prevent a causal placement interpretation. VGGSound-C EATA uses clean test-split media for Fisher estimation because the training-media mount was unavailable; the resulting test-set information leakage limits

comparison with a strictly source-only EATA. Batch-64 evidence contains a ranking inversion and remains separate. Three target-stream seeds and correlated corruption domains limit statistical independence. Continual streams, smaller streaming batches, additional backbones, and other capacity placements remain uncharacterized. Parameter matching and the short prefix profile establish no efficiency advantage.

Under the evaluated protocol, allocating a near-matched budget to Prompt+Adapter gives the strongest gains on paired corruptions and VGGSound-C, while single-modality KS50-C differences are small. The derivations establish identity initialization, conditional gradient activation, and a local rank bound; accuracy comparisons remain empirical. Future work should separate placement from bottleneck width and batch size, and test whether these allocation patterns persist without episodic resets.

AI USE STATEMENT

Generative AI assisted manuscript drafting, editing, and compression; formalization of the method and derivation of adapter gradients and local Jacobian properties; figure design and drawing specifications; and experiment-pipeline monitoring and data-aggregation scripts. The experimental results reported here come from the registered runs. The authors are responsible for verifying the mathematical derivations, the correspondence between implementation and description, the figures, and the experimental evidence before submission.

ETHICS STATEMENT

This work uses publicly available datasets (Kinetics and VGGSound) and synthetic corruptions, involves no human subjects, and raises no foreseeable privacy, security, or fairness concerns beyond those inherent to audio-visual classification models. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The experimental protocol, including datasets, corruption domains, seeds, and registered hyperparameter settings, is described in Section 4 and in the registered-settings table of the appendix. Every reported result is backed by an immutable registry entry with per-sample predictions that are independently reaggregated before inclusion, and the integrity test for the adapter initialization is executable. An anonymized code package is provided as supplementary material, and the full pipeline, manifests, and aggregation scripts will be released publicly upon publication.

REFERENCES

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset, 2020. URL https://arxiv.org/abs/2004.14368.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder, 2023. URL https://arxiv.org/abs/2210.07839.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning, 2022. URL https://arxiv.org/abs/2110.04366.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019. URL https://arxiv.org/abs/1902.00751.

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. URL https://arxiv.org/abs/2106.09685.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022. URL https://arxiv.org/abs/2203.12119.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=9w3iw8wDuE.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021. URL https://arxiv.org/abs/2104.08691.

Jiacheng Li and Songhe Feng. Bridging modalities via progressive re-alignment for multimodal test-time adaptation, 2025. URL https://arxiv.org/abs/2511.22862.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=TPZRq4FALB.

A BATCH-64 CAPACITY EVIDENCE

The following table records the earlier batch-64 capacity comparison of the three capacity configurations. It is generated from registry-approved summaries and is intentionally kept separate from the batch-32 main protocol; numbers across the two tables are not directly comparable.

Table 4: Earlier batch-64 capacity evidence (3 methods, severity 5, seeds 111/222/333), retained for reference. This batch-64 protocol differs from the batch-32 main protocol (Table 1) in batch size and baseline coverage, so numbers are not directly comparable across tables. The VGGSound-C video inversion (Prompt78 slightly above Prompt+Adapter) observed under batch 64 does not persist under the batch-32 common-batch protocol.

Dataset	Modality	BriMPR	Prompt78	Prompt+Adapter
KS50-C	Video	64.35±0.02	65.57±0.01	**66.20±0.08**
KS50-C	Audio	71.82±0.21	72.41±0.20	**72.60±0.18**
KS50-C	Paired	38.45±0.04	41.01±0.03	**42.84±0.07**
VGGSound-C	Video	57.18±0.01	**57.18±0.04**	56.82±0.09
VGGSound-C	Audio	36.34±0.05	36.04±0.06	**36.71±0.05**
VGGSound-C	Paired	20.10±0.03	20.37±0.04	**22.18±0.05**

B ADDITIONAL BATCH-32 RESULT VIEWS

The following plots show the same batch-32 comparison as Table 1 through paired differences and domain win rates.

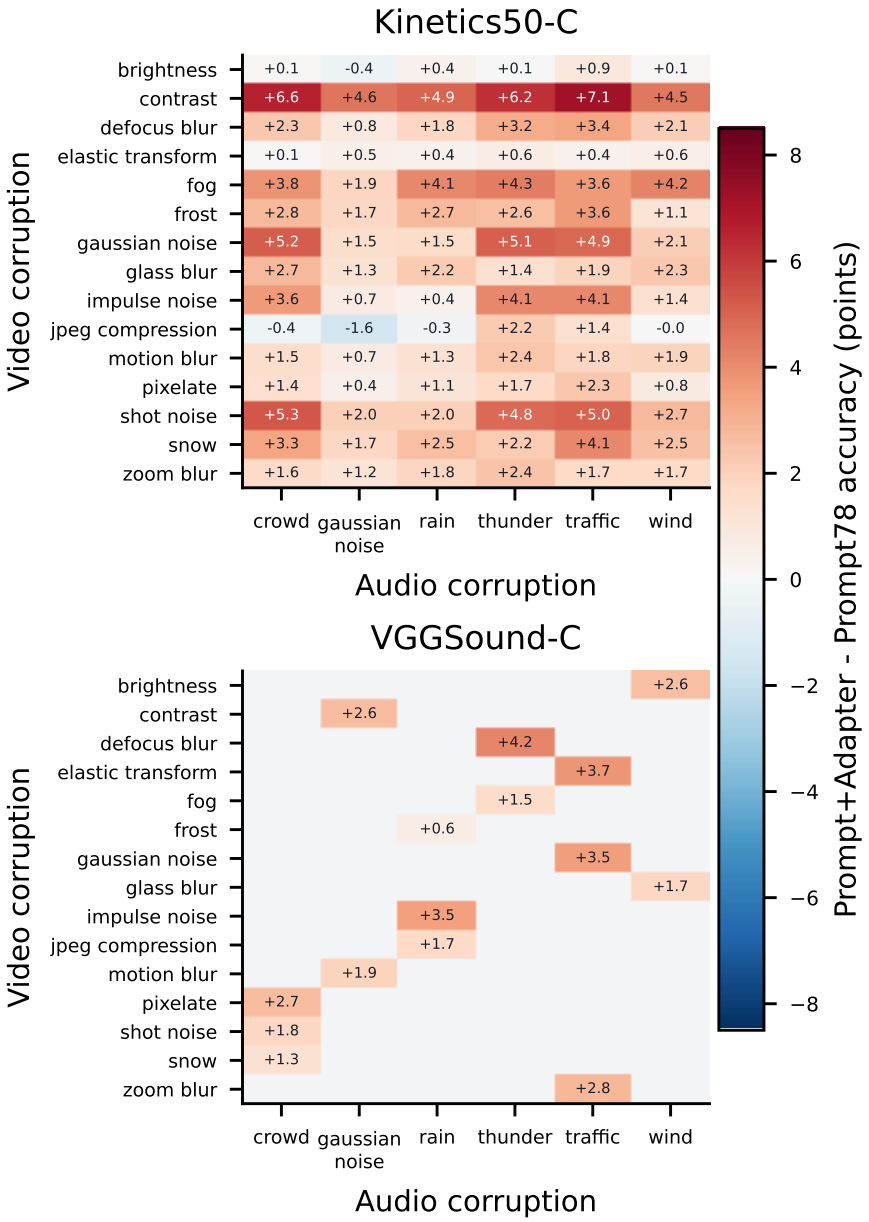

Figure 4: Paired-corruption accuracy differences between Prompt+Adapter and Prompt78 under the batch-32 protocol. Positive values favor Prompt+Adapter. VGGSound-C evaluates each video corruption with one fixed audio corruption; unmeasured cells are left blank.

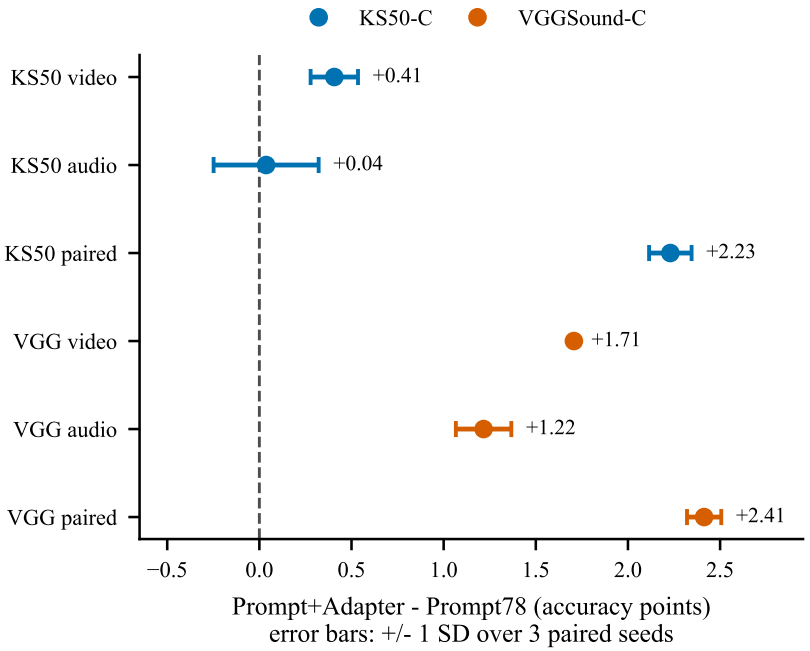

Figure 5: Paired accuracy differences between Prompt+Adapter and Prompt78 for the validated dataset-modality settings. Positive values favor Prompt+Adapter.

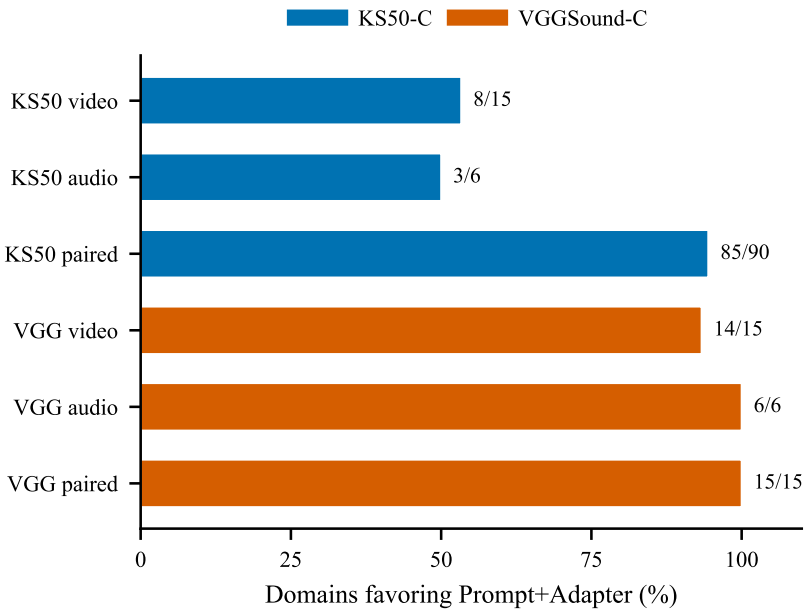

Figure 6: Win rates of Prompt+Adapter against Prompt78 over registered corruption domains for video, audio, and paired shifts.

C Derivation and Scope of the Local Adapter Analysis

This analysis concerns the residual map in Eq. 2. It does not assume Gaussian features or transfer a distribution-estimation theorem from BriMPR to the capacity comparison.

C.1 Identity and gradient activation

At initialization, $U = 0$ and $b_u = 0$, so substitution gives $A(h) = h$ regardless of the value of D. Applying this equality to every adapter preserves the complete model with the same prompts and frozen weights. Changing the number or initialization of prompts is a different operation and is not covered by the equality.

Let $v = Dh + b_d$, $a = \sigma(v)$, and let g be the incoming loss gradient. The differential of the output is

$$dA = dh + (dU)a + US\big((dD)h + D\,dh + db_d\big) + db_u, \qquad S = \mathrm{diag}(\sigma'(v)). \tag{9}$$

Collecting the coefficients in $d\mathcal{L} = g^\top dA$ yields Eq. 7. At $U = 0$, both gradients involving the down projection vanish exactly. A nonzero random D makes nonzero activations available to the up projection, but does not guarantee a nonzero summed gradient: a zero loss gradient or cancellation can still prevent it. Once U changes, $SU^\top g$ can be nonzero. This is the precise sense in which the second backward pass can activate the down projection.

For illustration only, under plain gradient descent with step size η and token index q, the first up-projection step is

$$U_1 = -\eta \sum_q g_{0,q} a_{0,q}^\top, \qquad b_{u,1} = -\eta \sum_q g_{0,q}, \qquad \nabla_D \mathcal{L}_0 = 0. \tag{10}$$

The experiments use Adam rather than this illustrative update. Adam changes the nonzero step magnitudes, but the zero down-projection loss gradient and its conditional activation follow from the same computational graph. No optimization convergence or generalization guarantee follows.

C.2 Local rank and propagation

Holding the adapter parameters fixed and differentiating with respect to a single token gives $J_A = I + USD$. The rank inequality $\mathrm{rank}(XYZ) \le \min(\mathrm{rank}(X), \mathrm{rank}(Y), \mathrm{rank}(Z))$ and submultiplicativity of the spectral norm give Eq. 8. Biases affect S or the output offset but do not add an input-dependent Jacobian term. For a sequence of n tokens with no mixing inside the adapter, the sequence Jacobian is block diagonal; its residual rank can reach nr. Thus the bound r applies to a single token, not to the full sequence, the entire network, or the parameter-space gradient.

For a differentiable downstream map G, Taylor's definition of differentiability yields

$$G(h + \delta h) - G(h) = J_G(h)\delta h + o(\|\delta h\|). \tag{11}$$

If several adapters introduce small perturbations, fix the prompts and linearize around the network with identity adapters. With the adapter perturbations stacked in δ, the first-order output perturbation is

$$\delta y = \sum_\ell J_{>\ell}\, \delta H_\ell + o(\|\delta\|), \tag{12}$$

where δH_ℓ is the sequence perturbation at layer ℓ and $J_{>\ell}$ is the downstream derivative at that reference forward pass. This expression accounts for propagation through the later frozen operations; cross-perturbation effects are absorbed in the remainder. It explains why insertion locations define different local response maps. It neither identifies which location will improve accuracy nor establishes that any empirical placement difference is causal.

C.3 Capacity accounting

Each modality contributes Lpd prompt parameters. Every adapter contributes $rd + r + dr + d$ parameters including both biases. With $2L+1$ adapted blocks, these counts give Eq. 3. For $d = 768$, $L = 11$, and $r = 32$, one adapter has 49,952 parameters, and 23 adapters have 1,148,896. Adding the 168,960 prompt parameters gives 1,317,856. The difference from Prompt78 is $32/1{,}317{,}888 \simeq 2.43 \times 10^{-5}$ of its budget. These are architectural counts, not measurements of memory or runtime.

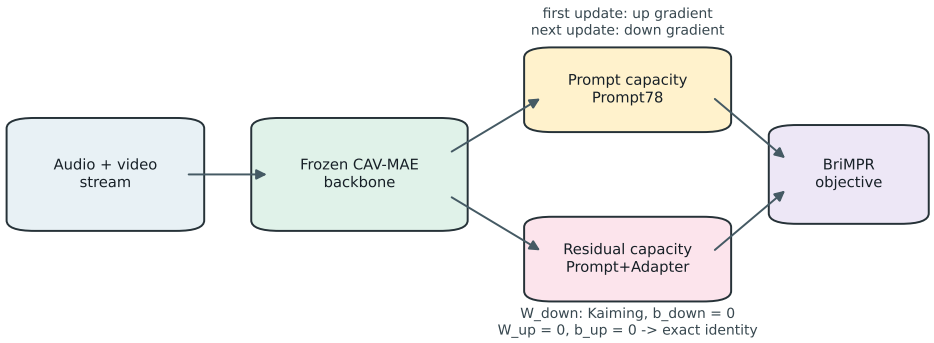

Figure 7: Placement of trainable capacity in the frozen audio-visual backbone. Left: deep prompt capacity inserted into every modality-specific block (Prompt78). Right: a small prompt set retained in every block plus identity-initialized bottleneck adapters attached after each adapted block (Prompt+Adapter). All backbone weights remain frozen in both configurations.

D Capacity Architecture and Initialization Details

Figure 7 shows the placement of the trainable capacity relative to the frozen CAV-MAE backbone. Prompt tokens are inserted into the input sequence of every modality-specific transformer block, while bottleneck adapters are applied sequentially after the self-attention and feed-forward sublayers' output of each adapted block. Under Prompt+Adapter, the 23 adapted blocks are the 11 audio-specific blocks, the 11 visual-specific blocks, and the single multimodal fusion block; the remaining computation, including all attention weights, feed-forward weights, normalization parameters, and the classification head, stays frozen.

D.1 Gradient-flow integrity test

Identity initialization makes the residual branch exactly zero at initialization, which also implies that not every adapter parameter receives a nonzero gradient on the first backward pass. Because W_{up} is initialized to zero, the first update can only move W_{up} and b_{up}; the down projection becomes reachable only after the up projection is nonzero. The executable integrity test asserts this schedule explicitly rather than assuming it:

1. **Identity check.** With the registered initialization, the adapter output equals its input exactly for a batch of random activations (bit-exact equality, not a tolerance).
2. **First-step check.** After one backward pass on the registered objective, the up-projection parameters have nonzero gradients and the down-projection parameters have exactly zero gradients; after the optimizer step, W_{up} and b_{up} are nonzero.
3. **Second-step check.** After the second backward pass, both projections receive nonzero gradients, confirming that the full adapter is trainable from the second update onward.

The test runs standalone against the same initialization code path used by the registered experiments, so a regression in the initialization or the wiring of the residual branch fails the gate before any experiment output can be promoted to the registry.

D.2 Episodic update loop

Figure 8 restates the per-batch update procedure of Section 3 as pseudocode. The same loop applies to every method; only the trainable parameter set ϕ and the method-specific update rule inside its registered wrapper differ. Predictions are always logged before the update on the same batch (test-before-train), and the episodic reset at stream onset restores the trainable parameters, the optimizer state, and the AMP scaler state from the trusted checkpoint.

Episodic online adaptation loop (per corruption domain, per seed)

1. Restore checkpoint: frozen backbone θ, trainable capacity ϕ, optimizer state, and AMP scaler state.
2. For each target batch $x_t = (x_t^a, x_t^v)$ in the stream:
   (a) Forward pass with the selected capacity; record logits, labels, corruption ID, severity, seed, batch index, and stream index.
   (b) Evaluate the fixed BriMPR objective on the same batch.
   (c) Apply the registered update steps to ϕ only (three joint backward steps for the capacity configurations).
3. Aggregate accuracy per domain and seed; independently reaggregate the saved predictions before registry promotion.

Figure 8: Per-batch online update procedure. Steps (a)–(c) record predictions before any gradient is applied on the same batch, so the reported accuracy never benefits from having trained on the evaluated samples.

E Capacity Accounting

Table 5 decomposes the trainable-parameter count of every configuration. Prompt78 allocates 78 tokens per modality per block: $78 \times 2 \times 768 \times 11 = 1{,}317{,}888$. Prompt+Adapter retains 10 tokens per modality per block ($10 \times 2 \times 768 \times 11 = 168{,}960$ parameters) and adds a rank-32 bottleneck adapter on each of the 23 adapted blocks. One adapter carries $768 \times 32 + 32$ parameters in the down projection and $32 \times 768 + 768$ in the up projection, i.e. 49,952 parameters; over 23 blocks this gives 1,148,896 parameters, and the two components total $168{,}960 + 1{,}148{,}896 = 1{,}317{,}856$. The two configurations therefore differ by exactly 32 trainable parameters out of approximately 1.32 million (a relative difference of about 2.4×10^{-5}), while the frozen backbone holds roughly 90 million parameters in both cases. The standalone BriMPR baseline tunes only the 10 deep prompts per modality (168,960 parameters) and uses no adapters.

Table 5: Trainable-parameter accounting. Prompt counts are per modality per block; every prompt token is a 768-dimensional vector; adapters use bottleneck rank r=32 on a 768-dimensional hidden state.

Component	Prompt78	Prompt+Adapter	BriMPR
Prompt tokens per modality per block	78	10	10
Prompt parameters	1,317,888	168,960	168,960
Adapted blocks with adapters	0	23	0
Adapter parameters per block	—	49,952	—
Adapter parameters total	0	1,148,896	0
Trainable parameters	1,317,888	1,317,856	168,960
Relative to Prompt78	100.000%	99.998%	12.82%

We note that a pure zero-prompt adapter configuration cannot close this gap at an integer bottleneck rank: removing the 168,960 prompt parameters would require raising the per-block adapter budget to about 57,299 parameters, which corresponds to a non-integer rank of approximately 36.7. Retaining the 10-prompt base both preserves the prompt-modulated interface on which the BriMPR objective is defined and keeps the comparison within 32 parameters.

F Corruption Domain Taxonomy

The video corruptions follow the fifteen types of the established robustness benchmark of Hendrycks & Dietterich (2019): gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transform, pixelate, and JPEG compression. The audio corruptions are six environment-noise types: gaussian noise, traffic, crowd, rain, thunder,

and wind. All corruptions are applied offline at a fixed severity per evaluation, so every method sees exactly the same corrupted clips.

On KS50-C, the paired setting is the full cross product of the 15 video and 6 audio corruptions, giving 90 paired domains; together with the 15 video-only and 6 audio-only domains this yields 111 domains per seed. On VGGSound-C, the paired setting assigns each of the 15 video corruptions one fixed audio corruption, giving 15 paired domains; the exact assignment is enumerated in Table 10. Domains are evaluated as episodic streams: each domain and seed starts from the same trusted checkpoint, and no state is carried across domains.

G Statistical Testing Procedure

All inferential statements rely on a sign-flip permutation test over per-domain mean differences between Prompt+Adapter and Prompt78. For each corruption domain we first average the three target-stream seeds to obtain one paired accuracy difference per domain. Under the null hypothesis of no systematic placement effect, the sign of each per-domain difference is equally likely to be positive or negative, so the null distribution of the mean difference is obtained by flipping the sign of each per-domain difference independently. For settings with at most 20 domains we enumerate all 2^n sign assignments exactly; for the 90-domain KS50-C paired setting we draw 2×10^5 Monte-Carlo samples from the same null distribution. Reported p-values are two-sided. This test uses the domain as the unit of evidence rather than the seed, which matters because the three target-stream seeds alone would provide little power; the price is that corruption domains are not fully independent experimental units, which we state as a limitation in the main text. Domain-win counts quoted in the text (85/90, 15/15, 14/15, 6/6, 8/15, 3/6) are the raw counts underlying these tests and can be verified cell by cell against Tables 6–10.

H Per-Domain Results

The following tables decompose every aggregate number of Table 1 into per-domain accuracies, generated from the same registry-approved summaries. Bold marks the best adapted method per row in the single-modality tables and the domains where Prompt+Adapter is ahead of Prompt78 in the paired tables.

Table 6: Per-domain severity-5 accuracy (mean ± sample standard deviation over seeds 111/222/333) on KS50-C video-only corruptions under the batch-32 protocol. Source performs no adaptation and has no seed variance. P78 = Prompt78; P+A = Prompt+Adapter.

Corruption	Source	Tent	EATA	SAR	READ	BriMPR	P78	P+A
brightness	75.18	60.48±0.05	75.45±0.08	61.50±0.14	75.21±0.23	76.41±0.21	**76.57±0.08**	76.26±0.10
contrast	51.99	41.80±0.72	52.35±0.20	45.47±0.08	54.60±0.54	58.68±0.15	59.99±0.15	**61.77±0.57**
defocus blur	67.44	53.85±0.20	67.82±0.08	54.62±0.23	67.88±0.12	68.95±0.54	**69.51±0.32**	69.33±0.41
elastic transform	64.84	58.61±0.28	67.40±0.40	59.18±0.16	69.03±0.26	73.20±0.18	74.40±0.04	**74.87±0.34**
fog	45.90	28.82±0.45	49.82±0.69	46.96±0.06	62.53±0.18	66.37±0.23	**68.07±0.05**	67.64±0.21
frost	60.06	52.11±0.23	62.45±0.20	53.77±0.06	65.02±0.11	66.61±0.13	**67.69±0.27**	66.68±0.54
gaussian noise	47.00	36.55±6.79	47.17±0.24	41.05±4.24	50.42±0.11	50.58±0.16	51.59±0.31	**52.99±0.17**
glass blur	62.08	54.57±0.22	64.99±0.28	55.95±0.08	66.07±0.19	69.02±0.35	70.26±0.19	**71.44±0.27**
impulse noise	46.96	31.44±0.38	46.93±0.04	38.10±0.15	50.36±0.55	50.43±0.02	51.84±0.43	**53.28±0.13**
jpeg compression	62.21	51.15±0.10	64.22±0.37	52.50±0.13	66.41±0.28	69.82±0.17	**70.76±0.46**	70.69±0.44
motion blur	70.36	57.29±0.04	71.07±0.17	57.89±0.14	71.18±0.04	71.36±0.13	71.74±0.10	**72.61±0.31**
pixelate	66.55	53.14±0.20	68.30±0.04	54.16±0.13	68.65±0.17	73.56±0.22	**75.01±0.28**	74.60±0.33
shot noise	47.89	31.71±0.03	47.68±0.02	38.58±0.13	50.31±0.45	51.23±0.27	52.61±0.34	**54.22±0.50**
snow	61.80	51.65±0.12	64.57±0.18	53.49±0.06	65.22±0.05	66.80±0.12	**67.63±0.40**	65.99±0.37
zoom blur	66.63	56.43±0.04	68.40±0.16	57.27±0.16	69.69±0.19	69.52±0.25	70.36±0.11	**71.75±0.17**
Mean over domains	59.79	47.97±0.37	61.24±0.05	51.37±0.24	63.51±0.02	65.50±0.09	66.54±0.04	66.94±0.12

Table 7: Per-domain severity-5 accuracy (mean ± sample standard deviation over seeds 111/222/333) on KS50-C audio-only corruptions under the batch-32 protocol. Source performs no adaptation and has no seed variance. P78 = Prompt78; P+A = Prompt+Adapter.

Corruption	Source	Tent	EATA	SAR	READ	BriMPR	P78	P+A
crowd	67.80	11.99±0.10	36.20±0.17	34.37±0.33	70.38±0.22	72.36±0.15	**73.10±0.13**	72.90±0.44
gaussian noise	73.93	56.20±12.51	59.58±10.17	59.52±9.90	74.20±0.24	74.86±0.09	**74.98±0.03**	74.61±0.13
rain	70.36	16.45±0.11	43.77±0.05	43.81±0.05	71.33±0.20	71.56±0.13	**71.95±0.31**	70.87±0.35
thunder	67.88	19.02±0.43	42.82±0.27	42.73±0.15	72.56±0.36	73.48±0.40	74.25±0.28	**74.95±0.19**
traffic	65.49	17.29±0.71	36.29±0.33	35.28±0.26	69.63±0.22	70.10±0.38	71.01±0.10	**71.18±0.29**
wind	70.36	35.10±1.12	47.65±0.14	47.69±0.09	70.84±0.13	71.49±0.32	70.79±0.21	**71.78±0.07**
Mean over domains	69.30	26.01±2.00	44.39±1.67	43.90±1.75	71.49±0.01	72.31±0.18	72.68±0.10	72.72±0.14

Table 8: Per-domain severity-5 accuracy (mean ± sample standard deviation over seeds 111/222/333) on VGGSound-C video-only corruptions under the batch-32 protocol. Source performs no adaptation and has no seed variance. P78 = Prompt78; P+A = Prompt+Adapter.

Corruption	Source	Tent	EATA	SAR	READ	BriMPR	P78	P+A
brightness	58.94	38.68±1.95	58.65±0.04	44.69±0.08	59.22±0.08	57.73±0.07	57.41±0.12	**60.23±0.05**
contrast	53.66	18.44±1.02	54.17±0.07	40.25±0.30	54.77±0.20	51.44±0.32	52.72±0.18	**56.75±0.06**
defocus blur	57.18	41.30±0.12	56.79±0.02	43.07±0.19	56.97±0.17	56.21±0.12	56.05±0.21	**57.90±0.10**
elastic transform	56.99	42.74±0.06	57.96±0.09	44.92±0.07	57.76±0.07	57.68±0.08	57.04±0.34	**59.81±0.08**
fog	55.57	41.14±0.20	56.43±0.05	42.47±0.29	57.21±0.04	**58.06±0.15**	57.74±0.14	57.89±0.12
frost	56.47	40.96±0.09	56.15±0.12	42.34±0.26	57.25±0.11	57.12±0.14	56.57±0.07	**57.31±0.14**
gaussian noise	52.76	33.87±13.23	52.93±0.03	44.01±6.09	52.43±0.20	53.62±0.13	53.49±0.15	**53.68±0.14**
glass blur	57.36	41.96±0.17	56.69±0.07	43.49±0.02	57.23±0.12	56.57±0.19	56.17±0.24	**58.10±0.02**
impulse noise	52.78	24.93±1.73	52.87±0.03	39.92±0.13	52.38±0.15	53.53±0.19	53.70±0.18	**53.98±0.06**
jpeg compression	56.89	41.29±0.11	56.99±0.07	43.12±0.07	56.72±0.19	57.09±0.14	56.88±0.20	**58.56±0.24**
motion blur	58.63	42.57±0.18	58.05±0.13	44.45±0.16	58.59±0.06	57.52±0.40	56.67±0.15	**58.71±0.04**
pixelate	55.80	41.74±0.14	56.78±0.02	43.82±0.04	56.22±0.16	57.15±0.19	56.32±0.21	**59.35±0.04**
shot noise	52.66	28.48±2.78	52.94±0.04	39.79±0.17	52.39±0.05	54.00±0.06	**54.02±0.09**	53.71±0.06
snow	56.27	41.16±0.06	55.71±0.17	42.55±0.05	56.03±0.03	56.51±0.09	55.22±0.15	**57.15±0.14**
zoom blur	57.59	41.99±0.16	57.56±0.05	44.10±0.23	58.12±0.01	57.51±0.14	56.26±0.11	**58.72±0.12**
Mean over domains	55.97	37.42±1.19	56.04±0.02	42.87±0.37	56.22±0.06	56.12±0.07	55.75±0.05	57.46±0.04

Table 9: Per-domain severity-5 accuracy (mean ± sample standard deviation over seeds 111/222/333) on VGGSound-C audio-only corruptions under the batch-32 protocol. Source performs no adaptation and has no seed variance. P78 = Prompt78; P+A = Prompt+Adapter.

Corruption	Source	Tent	EATA	SAR	READ	BriMPR	P78	P+A
crowd	16.79	0.47±0.02	3.57±0.11	1.94±0.01	19.09±0.63	36.32±0.38	35.48±0.30	**36.66±0.13**
gaussian noise	37.11	2.93±2.22	23.77±11.91	20.13±1.96	26.60±0.51	**37.60±0.11**	36.70±0.31	37.36±0.07
rain	21.63	0.53±0.03	3.07±0.04	4.02±0.52	20.55±0.11	**31.66±0.09**	30.27±0.13	31.30±0.12
thunder	27.28	1.77±0.09	12.39±0.19	15.97±0.05	23.26±1.34	40.61±0.20	39.98±0.17	**40.75±0.21**
traffic	21.14	0.54±0.02	4.66±0.06	3.05±0.52	21.30±0.45	34.69±0.31	33.72±0.17	**35.34±0.25**
wind	25.54	0.68±0.04	7.73±0.05	7.03±1.15	20.20±1.33	33.60±0.11	31.87±0.05	**33.90±0.20**
Mean over domains	24.92	1.15±0.39	9.20±2.01	8.69±0.43	21.83±0.46	35.75±0.14	34.67±0.07	35.88±0.05

Table 10: Per-domain severity-5 accuracy (mean over seeds 111/222/333) on VGGSound-C paired audio-video corruptions under the batch-32 protocol. Each video corruption is evaluated with one fixed audio corruption. Δ is Prompt+Adapter minus Prompt78 computed from unrounded means; bold marks domains where Prompt+Adapter is ahead.

Video corruption	Audio corruption	BriMPR	Prompt78	Prompt+Adapter	Δ
brightness	wind	23.76	23.47	26.05	**+2.58**
contrast	gaussian noise	17.38	19.18	21.78	**+2.60**
defocus blur	thunder	27.84	24.61	28.78	**+4.17**
elastic transform	traffic	24.27	22.12	25.85	**+3.72**
fog	thunder	29.16	27.27	28.81	**+1.54**
frost	rain	16.34	15.48	16.12	**+0.64**
gaussian noise	traffic	8.82	8.42	11.97	**+3.55**
glass blur	wind	19.54	19.08	20.83	**+1.75**
impulse noise	rain	5.89	4.98	8.46	**+3.48**
jpeg compression	rain	18.63	17.30	19.02	**+1.72**
motion blur	gaussian noise	24.68	24.23	26.13	**+1.91**
pixelate	crowd	24.12	23.81	26.46	**+2.65**
shot noise	crowd	13.13	12.67	14.46	**+1.78**
snow	crowd	21.14	20.49	21.79	**+1.30**
zoom blur	traffic	21.08	20.42	23.25	**+2.82**

Table 11: Per-domain severity-5 accuracy (mean over seeds 111/222/333) on KS50-C paired audio-video corruptions under the batch-32 protocol. Δ is Prompt+Adapter minus Prompt78 computed from unrounded means; bold marks domains where Prompt+Adapter is ahead.

Video corruption	Audio corruption	BriMPR	Prompt78	Prompt+Adapter	Δ
brightness	crowd	60.34	61.10	61.19	**+0.09**
brightness	gaussian noise	62.75	63.80	63.40	-0.41
brightness	rain	58.80	59.39	59.75	**+0.35**
brightness	thunder	62.54	63.46	63.52	**+0.05**
brightness	traffic	57.80	58.73	59.60	**+0.87**
brightness	wind	59.46	60.06	60.14	**+0.08**
contrast	crowd	28.87	31.51	38.06	**+6.56**
contrast	gaussian noise	35.13	36.77	41.34	**+4.57**
contrast	rain	26.79	28.85	33.79	**+4.95**
contrast	thunder	29.36	31.95	38.15	**+6.19**
contrast	traffic	25.39	27.82	34.94	**+7.12**
contrast	wind	29.59	31.59	36.06	**+4.47**
defocus blur	crowd	44.07	45.55	47.85	**+2.30**
defocus blur	gaussian noise	49.68	50.43	51.27	**+0.84**
defocus blur	rain	43.54	44.15	45.93	**+1.78**
defocus blur	thunder	46.08	48.30	51.46	**+3.16**
defocus blur	traffic	41.04	42.35	45.72	**+3.37**
defocus blur	wind	44.40	44.85	46.97	**+2.12**
elastic transform	crowd	53.19	55.99	56.06	**+0.07**
elastic transform	gaussian noise	56.62	59.41	59.87	**+0.46**
elastic transform	rain	52.58	54.53	54.89	**+0.37**
elastic transform	thunder	57.12	60.08	60.72	**+0.64**
elastic transform	traffic	51.80	54.50	54.85	**+0.35**
elastic transform	wind	52.14	54.35	54.92	**+0.57**
fog	crowd	41.50	44.11	47.93	**+3.83**
fog	gaussian noise	45.11	47.47	49.32	**+1.85**
fog	rain	38.46	40.52	44.62	**+4.10**
fog	thunder	42.92	45.74	50.08	**+4.34**
fog	traffic	38.55	41.28	44.89	**+3.61**
fog	wind	39.48	41.93	46.12	**+4.19**
frost	crowd	42.74	44.38	47.15	**+2.77**
frost	gaussian noise	46.51	48.69	50.42	**+1.73**
frost	rain	40.58	42.69	45.42	**+2.73**
frost	thunder	46.03	48.32	50.91	**+2.58**
frost	traffic	39.73	41.38	45.03	**+3.65**
frost	wind	41.93	44.21	45.35	**+1.14**
gaussian noise	crowd	15.80	16.94	22.09	**+5.15**
gaussian noise	gaussian noise	23.02	24.28	25.76	**+1.49**
gaussian noise	rain	14.10	14.91	16.44	**+1.53**
gaussian noise	thunder	18.55	19.71	24.78	**+5.07**
gaussian noise	traffic	13.26	14.26	19.17	**+4.91**
gaussian noise	wind	15.94	17.41	19.52	**+2.11**
glass blur	crowd	46.03	48.16	50.89	**+2.73**
glass blur	gaussian noise	50.28	52.62	53.95	**+1.32**
glass blur	rain	45.55	46.32	48.53	**+2.20**
glass blur	thunder	50.49	53.08	54.43	**+1.35**
glass blur	traffic	44.32	46.36	48.23	**+1.87**
glass blur	wind	46.08	47.07	49.38	**+2.31**
impulse noise	crowd	15.31	17.57	21.19	**+3.62**
impulse noise	gaussian noise	22.53	23.98	24.71	**+0.73**
impulse noise	rain	14.18	14.94	15.33	**+0.39**
impulse noise	thunder	18.53	20.01	24.09	**+4.08**
impulse noise	traffic	12.88	14.19	18.25	**+4.06**
impulse noise	wind	16.34	18.17	19.56	**+1.39**
jpeg compression	crowd	46.51	49.41	49.05	-0.35
jpeg compression	gaussian noise	51.00	54.77	53.22	-1.55
jpeg compression	rain	44.73	48.36	48.09	-0.27

continued on next page

(continued)

Video corruption	Audio corruption	BriMPR	Prompt78	Prompt+Adapter	Δ
jpeg compression	thunder	46.51	49.84	52.07	**+2.23**
jpeg compression	traffic	43.25	46.16	47.55	**+1.39**
jpeg compression	wind	45.40	47.86	47.84	-0.03
motion blur	crowd	50.18	51.95	53.42	**+1.47**
motion blur	gaussian noise	55.00	56.20	56.92	**+0.72**
motion blur	rain	49.16	49.84	51.11	**+1.27**
motion blur	thunder	52.70	54.93	57.30	**+2.37**
motion blur	traffic	47.38	49.12	50.95	**+1.82**
motion blur	wind	49.50	50.00	51.89	**+1.89**
pixelate	crowd	53.11	55.72	57.12	**+1.41**
pixelate	gaussian noise	58.53	60.56	60.96	**+0.41**
pixelate	rain	51.51	53.61	54.68	**+1.07**
pixelate	thunder	53.87	57.50	59.18	**+1.68**
pixelate	traffic	50.07	53.18	55.46	**+2.28**
pixelate	wind	53.20	55.15	55.92	**+0.77**
shot noise	crowd	17.57	18.80	24.06	**+5.26**
shot noise	gaussian noise	23.41	24.83	26.86	**+2.03**
shot noise	rain	15.44	16.10	18.11	**+2.01**
shot noise	thunder	19.98	21.25	26.07	**+4.83**
shot noise	traffic	14.19	15.68	20.71	**+5.03**
shot noise	wind	18.03	19.05	21.79	**+2.74**
snow	crowd	39.06	41.36	44.70	**+3.34**
snow	gaussian noise	44.30	46.28	48.01	**+1.73**
snow	rain	37.23	39.29	41.82	**+2.53**
snow	thunder	44.19	46.35	48.53	**+2.18**
snow	traffic	36.25	38.33	42.42	**+4.08**
snow	wind	38.39	40.21	42.69	**+2.47**
zoom blur	crowd	48.47	50.41	52.04	**+1.64**
zoom blur	gaussian noise	53.89	55.10	56.26	**+1.16**
zoom blur	rain	47.74	49.08	50.88	**+1.80**
zoom blur	thunder	52.11	53.47	55.87	**+2.39**
zoom blur	traffic	46.99	48.80	50.49	**+1.69**
zoom blur	wind	48.77	49.72	51.41	**+1.69**

I PER-SEED RESULTS

Table 12 reports the three target-stream seeds individually for every one of the 48 cases of the batch-32 common-batch protocol, so that the seed-level variability behind each mean in Table 1 is directly inspectable. Seed standard deviations are small relative to the reported margins on paired corruptions; the two KS50-C single-modality settings, where the margins are smallest, are exactly the settings the permutation test places at parity.

Table 12: Per-seed severity-5 accuracy (%) for all 48 cases of the batch-32 common-batch protocol (target-stream seeds 111/222/333). Mean ± sample standard deviation is over the three seeds; margins quoted in the text are computed from unrounded means. Source performs no adaptation and is deterministic given the fixed corrupted streams, so its seed entries coincide.

Dataset	Modality	Method	Seed 111	Seed 222	Seed 333	Mean ± std
KS50-C	Video	Source	59.79	59.79	59.79	59.79±0.00
		Tent	48.48	47.61	47.82	47.97±0.37
		EATA	61.29	61.18	61.25	61.24±0.05
		SAR	51.71	51.24	51.16	51.37±0.24
		READ	63.53	63.50	63.49	63.51±0.02
		BriMPR	65.62	65.48	65.41	65.50±0.09
		Prompt78	66.59	66.53	66.49	66.54±0.04
		Prompt+Adapter	67.09	66.79	66.95	66.94±0.12
KS50-C	Audio	Source	69.30	69.30	69.30	69.30±0.00
		Tent	28.82	24.83	24.38	26.01±2.00
		EATA	46.75	43.14	43.27	44.39±1.67
		SAR	46.38	42.65	42.67	43.90±1.75
		READ	71.51	71.49	71.47	71.49±0.01
		BriMPR	72.44	72.42	72.06	72.31±0.18
		Prompt78	72.64	72.82	72.58	72.68±0.10
		Prompt+Adapter	72.68	72.57	72.90	72.72±0.14
KS50-C	Paired	Source	30.69	30.69	30.69	30.69±0.00
		Tent	9.83	9.80	9.76	9.80±0.03
		EATA	19.63	19.66	19.65	19.65±0.01
		SAR	19.61	19.50	19.50	19.54±0.05
		READ	37.55	37.69	37.68	37.64±0.06
		BriMPR	40.67	40.75	40.69	40.70±0.03
		Prompt78	42.55	42.50	42.57	42.54±0.03
		Prompt+Adapter	44.66	44.74	44.91	44.77±0.10
VGGSound-C	Video	Source	55.97	55.97	55.97	55.97±0.00
		Tent	38.82	35.90	37.53	37.42±1.19
		EATA	56.04	56.07	56.02	56.04±0.02
		SAR	43.37	42.48	42.74	42.87±0.37
		READ	56.31	56.18	56.17	56.22±0.06
		BriMPR	56.15	56.18	56.02	56.12±0.07
		Prompt78	55.68	55.79	55.78	55.75±0.05
		Prompt+Adapter	57.40	57.50	57.47	57.46±0.04
VGGSound-C	Audio	Source	24.92	24.92	24.92	24.92±0.00
		Tent	1.70	0.92	0.84	1.15±0.39
		EATA	12.04	7.89	7.68	9.20±2.01
		SAR	9.26	8.58	8.23	8.69±0.43
		READ	21.25	21.86	22.39	21.83±0.46
		BriMPR	35.55	35.81	35.87	35.75±0.14
		Prompt78	34.59	34.76	34.65	34.67±0.07
		Prompt+Adapter	35.95	35.82	35.88	35.88±0.05
VGGSound-C	Paired	Source	8.75	8.75	8.75	8.75±0.00
		Tent	0.52	0.53	0.53	0.53±0.00
		EATA	4.26	4.22	4.21	4.23±0.02
		SAR	2.93	2.92	2.92	2.92±0.00
		READ	12.14	12.33	12.19	12.22±0.08
		BriMPR	19.77	19.73	19.66	19.72±0.04
		Prompt78	19.02	18.77	18.92	18.90±0.10
		Prompt+Adapter	21.46	21.26	21.23	21.32±0.10

J EARLY-STREAM BEHAVIORAL PROFILE

Table 13 gives the per-seed breakdown of the 25-batch early-stream behavioral check summarized in the main text. The profile covers the gaussian-noise video corruption combined with traffic noise

at severity 5 and batch 32 on KS50-C. Prompt+Adapter leads the profiled prefix on every seed. We reiterate that this is a short stream prefix used to sanity-check that all three configurations update normally at stream onset; it is not a full-stream accuracy measurement and supports no efficiency, latency, or throughput claim.

Table 13: Early-stream behavioral profile: accuracy (%) over the first 25 batches of a paired gaussian-noise $\oplus$ traffic stream at severity 5 and batch 32 on KS50-C, per target-stream seed. This is a stream-prefix behavioral check, not a full-stream accuracy measurement.

Method	Seed 111	Seed 222	Seed 333	Mean $\pm$ std
BriMPR	9.00	10.25	11.12	10.12$\pm$0.87
Prompt78	11.25	11.50	11.62	11.46$\pm$0.16
Prompt+Adapter	10.62	13.00	13.12	12.25$\pm$1.15

K Registered Reproducibility Details

Table 14 records settings for the validated batch-32 evidence. The execution manifests, distributed with the anonymous code package, remain the authoritative source for method-specific arguments and expected prediction counts.

Table 14: Registered settings for the reported evidence.

Setting	Registered value
Target seeds	111, 222, 333
Source-statistics seeds	0, 1, 2 where preregistered
Main severity	5 (ablations: 1, 3)
Main batch size	32 (appendix: 64)
Optimizer and learning rate	Adam, 10^{-4}
Update steps per batch	3, each a single joint backward pass
AMP initial scale	1024
Contrastive temperature τ	0.07
Consistency weights	distance-dependent λ_a, λ_v, recomputed per batch
Source samples	32 instances from the source split, statistics precomputed per seed
Deep prompt layout	tokens inserted in all 11 modality-specific blocks per modality
Prompt78 capacity	$78 \times 2 \times 768 \times 11 = 1{,}317{,}888$ parameters
Prompt+Adapter capacity	10 prompts + r=32 adapters on 23 blocks = 1,317,856
BriMPR baseline capacity	10 deep prompts per modality = 168,960 parameters
Adapter topology	rank 32, GELU, sequential post-block on 11 audio + 11 visual + 1 fusion blocks
Baseline hyperparameters	EATA: Fisher size 2000, α=1.0, ϵ=0.1; SAR: margin $0.40 \ln 50$; READ: γ=e^{-1}; Tent: none
Domains (KS50-C)	15 video, 6 audio, 90 paired
Domains (VGGSound-C)	15 video, 6 audio, 15 paired
Compared methods	Source, Tent, EATA, SAR, READ, BriMPR, Prompt78, Prompt+Adapter
Compute and software	NVIDIA A100-PCIE-40GB; PyTorch 1.13.1 (CUDA 11.7), Python 3.10