CAN HYBRID WEIGHT-SPACE STOCHASTICITY ADVANCE ADVERSARIAL ROBUSTNESS IN MULTIMODAL FUSION?

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce **Hybrid Weight-Space Convolutional Stochasticity** (`HyWeCS`), a unified randomized defense for multimodal fusion that closes the long-standing gap between *provable* ℓ_2 certification and *empirical* robustness against strong ℓ_∞ attacks, while preserving generalization on natural- and medical data. Specifically, we formulate two complementary mechanisms at disjoint weight locations within every filter: **Orthogonally Retracted Random-Projection Weights** (`ORaW`) and **Stochastic Hybrid-Space Attention-Noise Weights** (`SHANW`), jointly exploiting projection diversity and self-modulated attention noise during training and inference. Reciprocal intermediate cross fusion reuses terminal stochastic weights without a separate feature-space stochastic module, thereby saving computational cost, with parameters and `GFLOPs` reduced by up to $\approx 60.5\%$ and $\approx 91.1\%$, respectively. Shared Lipschitz calibration establishes a uniform 2-Lipschitz score bound for formal certification. Across eight multimodal benchmarks, `HyWeCS` improves certified and empirical robustness by up to $\approx 12.4\%$ and $\approx 12.8\%$, respectively, over the strongest competing defenses, with improvements persisting under adaptive attacks. Code — HyWeCS

1 INTRODUCTION

Multimodal fusion architectures have achieved remarkable performance across high-impact applications, including emotion recognition (Fan et al., 2020; Wang et al., 2021) and clinical decision support (Dhar et al., 2026c; Hemker et al., 2024). However, their broader attack surface allows adversaries to perturb multiple modalities, thereby undermining reliability (Cheng et al., 2023; Dhar et al., 2025a). To mitigate this vulnerability, prior work has explored empirical defenses (Yang et al., 2021; Li et al., 2022; Zhang et al., 2026) and certified defenses (Wang et al., 2024; Yang et al., 2024). Although empirical defenses, including adversarial training (Madry et al., 2017), can improve robustness against specific attacks, they may remain vulnerable to adaptive attacks (Athalye et al., 2018; Dhar et al., 2026a). These limitations renew interest in *certified robustness*, which formally guarantees prediction invariance within a specified perturbation range (Wang et al., 2024).

Among certified defenses, Randomized Smoothing[1] (`RS`) has emerged as one of the most scalable certification frameworks for modern vision architectures (Lecuyer et al., 2018; Cohen et al., 2019). Its central limitation is the accuracy–radius trade-off induced by a fixed noise scale: increasing noise can enlarge certified radii but degrade clean accuracy, whereas reducing noise preserves accuracy at the cost of weaker certificates (Dhar et al., 2026d). Adaptive variants (Alfarra et al., 2022; Súkeník et al., 2022; Hong et al., 2022; Wang et al., 2024) partially mitigate this trade-off, yet their robustness can deteriorate under stronger tailored attacks (Dhar et al., 2026a). Moreover, such robustness may not transfer from natural to medical data owing to differences in sample size, feature statistics, acquisition artifacts, and task structure (Ma et al., 2021; Dong et al., 2024).

Recent randomized defenses (Dhar et al., 2026d;a) have broadened convolutional stochasticity to both natural- and medical-domain benchmarks, yet remain focused on unimodal learners. However, their extension to multimodal fusion is nontrivial: adversaries can jointly perturb modalities with distinct geometries, scales, and statistics, limiting naive adaptations (Wang et al., 2024). Beyond this, these defenses confine convolutional stochasticity to a single locus: either non-smoothing-aware

[1]Randomized Smoothing defines a *smoothed* classifier by aggregating base-model predictions under i.i.d. Gaussian input perturbations at inference, yielding certified ℓ_2 robustness (Xia et al., 2024; Dhar et al., 2026d).

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.

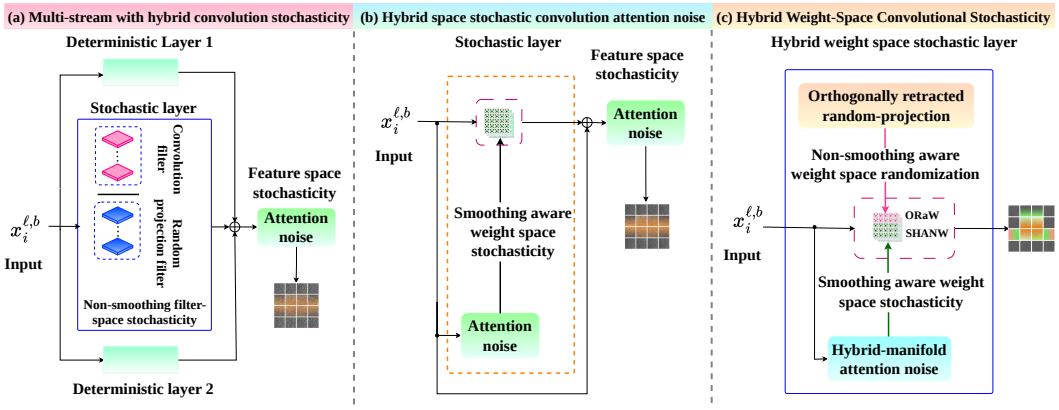


Figure 1: **Randomized-defense comparison.** **(a)** Non-smoothing-aware filter-space randomization (Dhar et al., 2026d). **(b)** Smoothing-aware weight-space randomization (Dhar et al., 2026a). Both focus on unimodal learners, restrict convolutional stochasticity to a single locus, and incur additional feature-space stochastic overhead. **(c)** HyWeCS jointly exploits complementary weight-space stochasticity and reuses terminal weights for reciprocal cross fusion (see Fig. 3(b)), improving robustness at lower cost (Table 1 and Figs. 7–9; App. I.3).

filter-space randomization (Dhar et al., 2026d) (Fig. 1(a)) or attention-noise-induced, smoothing-aware weight-space randomization (Dhar et al., 2026a) (Fig. 1(b)). Their complementary feature-space stochasticity increase computational cost (see Fig. 7–9; App. I.3), thereby limiting suitability for resource-constrained multimodal AI. These observations motivate the following question:

Can a multimodal defense jointly improve certified guarantees and empirical robustness at low computational cost across natural and medical domains?

We answer this question by introducing **Hybrid Weight-Space Convolutional Stochasticity** (HyWeCS), a plug-and-play randomized defense for multimodal fusion. As shown in Figs. 1(c)–2, HyWeCS replaces standard convolutions by incorporating two complementary stochastic mechanisms: **Orthogonally Retracted Random-Projection Weights** (ORaW) and **Stochastic Hybrid-Space Attention-Noise Weights** (SHANW). These mechanisms jointly realize non-smoothing-aware projection diversity and smoothing-aware hybrid-manifold attention noise at disjoint weight locations within each filter. Terminal HyWeCS weights are reused for reciprocal intermediate cross fusion without a separate feature-space stochastic path (see Fig. 3(b)). Under shared Lipschitz calibration, this design yields a globally 2-Lipschitz learner, thereby strengthening certified ℓ_2 and empirical ℓ_∞ robustness across eight natural- and medical-data at low computational cost (see Table 1–4 and Figs. 7–9; App. I.3). The key contributions of this paper can be summarized as follows:

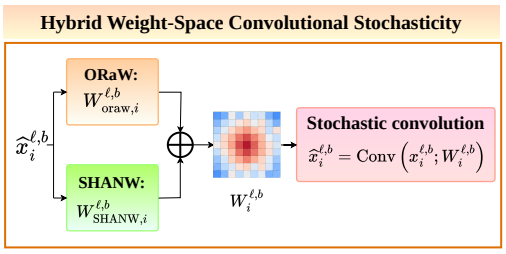


Figure 2: Overview of HyWeCS, comprising two complementary weight-space randomization mechanisms, ORaW and SHANW, that incorporate non-smoothing-aware and smoothing-aware stochasticity, respectively, at disjoint weight locations within each convolutional filter, thereby realizing hybrid weight-space stochasticity.

1. **Hybrid weight-space defense.** We introduce HyWeCS, a randomized hybrid weight-space defense for multimodal fusion that offers certified ℓ_2 guarantees and strong empirical ℓ_∞ robustness across diverse benchmarks at lower computational cost (**Table 1–4 and Figs. 7–9; App. I.3**).
2. **Complementary design.** HyWeCS jointly exploits ORaW and SHANW at disjoint complementary weight locations within every filter and reuses terminal stochastic weights for reciprocal intermediate cross fusion, avoiding a separate feature-space stochastic pathway (**Fig. 2–5; §2.3**).
3. **Theoretical guarantees.** We establish a uniform end-to-end 2-Lipschitz bound together with joint-randomness randomized-smoothing ℓ_2 certification, and provide theoretical support for empirical ℓ_∞ robustness (**Theorems 1–2 and corollary 1**).
4. **Comprehensive evaluation.** Across eight natural- and medical-domain multimodal benchmarks, HyWeCS consistently improves certified and empirical robustness over prior defenses, including

under adaptive attacks (**Fig. 3(c) and Tables 1–4**), while reducing computational cost relative to the compared hybrid defenses (**Figs. 7–9; App. I.3**).

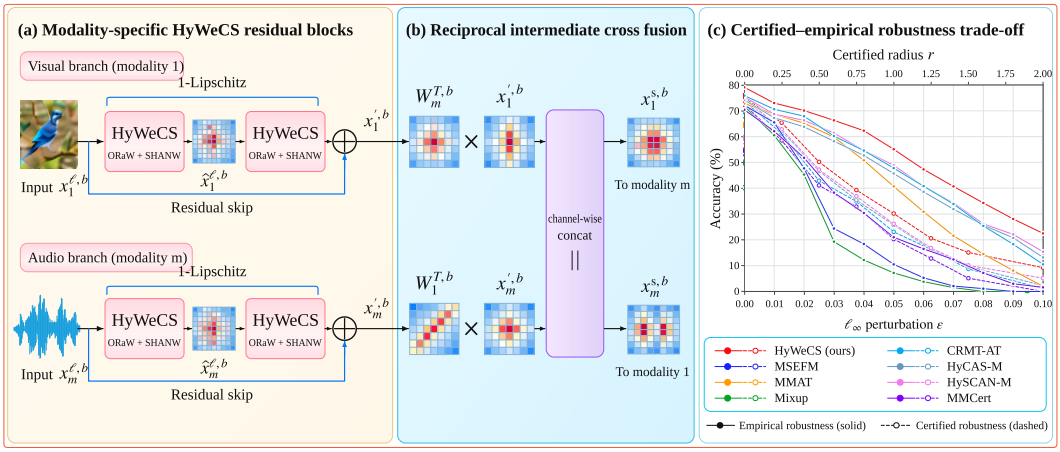


Figure 3: **Overview of the HyWeCS-integrated multimodal fusion learner. (a)** Modality-specific residual blocks replace standard convolutions with `HyWeCS` layers. **(b)** Reciprocal intermediate cross fusion reuses terminal stochastic `HyWeCS` weights to transform paired modality representations without a separate feature-space stochastic module. **(c)** Certified ℓ_2–empirical ℓ_∞ robustness trade-off on `UCF101` (Soomro et al., 2012).

Related work. We include an extensive discussion of related work in Appendix A.

2 Method

2.1 Problem Formulation

Let $X = \{x_i\}_{i=1}^m$ be a multimodal input with m heterogeneous modalities and label y, where $x_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ is an image-like tensor or $x_i \in \mathbb{R}^{D_i}$ is a vector. A multimodal fusion pipeline $\mathcal{F}(\cdot;\beta)$ predicts $\hat{y} = \mathcal{F}(\{x_i^S\}_{i=1}^m;\beta)$ from the shared representations $\{x_i^S\}_{i=1}^m$ and optimizes its loss

$$\mathcal{L}_{\texttt{MML}}(X, y; \beta, \Lambda) = \sum_{i=1}^{m} \lambda_i \, \mathcal{L}_i\big(\mathcal{F}(x_i^S;\beta), y\big), \tag{1}$$

where β denotes trainable parameters and $\Lambda = \{\lambda_i\}_{i=1}^m$ weights the modality-specific losses. We consider white-box adversaries that jointly perturb modalities (Dhar et al., 2026e; Tian & Xu, 2021; Zhang et al., 2022). For $\delta = \{\delta_i\}_{i=1}^m$, define $\|\delta\|_{\infty,\Pi} := \max_i \|\delta_i\|_\infty$. The adversarial objective is

$$\mathcal{L}_{\texttt{adv}}(\beta) = \mathbb{E}_{(X,y)\sim\mathcal{D}}\left[\max_{\|\delta\|_{\infty,\Pi}\le\epsilon} \mathcal{L}_{\texttt{MML}}(X+\delta, y; \beta, \Lambda)\right], \tag{2}$$

where $\mathcal{D}$ is the data distribution. An adversarial example $X^\star = X + \delta^\star$ uses a perturbation $\delta^\star$ that approximately solves the inner maximization.

2.2 Motivation and Overview

Recent hybrid randomized defenses (e.g., `HyCAS` (Dhar et al., 2026d) and `HySCAN` (Dhar et al., 2026a)) have explored complementary stochastic mechanisms for certified and empirical adversarial robustness. Despite their effectiveness, their convolutional stochasticity follows one dominant principle: either Lipschitz-constrained, non-smoothing-aware filter-space randomization (Fig. 1(a)) or attention-noise-induced, smoothing-aware weight-space randomization (Fig. 1(b)). Their joint realization within the weight space of each convolutional filter remains underexplored as a route to stronger adversarial robustness. Instead, their complementarity relies on an additional feature-space stochastic path, increasing computational cost (Fig. 1(a–b) and Figs. 7–9; App. I.3). Beyond this structural limitation, these defenses primarily focus on unimodal learners rather than multimodal fusion. Together, these limitations motivate an intra-weight-space defense that coordinates both heterogeneous stochastic principles within every filter to strengthen multimodal adversarial robustness without the overhead of a separate feature-space stochastic mechanism.

We bridge this gap by introducing **Hybrid Weight-Space Convolutional Stochasticity** (HyWeCS), a plug-and-play stochastic convolutional layer for strengthening adversarial robustness in multimodal fusion (see Figs. 2–3(a-b)). Its defining principle is to realize stochastic complementarity within each n-weight filter through two disjoint weight-generation mechanisms. **(i) Orthogonally Retracted Random-Projection Weights** (ORaW) couple random-projection resampling with Layer-wise Orthogonal Training (LOT) (Xu et al., 2022) to generate weights at r locations, combining stochastic projection diversity with deterministic certificate-compatible geometry. **(ii) Stochastic Attention-Noise Weights** (SHANW) introduce feature-conditioned stochasticity through progressive Euclidean–hyperbolic attention-noise mixing at the remaining $n - r$ locations.

Their disjoint integration coordinates non-smoothing and smoothing-aware stochasticity under shared Lipschitz calibration. Beyond modality-specific encoding, terminal HyWeCS weights are reciprocally reused at each residual-block stage to implement stochastic intermediate cross fusion. This reuse carries the same stochastic realization into cross-modal interaction without a separate feature-space stochastic module.

2.3 Implementation

An overview of the HyWeCS-integrated multimodal learner is shown in Fig. 3; training and adversarial optimization appear in Algs. 1–3 (Apps. F–G). We next formulate its within-filter stochastic convolution, residual-block integration, and reciprocal intermediate cross fusion.

Intra-weight-space allocation: rationale and construction. Consider layer ℓ in residual block b of modality i with kernel size $k_i^{\ell,b} \times k_i^{\ell,b}$ and $C_{\text{in},i}^{\ell,b}$ input channels, each output filter contains $n_i^{\ell,b} = (k_i^{\ell,b})^2 C_{\text{in},i}^{\ell,b}$ weights. Assigning entire filters to either ORaW or SHANW would leave each filter with only one stochastic principle. Their separately generated outputs would therefore be combined only at a later stage, rather than the two principles acting jointly within each filter. Conversely, applying both components at the same weight locations would modify the ORaW values obtained through LOT retraction. We therefore assign $r_i^{\ell,b}$ locations to ORaW and the remaining $n_i^{\ell,b} - r_i^{\ell,b}$ to SHANW. The resulting tensors $W_{\text{oraw},i}^{\ell,b}$ and $W_{\text{SHANW},i}^{\ell,b}$ have fixed complementary supports. This disjoint allocation preserves their distinct roles while jointly realizing orthogonally retracted random-projection stochasticity and progressive hybrid-space attention-noise smoothing within every filter. Their complementarity avoids a separate feature-space stochastic path and its additional computational overhead.

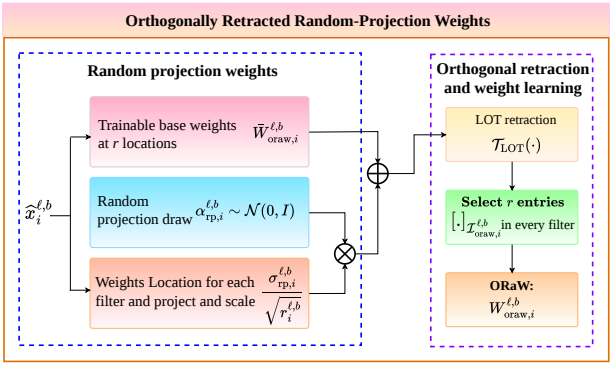


Figure 4: Internal structure of ORaW, showing how random-projection weight resampling is coupled with LOT retraction to learn orthogonally retracted stochastic weights at r weight locations within each convolutional filter.

For each fixed stochastic realization, let $\widehat{L}_i^{\ell,b}$ denote an input-independent upper bound on the Lipschitz constant of the unscaled input-conditioned layer. The HyWeCS weights are

$$W_i^{\ell,b} = \min\left\{1, \frac{\tau_i^{\ell,b}}{\widehat{L}_i^{\ell,b}}\right\}\left(W_{\text{oraw},i}^{\ell,b} + W_{\text{SHANW},i}^{\ell,b}\right), \tag{3}$$

where $0 < \tau_i^{\ell,b} \leq 1$ denotes the layer-wise Lipschitz bound. This formulation integrates the complementary weights under a shared scaling factor. Let $\Omega_i^{\ell,b} = (\alpha_{\text{rp},i}^{\ell,b}, \eta_i^{\ell,b})$ collect the ORaW random-projection and SHANW attention-noise draws. The resulting transformation is

$$\widehat{x}_i^{\ell,b} = \mathcal{G}_i^{\ell,b}\left(x_i^{\ell,b}; \Omega_i^{\ell,b}\right) = \text{Conv}\left(x_i^{\ell,b}; W_i^{\ell,b}\right). \tag{4}$$

Hence, $\text{Lip}_2(\mathcal{G}_i^{\ell,b}) \leq \tau_i^{\ell,b} \leq 1$ for every admissible stochastic draw. This design couples complementary weight-space stochasticity with a Lipschitz bound, supporting the multimodal robustness.

Residual-block realization. Figure 3 shows the mapping of a two-layer HyWeCS residual block

$$x_i'^{,b} = x_i^{\ell,b} + \mathcal{G}_i^{\ell=2,b}\Big(\varsigma\Big(\mathcal{G}_i^{\ell=1,b}(x_i^{\ell,b};\Omega_i^{\ell=1,b})\Big);\Omega_i^{\ell=2,b}\Big), \tag{5}$$

where ς is a non-expansive activation. The identity pathway preserves the incoming representation, while the two HyWeCS layers learn its stochastic residual.

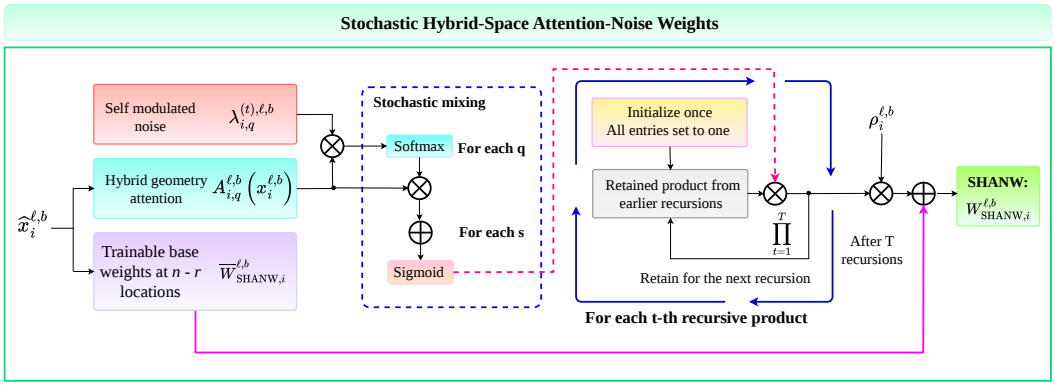


Figure 5: Internal structure of SHANW, showing how independently sampled self-modulated Gaussian noise adaptively fuses Euclidean and hyperbolic attention responses over T recursions to progressively generate stochastic weights at the complementary $n - r$ weight locations within each convolutional filter.

Reciprocal intermediate cross fusion. At stage b, terminal HyWeCS weights $W_i^{T,b}$ and $W_{i+1}^{T,b}$ reciprocally transform the paired representations $x_i'^{,b}$ and $x_{i+1}'^{,b}$ into a stochastic shared representation:

$$x_{i,i+1}^{\mathrm{s},b} = \big(W_i^{T,b}x_{i+1}'^{,b}\big)\|\big(W_{i+1}^{T,b}x_i'^{,b}\big), \tag{6}$$

where $\|$ is concatenation and terminal layer $T = \ell$. The two transformed representations are propagated to modalities i and $i+1$, respectively. This weight reuse extends hybrid weight-space stochasticity to cross-modal interaction without separate feature-space attention noise (**see Fig. 3(b)**).

Joint deterministic–stochastic formulation. Let Θ denote the trainable parameters and Γ_{det} the deterministic structure induced by LOT and Lipschitz scaling. The joint internal draw $\Omega_{\mathrm{H}} = (\boldsymbol{\alpha}_{\mathrm{rp}}, \boldsymbol{\eta}) \sim \nu_{\mathrm{H}}$ collects ORaW random-projection and SHANW attention-noise draws, with input-independent distribution ν_{H}. Let $\widetilde{\boldsymbol{s}}_\Theta(\cdot;\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}})$ denote the unscaled score mapping. Given an input-independent Lipschitz upper bound $\widehat{L}_\Theta > 0$, uniform over internal realizations, we set $\boldsymbol{s}_\Theta = \min\{1, L_\star/\widehat{L}_\Theta\}\widetilde{\boldsymbol{s}}_\Theta$, where $L_\star = 2$ is the global Lipschitz bound used for certification. This positive scaling preserves class-score ordering. For multimodal input X and class set $\mathcal{Y}$, sample $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\sigma^2 I)$ independently of Ω_{H}, where $\sigma > 0$ is the smoothing scale and I is the identity matrix on the vectorized input space. The expected score vector and its associated prediction:

$$\overline{\boldsymbol{s}}_\Theta^{\mathrm{H}}(X) = \mathbb{E}_{\boldsymbol{\varepsilon},\Omega_{\mathrm{H}}}\boldsymbol{s}_\Theta(X+\boldsymbol{\varepsilon};\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}}), \qquad \text{and} \qquad \widehat{y}_\Theta^{\mathrm{H}}(X) = \arg\max_{c\in\mathcal{Y}}\Big[\overline{\boldsymbol{s}}_\Theta^{\mathrm{H}}(X)\Big]_c, \tag{7}$$

where $[\cdot]_c$ selects the class-c score. For probability-based prediction, we aggregate class decisions, hence, the joint-randomness smoothed classifier is

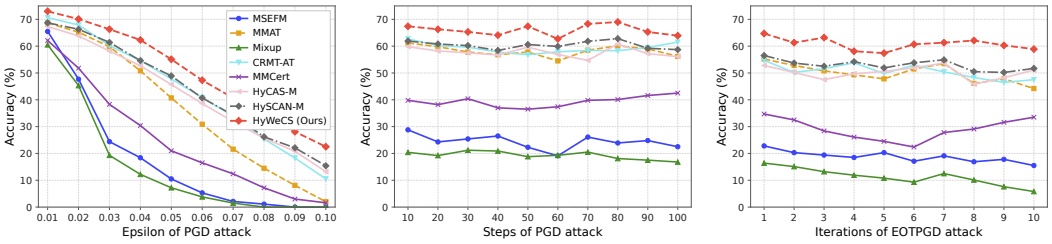
$$p_c^{\mathrm{H}}(X) = \mathbb{P}_{\boldsymbol{\varepsilon},\Omega_{\mathrm{H}}}\Big[\arg\max_{j\in\mathcal{Y}}\boldsymbol{s}_{\Theta,j}(X+\boldsymbol{\varepsilon};\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}}) = c\Big], \qquad \text{and} \qquad g_\Theta^{\mathrm{H}}(X) = \arg\max_{c\in\mathcal{Y}}p_c^{\mathrm{H}}(X), \tag{8}$$

where $p_c^{\mathrm{H}}(X)$ is the class-c prediction probability.

Robustness guarantees. For multimodal inputs X, we define the product ℓ_2 norm

$$\|X - X'\|_{2,\Pi} := \left(\sum_{i=1}^{m}\|\operatorname{vec}(x_i - x_i')\|_2^2\right)^{1/2}, \tag{9}$$

where $\operatorname{vec}(\cdot)$ denotes vectorization. We use $\mathrm{Lip}_\Pi(\cdot)$ for the global Lipschitz constant with respect to this input norm and the Euclidean output norm.



(a) (b) (c)

Figure 6: **Robustness under increasing attack strength on `UCF101` benchmark.** We vary **(a)** the `PGD` perturbation strength ϵ, **(b)** the number of `PGD` steps, and **(c)** the number of `EOT` iterations in `EOTPGD`.

Theorem 1 (Uniform Lipschitz control of `HyWeCS`)**.** *Assume each input-conditioned layer satisfies* $\mathrm{Lip}_2(\mathcal{G}_i^{\ell,b}) \le \tau_i^{\ell,b} \le 1$ *for every admissible draw, and let* $L_\varsigma := \mathrm{Lip}_2(\varsigma) \le 1$*. The residual mapping* $\mathcal{B}_i^b : x_i^b \mapsto x_i'^{,b}$ *then satisfies* $\mathrm{Lip}_2(\mathcal{B}_i^b) \le 1 + \tau_i^{2,b}\tau_i^{1,b}L_\varsigma \le 2$*. If reciprocal cross fusion, including the feature dependence of its reused weights, is non-expansive under the paired product norm, the residual–fusion stage is at most 2-Lipschitz. Given an input-independent Lipschitz upper bound* $\widehat{L}_\Theta$ *on the unscaled network, uniform over internal realizations, global score scaling yields*

$$\sup_{\Omega_{\mathrm{H}}} \mathrm{Lip}_\Pi(\boldsymbol{s}_\Theta(\cdot;\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}})) \le L_\star, \qquad \mathrm{Lip}_\Pi\left(\overline{\boldsymbol{s}}_\Theta^{\mathrm{H}}\right) \le L_\star. \tag{10}$$

Proof. See App. E.2.

Corollary 1 (Certified ℓ_2 guarantees of `HyWeCS`)**.** *Under Theorem 1, let* $c_{\mathrm{e}}(X) := \widehat{y}_\Theta^{\mathrm{H}}(X)$*. Its expected-logit margin and certified radius are*

$$\Delta_\Theta^{\mathrm{H}}(X) = \left[\overline{\boldsymbol{s}}_\Theta^{\mathrm{H}}(X)\right]_{c_{\mathrm{e}}(X)} - \max_{j \ne c_{\mathrm{e}}(X)} \left[\overline{\boldsymbol{s}}_\Theta^{\mathrm{H}}(X)\right]_j \qquad \textit{and} \qquad r_{\exp}(X) = \frac{\Delta_\Theta^{\mathrm{H}}(X)}{2L_\star}. \tag{11}$$

Thus, $\widehat{y}_\Theta^{\mathrm{H}}(X+\delta) = c_{\mathrm{e}}(X)$ *whenever* $\|\delta\|_{2,\Pi} < r_{\exp}(X)$*. For* $L_\star = 2$*,* $r_{\exp}(X) = \Delta_\Theta^{\mathrm{H}}(X)/4$*. For the joint-randomness smoothed prediction* $c_{\mathrm{s}}(X) := g_\Theta^{\mathrm{H}}(X)$*, we define* $p_A := p^{\mathrm{H}}_{c_{\mathrm{s}}(X)}(X)$ *and* $p_B := \max_{c \ne c_{\mathrm{s}}(X)} p_c^{\mathrm{H}}(X)$*. When* $p_A > p_B$*, randomized smoothing (Cohen et al., 2019) gives*

$$r_{\mathrm{RS}}(X) := \frac{\sigma}{2}\left[\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right], \tag{12}$$

where Φ *is the standard Gaussian cumulative distribution function. Then* $g_\Theta^{\mathrm{H}}(X+\delta) = c_{\mathrm{s}}(X)$ *whenever* $\|\delta\|_{2,\Pi} < r_{\mathrm{RS}}(X)$*. If* $c_{\mathrm{e}}(X) = c_{\mathrm{s}}(X)$*, both prediction rules preserve the shared class for* $\|\delta\|_{2,\Pi} < \min\{r_{\exp}(X), r_{\mathrm{RS}}(X)\}$*.*

Proof. See App. E.3.

Empirical robustness. For a labeled multimodal input (X, y), define the expected loss under Gaussian input noise and joint `ORaW-SHANW` randomness, together with its adversarial counterpart:

$$\overline{\mathcal{L}}_{X,y}(\delta) = \mathbb{E}_{\boldsymbol{\varepsilon},\Omega_{\mathrm{H}}}\left[\mathcal{L}_{\mathrm{MML}}(\boldsymbol{s}_\Theta(X+\delta+\boldsymbol{\varepsilon};\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}}), y)\right], \quad \text{and} \quad \overline{\mathcal{L}}^{\mathrm{adv}}_{X,y} = \sup_{\|\delta\|_{\infty,\Pi} \le \epsilon_\infty} \overline{\mathcal{L}}_{X,y}(\delta), \tag{13}$$

where $\epsilon_\infty \ge 0$ denotes the joint ℓ_∞ perturbation radius.

Theorem 2 (Empirical ℓ_∞ loss bound of `HyWeCS`)**.** *Let* $\mathcal{L}_{\mathrm{MML}}(\cdot, y)$ *be* L_ℓ*-Lipschitz in the Euclidean score norm and* $\overline{\mathcal{L}}_{X,y}$ *be* L_s*-smooth in the vectorized multimodal input. Under Theorem 1, for total input dimension* $D := \sum_{i=1}^m \dim(x_i)$*, then*

$$\overline{\mathcal{L}}^{\mathrm{adv}}_{X,y} \le \overline{\mathcal{L}}_{X,y}(0) + \epsilon_\infty\sqrt{D}\,L_\ell L_\star + \frac{L_s D}{2}\epsilon_\infty^2. \tag{14}$$

Proof. See App. E.4.

This guarantee shows that the Lipschitz control supporting certification also limits adversarial inflation of the multimodal fusion loss under ℓ_∞ perturbations.

Training objective. We train the `HyWeCS`-integrated learner by optimizing

$$\min_\Theta \mathbb{E}_{(X,y)\sim\mathcal{D}}\left[\overline{\mathcal{L}}_{X,y}(0) + \lambda_{\mathrm{adv}}\overline{\mathcal{L}}^{\mathrm{adv}}_{X,y}\right], \tag{15}$$

Table 1: Certified accuracy (%) on the RAVDESS and UCF101 benchmarks at fixed ℓ_2 radii r for different smoothing noise levels σ. Boldface denotes the best result in each column across all noise–radius pairs.

		RAVDESS								UCF101							
Approaches	σ	Certified accuracy at predetermined ℓ_2 radius r (%)								Certified accuracy at predetermined ℓ_2 radius r (%)							
		0.00	0.25	0.50	0.75	1.00	1.25	1.50	2.0	0.00	0.25	0.50	0.75	1.00	1.25	1.50	2.0
CRMT-RS	0.25	75.1	63.8	50.5	36.4	20.8	9.4	1.08	0.0	73.8	61.1	42.9	34.9	20.7	10.5	0.0	0.0
	0.50	63.8	53.5	41.6	30.1	21.7	14.2	3.40	0.0	60.2	51.7	39.9	34.9	23.1	12.8	5.14	0.0
	1.00	45.3	36.7	30.6	26.4	21.1	16.5	10.9	2.12	44.3	35.9	32.6	26.1	21.7	16.3	8.72	1.15
MMCert	0.25	73.7	62.3	48.5	33.9	17.2	5.25	0	0.0	72.4	58.5	41.1	32.9	18.3	7.72	0.0	0.0
	0.50	63.4	49.1	40.2	28.6	19.5	10.3	2.12	0.0	58.7	48.2	38.6	31.9	20.3	10.2	1.15	0.0
	1.00	43.1	35.1	29.2	24.9	20.8	15.2	9.85	1.08	42.8	34.6	31.1	25.7	18.2	12.8	5.14	0.0
HyCAS-M	0.25	76.1	64.8	52.7	38.5	23.9	14.5	1.08	0.0	74.7	62.3	46.7	35.2	21.4	12.3	1.15	0.0
	0.50	65.5	52.9	42.1	33.7	25.1	16.3	8.10	2.12	62.9	52.2	42.1	35.4	25.8	14.7	8.15	1.15
	1.00	46.2	37.7	31.3	26.1	21.4	16.8	10.9	2.16	47.5	37.2	32.6	27.4	22.1	16.3	9.71	2.27
HySCAN-M	0.25	76.4	65.2	53.5	39.1	24.6	15.6	8.10	1.08	74.9	62.8	47.2	35.9	22.4	13.2	2.59	0.0
	0.50	66.3	53.5	42.8	34.9	27.2	17.1	10.9	3.10	63.2	52.7	42.8	36.9	26.2	15.4	9.22	2.27
	1.00	47.4	39.1	32.1	27.2	22.3	17.9	12.5	4.12	47.8	37.9	33.2	27.9	22.7	16.8	10.1	5.14
HyWeCS (Ours)	0.25	**76.8**	**67.9**	**56.1**	**42.8**	**30.9**	**21.7**	**12.5**	**4.12**	**75.1**	**65.4**	**50.2**	**39.3**	**27.6**	**19.5**	**10.1**	**2.59**
	0.50	**67.2**	**58.7**	**47.7**	**40.1**	**32.5**	**24.8**	**15.6**	**8.35**	**65.4**	**54.5**	**46.9**	**37.7**	**30.2**	**20.1**	**12.9**	**6.74**
	1.00	**51.2**	**42.8**	**35.6**	**30.7**	**26.1**	**22.9**	**17.1**	**10.6**	**48.9**	**40.3**	**34.8**	**29.1**	**25.4**	**20.6**	**15.1**	**9.22**

where $\lambda_{\text{adv}} \geq 0$ controls the certified–empirical robustness trade-off, and $\overline{\mathcal{L}}_{X,y}(0)$ denotes the expected loss at $\delta = 0$. Setting $\lambda_{\text{adv}} = 0$ yields certification-compatible training, while positive values incorporate ℓ_∞ adversarial optimization. We next describe how ORaW and SHANW generate their weight-space stochasticity within HyWeCS.

2.3.1 Orthogonally Retracted Random-Projection Weights

Random projections diversify convolutional realizations (Dhar et al., 2026e; Dong & Xu, 2023) but do not inherently enforce orthogonality, whereas deterministic orthogonal tensors (e.g., LOT (Xu et al., 2022)) alone do not provide projection resampling. To couple stochastic projection diversity with certificate-compatible geometry in multimodal fusion, we therefore propose **Orthogonally Retracted Random-Projection Weights** (ORaW), which assigns LOT-retracted random-projection values to $r_i^{\ell,b}$ weight locations within every filter. As shown in Fig. 4, let $\overline{W}_{\text{oraw},i}^{\ell,b}$ denote the trainable base weights and $\alpha_{\text{rp},i}^{\ell,b} \sim \mathcal{N}(0, I)$ the random-projection draw. The $r_i^{\ell,b}$ coefficients assigned to ORaW are

$$W_{\text{oraw},i}^{\ell,b} = \mathcal{T}_{\text{LOT}} \left(\overline{W}_{\text{oraw},i}^{\ell,b} + \frac{\sigma_{\text{rp},i}^{\ell,b}}{\sqrt{r_i^{\ell,b}}} \alpha_{\text{rp},i}^{\ell,b} \right) \Bigg|_{\mathcal{I}_{\text{oraw},i}^{\ell,b}}, \tag{16}$$

where $\sigma_{\text{rp},i}^{\ell,b} \geq 0$ controls the stochastic projection magnitude, $\mathcal{T}_{\text{LOT}}$ denotes the deterministic retraction, and $\mathcal{I}_{\text{oraw},i}^{\ell,b}$ indexes the allocated weight locations. Thus, each stochastic evaluation resamples the projection before retraction and coefficient selection. The retained weights are then integrated with complementary SHANW weights through Eq. 3. Further details appear in Apps. B.2 and C.

2.3.2 Stochastic Hybrid-Space Attention-Noise Weights

Stochastic attention noise incorporates feature-conditioned perturbations into trainable weights. However, restricting this modulation to Euclidean responses confines stochastic smoothing to coordinate-space variation, without explicitly incorporating geodesic relations associated with hierarchical multimodal representations. To diversify weight-space stochasticity by incorporating both geometries, we therefore propose **Stochastic Hybrid-Space Attention-Noise Weights** (SHANW). As shown in Fig. 5, at each recursion, SHANW uses independently sampled, self-modulated Gaussian draws to adaptively fuse Euclidean and hyperbolic attention responses, progressively learning the complementary $n_i^{\ell,b} - r_i^{\ell,b}$ stochastic weights.

Progressive stochastic hybrid-space weights. For each filter, given $x_i^{\ell,b}$, let $A_{i,s}^{\ell,b}(x_i^{\ell,b}) \in [0,1]^{d_i^{\ell,b}}$ denote the weight-aligned attention response learned in space $s \in \mathcal{S} = \{\text{E}, \text{H}\}$ (Euclidean or hyperbolic), where $d_i^{\ell,b} := n_i^{\ell,b} - r_i^{\ell,b}$. As shown in Fig. 5, at recursion t, we independently draw $\eta_{i,s}^{(t),\ell,b} \sim \mathcal{N}(0, I)$ for each geometry, with I the $d_i^{\ell,b}$-dimensional identity, and form the self-

modulated noise as

$$\lambda_{i,s}^{(t),\ell,b} := \eta_{i,s}^{(t),\ell,b} \odot \left(\pi_{i,s}^{\ell,b} + \eta_{i,s}^{(t),\ell,b} \odot \pi_{i,s}^{\ell,b}\right), \tag{17}$$

where $\pi_{i,s}^{\ell,b}$ is a trainable geometry-specific noise scale. With trainable base weights $\overline{W}_{\mathrm{SHANW},i}^{\ell,b} \in \mathbb{R}^{d_i^{\ell,b}}$ and stochastic magnitude $\rho_i^{\ell,b} = \mathrm{SoftPlus}(\rho_{\mathrm{logit},i}^{\ell,b})$, the SHANW weights are

$$W_{\mathrm{SHANW},i}^{\ell,b} = \overline{W}_{\mathrm{SHANW},i}^{\ell,b} + \rho_i^{\ell,b} \prod_{t=1}^{T} \sigma\left(\sum_{s\in\mathcal{S}} \left[\mathrm{softmax}_{q\in\mathcal{S}}\left(A_{i,q}^{\ell,b}(x_i^{\ell,b}) \odot \lambda_{i,q}^{(t),\ell,b}\right)\right]_s \odot A_{i,s}^{\ell,b}(x_i^{\ell,b})\right). \tag{18}$$

Here, T is the recursion depth and σ denotes the sigmoid. Self-modulated noise perturbs the geometry-selection logits, whose softmax weights fuse the two attention responses at each weight location. The sigmoid-modulated mixtures are multiplied elementwise across recursions to form such weights that are integrated with ORaW via Eq. 3. Further details appear in Apps. B.3 and D.

3 Experimental Results

Datasets. We evaluate HyWeCS on eight multimodal benchmarks. The natural-domain suite uses audio–visual modalities for RAVDESS (Livingstone & Russo, 2018), KS (Arandjelovic & Zisserman, 2017), and VGGS (Chen et al., 2020), and RGB–optical-flow modalities for UCF101 (Soomro et al., 2012). The medical-domain suite uses imaging–multi-omics modalities for BRCA (Lingle et al., 2016) and BLCA (Network et al., 2014), and image–tabular modalities for HAM10000 (Tschandl et al., 2018) and SIPaKMeD (Plissiti et al., 2018).

Baselines. Modality-specific encoders use ResNet18 and ResNet50 for natural and medical data, respectively. Baselines include CRMT−RS and CRMT−AT (Yang et al., 2024), MMCert (Wang et al., 2024), MSEFM (Tian & Xu, 2021), MMAT (Li et al., 2022), Mixup (Li et al., 2022), HyCAS−M (Dhar et al., 2026d), and HySCAN−M (Dhar et al., 2026a) (using late-fusion for multimodal adaptations).

Table 2: **Certified robustness on medical-domain multimodal benchmarks.** Certified accuracy (%) on BRCA, BLCA, and HAM10000 across fixed ℓ_2 radii r under different Gaussian smoothing noise levels σ.

		BRCA			BLCA			HAM10000		
Approaches	σ	ℓ_2 radius r			ℓ_2 radius r			ℓ_2 radius r		
		0.0	0.50	1.0	0.0	0.50	1.0	0.0	0.50	1.0
CRMT-RS	0.25	50.9	39.5	10.2	49.6	36.2	4.53	94.2	54.7	12.9
	0.50	44.3	34.9	11.7	41.8	30.1	7.27	88.9	51.4	15.6
	1.0	36.4	30.9	12.5	34.6	27.3	15.9	80.9	56.2	21.8
MMCert	0.25	48.1	36.4	5.54	46.5	27.9	2.35	92.8	49.7	8.42
	0.50	40.9	34.7	10.3	38.5	31.1	8.87	85.3	51.9	15.4
	1.0	34.9	28.6	14.1	28.7	22.4	14.6	74.2	54.4	21.3
HyCAS-M	0.25	50.4	40.1	11.9	48.9	31.6	8.87	94.2	55.4	19.7
	0.50	45.7	36.2	13.4	40.9	33.6	11.8	88.4	55.7	21.3
	1.0	36.8	30.7	17.5	31.2	25.7	18.9	78.1	57.4	25.8
HySCAN-M	0.25	51.7	41.5	12.6	49.4	32.1	9.38	95.2	56.7	21.2
	0.50	46.9	37.4	15.2	42.7	35.1	13.3	90.2	57.4	24.5
	1.0	38.5	32.1	19.7	33.2	27.4	20.1	79.3	59.2	26.5
HyWeCS (Ours)	0.25	**56.6**	**48.1**	**19.1**	**54.9**	**43.4**	**17.2**	**97.9**	**67.8**	**32.4**
	0.50	**50.2**	**42.7**	**22.4**	**48.9**	**38.2**	**21.9**	**93.8**	**62.1**	**35.2**
	1.0	**42.7**	**36.2**	**26.5**	**40.7**	**34.9**	**25.2**	**86.9**	**63.4**	**38.9**

Training details. We measure certified accuracy at ℓ_2 radii with Gaussian smoothing scales $\sigma \in \{0.25, 0.50, 1.0\}$. White-box ℓ_∞ robustness is evaluated at perturbation strength $\epsilon \in \{8/255, 16/255\}$ using adaptive attacks: EOTPGD, BPDA (Athalye et al., 2018; Tramer et al., 2020), and BPDA−EOT (Athalye et al., 2018; Tramer et al., 2020) across all eight benchmarks, with additional PGD (Madry et al., 2017) and AutoAttack (Croce & Hein, 2020) attacks. Such attacks use 20 steps; EOTPGD and BPDA−EOT average gradients over 10 independent stochastic realizations per step. Experiments use five random seeds, with means and standard deviations provided where applicable. Certification training and inference, adversarial training, implementation settings, and evaluation details appear in Apps. F–H.

3.1 Adversarial Robustness

HyWeCS consistently improves the adversarial robustness of multimodal fusion learners across natural and medical domains (Tables 1–4). Compared with the strongest baseline at each matched setting, it improves certified ℓ_2 robustness by up to 7.7% on natural-domain benchmarks and 12.4% on medical-domain benchmarks, with improvements persisting at largest certified radii by up to $\approx 12.4\%$ (Tables 1–2). The improvements also extend to empirical robustness against PGD, AutoAttack, and adaptive BPDA, EOTPGD, and BPDA−EOT attacks. Specifically, at $\epsilon = 8/255$, empirical ℓ_∞ robustness improves by 3.7%–7.3% on natural-domain benchmarks and 2.1%–6.5% on medical-domain benchmarks (Tables 3–4). At the larger perturbation strength $\epsilon = 16/255$, improvements persist across all eight benchmarks, ranging from 4.9% to 11.8% under PGD−EOT and

Table 3: Empirical robustness (accuracy, %) against ℓ_∞ diverse attacks (PGD, AutoAttack, BPDA, and BPDA−EOT) on KS, UCF101, and VGGS datasets at perturbation strength $\epsilon \in \{8/255\}$.

Methods	Clean			AutoAttack			PGD			BPDA			BPDA-EOT		
	KS	UCF101	VGGS	KS	UCF101	VGGS	KS	UCF101	VGGS	KS	UCF101	VGGS	KS	UCF101	VGGS
MSEFM	62.7	72.1	49.2	35.7	27.1	12.7	37.6	24.3	11.5	34.1	21.5	10.4	28.7	18.5	9.7
MMAT	65.6	72.8	50.9	49.2	57.5	41.4	50.7	59.8	40.9	49.8	57.2	39.1	45.4	51.8	33.7
Mixup	66.9	71.7	50.7	31.8	17.8	14.6	32.7	19.2	13.5	25.7	18.7	11.7	22.3	14.5	10.4
CRMT-AT	76.2	75.9	53.8	57.3	59.1	41.4	60.2	60.2	41.4	57.2	55.9	40.5	53.8	51.7	36.2
MMCert	68.7	70.4	48.5	42.1	35.7	26.3	43.9	38.5	29.5	39.1	35.3	24.0	36.2	32.4	21.7
HyCAS-M	71.5	74.2	52.2	52.8	56.1	40.5	49.4	58.1	39.5	47.3	54.1	36.9	42.1	48.5	35.3
HySCAN-M	74.9	75.4	53.8	56.9	58.7	41.4	57.5	60.2	41.4	53.7	55.9	40.5	50.5	51.7	37.1
HyWeCS (Ours)	**78.7**	**79.1**	**55.6**	**62.8**	**64.4**	**45.2**	**65.1**	**66.3**	**46.8**	**63.9**	**64.1**	**44.2**	**60.2**	**59.1**	**41.4**

Table 4: Empirical robustness (accuracy, %) against stronger attacks (e.g., EOTPGD and BPDA−EOT) on multimodal datasets: BRCA BLCA, HAM10000 , and SIPAKMeD at $\epsilon \in \{8/255\}$.

Methods	EOTPGD				BPDA-EOT			
	BRCA	BLCA	HAM10000	SIPAKMeD	BRCA	BLCA	HAM10000	SIPAKMeD
MSEFM	30.1 ± 5.27	29.5 ± 5.81	58.9 ± 6.37	42.2 ± 7.48	26.4 ± 4.97	23.3 ± 6.04	50.9 ± 4.82	40.5 ± 5.47
MMAT	33.5 ± 4.17	32.9 ± 2.91	62.8 ± 3.47	43.9 ± 3.83	30.1 ± 3.42	28.9 ± 3.71	53.7 ± 4.16	45.4 ± 3.54
Mixup	28.9 ± 5.74	25.6 ± 6.81	57.9 ± 6.75	36.8 ± 7.63	24.9 ± 5.41	23.4 ± 4.74	48.5 ± 6.07	34.8 ± 4.51
CRMT-AT	34.6 ± 4.27	33.2 ± 2.75	63.8 ± 3.08	44.9 ± 3.52	31.7 ± 4.12	29.7 ± 3.85	55.2 ± 3.43	42.2 ± 3.10
MMCert	31.7 ± 4.82	30.4 ± 6.51	60.7 ± 5.29	40.7 ± 4.51	27.9 ± 4.74	24.5 ± 5.28	51.8 ± 5.08	38.6 ± 4.15
HyCAS-M	35.9 ± 3.74	34.2 ± 3.12	65.1 ± 2.93	52.5 ± 2.14	33.5 ± 2.42	30.9 ± 2.81	57.9 ± 3.14	49.3 ± 3.81
HySCAN-M	37.2 ± 2.91	36.5 ± 2.87	65.7 ± 2.45	55.9 ± 2.10	34.7 ± 2.15	32.5 ± 2.47	59.8 ± 2.74	53.9 ± 3.39
HyWeCS (Ours)	$\mathbf{41.2 \pm 1.85}$	$\mathbf{40.7 \pm 2.24}$	$\mathbf{67.8 \pm 1.72}$	$\mathbf{62.1 \pm 1.08}$	$\mathbf{38.5 \pm 2.17}$	$\mathbf{36.9 \pm 2.45}$	$\mathbf{63.2 \pm 2.51}$	$\mathbf{60.4 \pm 2.97}$

from 4.7% to 12.8% under BPDA−EOT (Fig. 10; App. I.1). Stronger robustness also accompanies higher clean accuracy than every compared baseline on the natural-domain benchmarks (Table 3).

These improvements support jointly exploiting ORaW's orthogonally retracted projection diversity and SHANW's hybrid-space attention noise. Their coordination diversifies weight-space stochasticity under shared Lipschitz calibration, thereby improving certified and empirical robustness over either component alone and evaluated stochastic replacements (Tables 7 and 8; Apps. J.1–J.2).

3.2 Robustness Across Attack Strengths and Trade-offs

Stability under increasing attack strength. Figure 6 shows that HyWeCS degrades more gradually as the perturbation strength, number of PGD steps, and number of EOT iterations increase, while remaining on the upper performance envelope. Its advantage is therefore not tied to a particular perturbation strength, optimization length, or approximation of the expected attack gradient.

Certified–empirical robustness–efficiency trade-off. Figure 3(c) shows that the HyWeCS integrated learner maintains a favorable certified ℓ_2–empirical ℓ_∞ robustness trade-off on UCF101 benchmarks. These results show that ORaW-driven orthogonal projection diversity and SHANW-driven hybrid-space smoothing reinforce one another under coordinated layer-wise and global 2-Lipschitz calibration, instead of improving one robustness regime at the expense of the other.

Reciprocal intermediate cross fusion reuses terminal stochastic HyWeCS weights for cross-modal interaction without a separate feature-space stochastic module, thereby reducing parameters and GFLOPs by up to 60.5% and 91.1%, respectively, while preserving modality-specific and cross-modal stochasticity. The HyWeCS-integrated multimodal fusion learner therefore achieves a favorable certified and empirical robustness–efficiency trade-off across ResNet18 and ResNet50 backbones, including under $\epsilon \in \{8/255, 16/255\}$ (Figs. 7–10; App. I). ***Additional results, ablations, and extended discussion with asymptotic analysis appear in Apps. I, J, and K, respectively.***

4 Conclusion

We introduced HyWeCS, a hybrid weight-space randomized defense for robust multimodal fusion. By jointly exploiting ORaW and SHANW at complementary weight locations, it strengthens certified and empirical robustness without a separate feature-space stochastic module. Across eight benchmarks, it improves certified and empirical robustness by up to 12.4% and 12.8%, respectively, while reducing parameters and GFLOPs by up to 60.5% and 91.1%. Future work will scale HyWeCS to vision–language models and develop margin certificates for robust cross-modal reasoning.

AI USE STATEMENT

We selected few sentences at a time and asked GPT to check the grammar.

REFERENCES

Abdulmateen Adebiyi, Nader Abdalnabi, Emily Hoffman Smith, Jesse Hirner, Eduardo J Simoes, Mirna Becevic, and Praveen Rao. Accurate skin lesion classification using multimodal learning on the ham10000 dataset. *MedRxiv*, pp. 2024–05, 2024.

Motasem Alfarra, Juan C. Perez, Ali Thabet, Adel Bibi, Philip H. S. Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Alexandre Araujo, Aaron Havens, Blaise Delattre, Alexandre Allauzen, and Bin Hu. A unified algebraic perspective on lipschitz neural networks. *arXiv preprint arXiv:2303.03169*, 2023.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Zhiyuan Cheng, Hongjun Choi, James Liang, Shiwei Feng, Guanhong Tao, Dongfang Liu, Michael Zuzak, and Xiangyu Zhang. Fusion is not enough: Single modal attacks on fusion models for 3d object detection. *arXiv preprint arXiv:2304.14614*, 2023.

Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Self-attention fusion for audiovisual emotion recognition with incomplete data. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 2822–2828. IEEE, 2022.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Adriteyo Das, Vedant Agarwal, and Nisha P Shetty. Comparative analysis of multimodal architectures for effective skin lesion detection using clinical and image data. *Frontiers in Artificial Intelligence*, 8:1608837, 2025.

Joy Dhar, Puneet Goyal, Maryam Haghighat, Nayyar Zaidi, Ferdous Sohel, Bao Q Vo, and KC Santosh. Towards building robust models for unimodal and multimodal medical imaging data. *Information fusion*, pp. 103822, 2025a.

Joy Dhar, Nayyar Zaidi, Maryam Haghighat, Sudipta Roy, Puneet Goyal, Azadeh Alavi, and Vikas Kumar. Multimodal fusion learning with dual attention for medical imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4362–4371, 2025b. doi: 10.1109/WACV61041.2025.00428.

Joy Dhar, Manish Kumar Pandey, Behzad Bozorgtabar, Nayyar Zaidi, Wenyu Zhang, Wei-Hong Li, Tingting Mu, Dwarikanath Mahapatra, Mahsa Baktashmotlagh, Trung Le, et al. Does a hybrid space-aware randomized defense improve empirical and certified adversarial robustness? In *Forty-third International Conference on Machine Learning*, 2026a.

Joy Dhar, Manish Kumar Pandey, Debashis Das Chakladar, Maryam Haghighat, Azadeh Alavi, Sajib Mistry, and Nayyar Zaidi. Hypca-net: Advancing multimodal fusion in medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1831–1840, 2026b.

Joy Dhar, Manish Kumar Pandey, Nayyar Zaidi, Chen Chen, Maryam Haghighat, Ferdous Sohel, and Puneet Goyal. Advancing multimodal fusion on heterogeneous medical data with hybrid geometry attention. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 863–874, 2026c.

Joy Dhar, Song Xia, Manish Kumar Pandey, Maryam Haghighat, Azadeh Alavi, Ferdous Sohel, Wenyu Zhang, and Nayyar Zaidi. Certified vs. empirical adversarial robustness via hybrid convolutions with attention stochasticity. In *The Fourteenth International Conference on Learning Representations*, 2026d. URL `https://openreview.net/forum?id=sYk9GaFHEf`.

Joy Dhar, Nayyar Zaidi, and Maryam Haghighat. Effective and robust multimodal medical image analysis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 188–199, 2026e.

Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57 (3):1–38, 2024.

Minjing Dong and Chang Xu. Adversarial robustness via random projection filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4077–4086, 2023.

Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *European Conference on Computer Vision*, pp. 340–356. Springer, 2020.

Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 588–597, 2019. doi: 10.1109/CVPR.2019.00068.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Healnet: Multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37:64479–64498, 2024.

Hanbin Hong, Binghui Wang, and Yuan Hong. UniCR: Universally approximated certified robustness via randomized smoothing. In *European Conference on Computer Vision (ECCV)*, 2022.

Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5492–5501, 2019.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

Juncheng B Li, Shuhui Qu, Xinjian Li, Po-Yao Bernie Huang, and Florian Metze. On adversarial robustness of large-scale audio visual learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 231–235. IEEE, 2022.

Wilma Lingle, Bradley J. Erickson, Margarita L. Zuley, Rose Jarosz, Ermelinda Bonaccio, Joe Filippini, Jose M. Net, Len Levi, Elizabeth A. Morris, Gloria G. Figler, Pierre Elnajjar, Shanah Kirk, Yueh Lee, Maryellen Giger, and Nicholas Gruszauskas. The cancer genome atlas breast invasive carcinoma collection (tcga-brca), 2016. URL https://www.cancerimagingarchive.net/collection/tcga-brca/.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Saiyue Lyu, Shadab Shaikh, Frederick Shpilevskiy, Evan Shelhamer, and Mathias Lécuyer. Adaptive randomized smoothing: Certified adversarial robustness for multi-step defences. *Advances in Neural Information Processing Systems*, 37:134043–134074, 2024.

Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

Yanxiang Ma, Minjing Dong, and Chang Xu. Adversarial robustness through random weight sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315, 2014.

Maria E Plissiti, Panagiotis Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, Orestis Krikoni, and Avraam Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3144–3148. IEEE, October 2018.

David Restrepo, Chenwei Wu, Sebastián Andrés Cajas, Luis Filipe Nakayama, Leo Anthony Celi, and Diego M López. Multimodal deep learning for low-resource settings: a vector embedding alignment approach for healthcare applications. *arXiv preprint arXiv:2406.02601*, 2024.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2022.

Nusrat Nizam Suzana and Muhammad Nazrul Islam. A multimodal approach for pap-smear image-based classification of cervical cancer. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pp. 1–6. IEEE, 2025.

Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5601–5611, 2021.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Shubham Ugare, Tarun Suresh, Debangshu Banerjee, Gagandeep Singh, and Sasa Misailovic. Incremental randomized smoothing certification. *arXiv preprint arXiv:2305.19521*, 2023.

Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11794–11803, 2021.

Yanting Wang, Hongye Fu, Wei Zou, and Jinyuan Jia. Mmcert: Provable defense against adversarial attacks to multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24655–24664, 2024.

Song Xia, Yi Yu, Xudong Jiang, and Henghui Ding. Mitigating the curse of dimensionality for certified robustness via dual randomized smoothing. In *ICLR*, 2024.

Kun Xiang, Xing Zhang, Jinwen She, Jinpeng Liu, Haohan Wang, Shiqi Deng, and Shancheng Jiang. Toward robust diagnosis: A contour attention preserving adversarial defense for covid-19 detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pp. 2928–2937, 2023.

Xiaojun Xu, Linyi Li, and Bo Li. Lot: Layer-wise orthogonal training on improving l2 certified robustness. *Advances in Neural Information Processing Systems*, 35:18904–18915, 2022.

Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3340–3349, 2021.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*, 2024.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5005–5013, 2022.

Junrui Zhang, Xinyu Zhao, Jie Peng, Chenjie Wang, Jianmin Ji, and Tianlong Chen. Vulnerability-aware robust multimodal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28265–28273, 2026.

APPENDIX

A RELATED WORK

A.1 MULTIMODAL LEARNING UNDER ADVERSARIAL PERTURBATIONS

Multimodal learning jointly exploits heterogeneous information sources to leverage complementary evidence across modalities, with applications spanning audio–visual recognition, scene understanding, sentiment analysis, and clinical decision support (Fan et al., 2020; Kazakos et al., 2019; Chumachenko et al., 2022; Dhar et al., 2025b; 2026b). Despite these advantages, multimodal fusion does not inherently improve adversarial robustness. Perturbing a single modality can compromise the fused prediction, while coordinated perturbations across modalities can further exploit interactions between modality-specific representations and the fusion mechanism (Cheng et al., 2023; Tian & Xu, 2021; Zhang et al., 2022; Wang et al., 2024; Yang et al., 2024). For example, `MSEFM` (Tian & Xu, 2021) shows that audio–visual integration can remain vulnerable under multimodal attacks, whereas `MMAT` (Li et al., 2022) and `CRMT` (Yang et al., 2024) improve robustness through multimodal adversarial training and robustness-aware representation learning. These studies establish multimodal robustness as a distinct problem from simply applying a unimodal defense independently to each modality.

A.2 EMPIRICAL AND STOCHASTIC ADVERSARIAL ROBUSTNESS

Empirical defenses primarily improve robustness against explicitly evaluated attack families, commonly under ℓ_∞ perturbations. Adversarial training remains a dominant strategy (Madry et al., 2017), while complementary approaches introduce regularization or architectural stochasticity. `PNI` (He et al., 2019) learns Gaussian perturbation magnitudes for network weights or activations; `RPF` (Dong & Xu, 2023) replaces a subset of convolutional filters with random-projection filters; and `CTRW` (Ma et al., 2023) optimizes constrained random weights to balance clean and adversarial performance. Domain-specific empirical defenses have also been explored, such as `CAP` (Xiang et al., 2023), which incorporates contour-aware regularization for robust medical-image prediction. These approaches can substantially improve empirical robustness, but their robustness is established through attack-based evaluation rather than a formal prediction certificate.

In multimodal learning, empirical defenses must additionally account for heterogeneous modality statistics and cross-modal interactions. `MSEFM` (Tian & Xu, 2021) introduces an audio–visual defense against multimodal attacks, while `MMAT` (Li et al., 2022) strengthens multimodal prediction through adversarial training and robustness-aware fusion. `CRMT` (Yang et al., 2024) further connects modality preference to multimodal robustness and develops a certifiable robust multimodal training procedure. However, empirical robustness alone does not guarantee prediction invariance against all admissible perturbations within a prescribed region, and stochastic defenses require attacks that explicitly account for their internal randomness (Athalye et al., 2018; Tramer et al., 2020).

A.3 CERTIFIED AND HYBRID RANDOMIZED ROBUSTNESS

Certified defenses provide mathematically verifiable robustness within a specified perturbation set. One direction controls network sensitivity through Lipschitz-constrained parameterizations, including spectral normalization (Miyato et al., 2018), Layer-wise Orthogonal Training (`LOT`) (Xu et al., 2022), and related Lipschitz constructions (Araujo et al., 2023). A complementary direction is randomized smoothing, which constructs a smoothed classifier by aggregating predictions under Gaussian input noise and provides formal ℓ_2 certificates (Cohen et al., 2019; Salman et al., 2019; Jeong & Shin, 2020). Subsequent methods improve the smoothing mechanism, base classifier, or noise adaptation (Carlini et al., 2022; Xia et al., 2024; Ugare et al., 2023; Lyu et al., 2024).

Certified robustness for multimodal learning remains comparatively less explored. `MMCert` (Wang et al., 2024) develops a certified defense specifically for multimodal models and derives guarantees for bounded modality-wise perturbations. `CRMT` (Yang et al., 2024) instead studies the role of modality preference and develops certifiable multimodal training. These formulations establish important multimodal guarantees, but they do not jointly formulate smoothing-aware and non-smoothing-aware stochastic principles within the weights of every convolutional filter while focusing on both certified ℓ_2 and empirical ℓ_∞ robustness.

Table 5: **Comparison of representative adversarial robustness methods.** We compare certified and empirical robustness, smoothing-aware and non-smoothing-aware weight-space stochasticity, separate feature-space stochasticity, learning setting, evaluation domain, and cost effectiveness.

Method	Certified Robustness	Empirical Robustness	Smoothing Aware Weight-Space Stochasticity	Non-Smoothing Aware Weight Space Stochasticity	Separate Feature-Space Stochasticity	Unimodal Learning	Multimodal Learning	Natural Domain	Medical Domain	Cost Effectiveness
Multimodal defenses										
CRMT	✓	✓	–	–	–	–	✓	✓	–	–
MMCert	✓	–	–	–	–	–	✓	✓	–	–
MSEFM	–	✓	–	–	–	–	✓	✓	–	–
MMAT	–	✓	–	–	–	–	✓	✓	–	–
Mixup	–	✓	–	–	–		✓	✓	–	–
Unimodal defenses										
LOT	✓	–	–	–	–	✓	–	✓	–	✓
PNI	–	✓	–	✓	–	✓	–	✓	–	✓
CAP	–	✓	–	–	–	✓	–	–	✓	–
RPF	–	✓	–	✓	–	✓	–	✓	–	✓
CTRW	–	✓	–	✓	–	✓	–	✓	–	✓
HyCAS	✓	✓	–	✓	✓	✓	–	✓	✓	–
HySCAN	✓	✓	✓	–	✓	✓	–	✓	✓	–
HyWeCS (Ours)	✓	✓	✓	✓	–	–	✓	✓	✓	✓

More recently, hybrid randomized defenses have sought to narrow the certified–empirical robustness gap. HyCAS (Dhar et al., 2026d) jointly exploits Lipschitz-constrained convolutional geometry and randomized mechanisms to improve both certified and empirical robustness, whereas HySCAN (Dhar et al., 2026a) incorporates attention-noise-induced stochasticity into its randomized defense. However, their original formulations focus on unimodal learners. Moreover, their stochastic complementarity is realized across distinct network spaces: convolutional stochasticity follows one dominant principle, while an additional feature-space stochastic pathway supplies complementary randomization. This separation increases stochastic-defense overhead and leaves unexplored whether heterogeneous stochastic principles can instead be coordinated directly within the weight space of every convolutional filter.

A.4 Research Gaps

Table 5 highlights the distinction underlying HyWeCS. Rather than obtaining stochastic complementarity through separate weight- and feature-space pathways, HyWeCS jointly exploits two heterogeneous stochastic principles directly within every convolutional filter. ORaW introduces non-smoothing-aware random-projection diversity under LOT-retracted geometry, whereas SHANW introduces smoothing-aware, self-modulated Euclidean–hyperbolic attention noise at the complementary weight locations. Their disjoint integration therefore produces within-filter stochastic complementarity without altering the ORaW coefficients through overlapping perturbations.

The same stochastic realization is subsequently carried into cross-modal interaction by reciprocally reusing terminal HyWeCS weights at intermediate fusion stages. Unlike prior hybrid defenses that introduce an additional stochastic feature-space pathway, this reuse requires neither a separate feature-space stochastic module nor independently parameterized fusion weights. Shared Lipschitz calibration constrains the resulting input-conditioned mapping and supports a uniform end-to-end 2-Lipschitz bound together with joint-randomness randomized-smoothing ℓ_2 certification. The accompanying analysis further provides theoretical support for empirical ℓ_∞ robustness. Accordingly, HyWeCS focuses on certified and empirical robustness within a single multimodal defense while being evaluated across natural- and medical-domain benchmarks (**Table 1–4 and Figs. 7–9; App. I.3**).

B Preliminaries

B.1 Product-Space Lipschitz Calculus

For a multimodal input $X = (x_1, \ldots, x_m)$, channel-wise concatenation is an isometry under the corresponding product norm:

$$\|u_1\|u_2 - v_1\|v_2\|_2^2 = \|u_1 - v_1\|_2^2 + \|u_2 - v_2\|_2^2. \tag{19}$$

Consequently, concatenation itself does not enlarge the Lipschitz constant. Throughout the appendix, an upper bound $\widehat{L}_i^{\ell,b}$ is input-independent for a fixed stochastic realization. Its possible dependence on the realized internal randomness is suppressed to preserve the notation used in the main paper. Once the realization is calibrated through Eq. 3, its resulting layer bound is uniformly at most $\tau_i^{\ell,b}$.

Lemma 1 (Expectation preserves uniform Lipschitzness)**.** *Let $f(\cdot;\omega)$ be L-Lipschitz for every admissible ω, and assume that the expectation $\overline{f}(x) = \mathbb{E}_\omega[f(x;\omega)]$ exists. Then*

$$\mathrm{Lip}_2(\overline{f}) \le L. \tag{20}$$

Proof. For arbitrary x and x',

$$\begin{aligned}\left\|\overline{f}(x) - \overline{f}(x')\right\|_2 &= \left\|\mathbb{E}_\omega\left[f(x;\omega) - f(x';\omega)\right]\right\|_2 \\ &\le \mathbb{E}_\omega\left[\left\|f(x;\omega) - f(x';\omega)\right\|_2\right] \\ &\le L\|x - x'\|_2,\end{aligned}$$

where the first inequality follows from Jensen's inequality. Taking the supremum over $x \neq x'$ proves Eq. 20. □

Lemma 2 (Sensitivity of an input-conditioned convolution)**.** *Let*

$$\mathcal{F}(x) = \mathrm{Conv}\left(x; W(x)\right)$$

be an input-conditioned convolution on a domain $\mathcal{X}$ satisfying $\sup_{x\in\mathcal{X}} \|x\|_2 \le M$. Assume that

$$\sup_{x\in\mathcal{X}} \left\|\mathrm{Conv}\left(\cdot; W(x)\right)\right\|_{2\to 2} \le S \tag{21}$$

and

$$\left\|\mathrm{Conv}\left(\cdot; W(x) - W(x')\right)\right\|_{2\to 2} \le K_W\|x - x'\|_2. \tag{22}$$

Then

$$\mathrm{Lip}_2(\mathcal{F}) \le S + MK_W. \tag{23}$$

Proof. Adding and subtracting $\mathrm{Conv}(x'; W(x))$ gives

$$\begin{aligned}\mathcal{F}(x) - \mathcal{F}(x') = &\ \mathrm{Conv}\left(x - x'; W(x)\right) \\ &+ \mathrm{Conv}\left(x'; W(x) - W(x')\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\left\|\mathcal{F}(x) - \mathcal{F}(x')\right\|_2 &\le S\|x - x'\|_2 + \|x'\|_2 K_W\|x - x'\|_2 \\ &\le \left(S + MK_W\right)\|x - x'\|_2.\end{aligned}$$

Taking the supremum over the domain proves Eq. 23. □

B.2 Layer-Wise Orthogonal Training

Layer-wise Orthogonal Training (LOT) (Xu et al., 2022) parameterizes an orthogonal or semi-orthogonal convolution from an unconstrained latent tensor. Let

$$V \in \mathbb{R}^{C_{\mathrm{out}}\times C_{\mathrm{in}}\times k\times k}$$

be the latent tensor and let $\widehat{V}(\omega)$ denote its Fourier response at frequency ω.

Frequency-wise retraction. Under the required rank condition, the frequency response is retracted as

$$\widehat{Q}(\omega) = \begin{cases}\left[\widehat{V}(\omega)\widehat{V}(\omega)^*\right]^{-1/2}\widehat{V}(\omega), & C_{\mathrm{out}} \le C_{\mathrm{in}}, \\ \widehat{V}(\omega)\left[\widehat{V}(\omega)^*\widehat{V}(\omega)\right]^{-1/2}, & C_{\mathrm{out}} > C_{\mathrm{in}},\end{cases} \tag{24}$$

where $(\cdot)^*$ denotes conjugate transpose. The spatial convolutional tensor is reconstructed through

$$\mathcal{T}_{\mathrm{LOT}}(V) = \mathcal{F}^{-1}\left(\left\{\widehat{Q}(\omega)\right\}_\omega\right). \tag{25}$$

The first case satisfies $\widehat{Q}(\omega)\widehat{Q}(\omega)^* = I$, whereas the second satisfies $\widehat{Q}(\omega)^*\widehat{Q}(\omega) = I$. Thus, the resulting frequency-wise response is orthogonal or semi-orthogonal according to the channel dimensions.

Differentiable inverse square root. For $A = \widehat{V}\widehat{V}^*$ or $A = \widehat{V}^*\widehat{V}$, the inverse square root is evaluated using

$$Y_{k+1} = \frac{1}{2} Y_k \left(3I - Z_k Y_k\right), \qquad Z_{k+1} = \frac{1}{2} \left(3I - Z_k Y_k\right) Z_k, \tag{26}$$

initialized with $Y_0 = A$ and $Z_0 = I$ after rescaling the latent tensor so that $\|I - A\|_2 < 1$. Under this condition, $Z_k \to A^{-1/2}$. The input and kernel are zero-padded before the Fourier transformation, and the padded output is truncated after the inverse transformation. This prevents the implicit circular boundary interaction of direct Fourier convolution. For finite Newton iterations, the exact orthogonality statement below applies to the limiting retraction. Any numerical approximation error in the implemented realization must be included in the verified $\widehat{L}_i^{\ell,b}$ used by the shared `HyWeCS` calibration.

B.3 Lorentz Hyperbolic Geometry

Let $\mathbb{H}^{d_{\mathrm{H}}}_{\kappa}$ denote the Lorentz manifold of curvature $-\kappa$, where $\kappa > 0$:

$$\mathbb{H}^{d_{\mathrm{H}}}_{\kappa} = \left\{ z \in \mathbb{R}^{d_{\mathrm{H}}+1} : \langle z, z \rangle_{\mathcal{L}} = -\kappa^{-1},\ z_0 > 0 \right\}, \tag{27}$$

with Lorentz inner product

$$\langle z, q \rangle_{\mathcal{L}} = -z_0 q_0 + \sum_{j=1}^{d_{\mathrm{H}}} z_j q_j$$

and origin $o_\kappa = (\kappa^{-1/2}, 0, \ldots, 0)$.

For a tangent vector $\overline{v} = (0, v)$ at o_κ, the exponential map and geodesic distance are

$$\mathrm{Exp}^{\kappa}_{o_\kappa}(\overline{v}) = \cosh\!\left(\sqrt{\kappa}\|\overline{v}\|_{\mathcal{L}}\right) o_\kappa + \frac{\sinh(\sqrt{\kappa}\|\overline{v}\|_{\mathcal{L}})}{\sqrt{\kappa}\|\overline{v}\|_{\mathcal{L}}} \overline{v}, \qquad d_\kappa(z, q) = \frac{1}{\sqrt{\kappa}} \operatorname{arcosh}\left(-\kappa \langle z, q \rangle_{\mathcal{L}}\right). \tag{28}$$

We use the continuous extension $\mathrm{Exp}^{\kappa}_{o_\kappa}(0) = o_\kappa$.

B.4 Gaussian Smoothing with Auxiliary Internal Randomness

The standard Gaussian-smoothing argument remains valid when the base classifier contains additional internal randomness, provided that this randomness is sampled independently of the adversarial input displacement.

Lemma 3 (Gaussian shift with auxiliary randomness)**.** *Let $h(z; \omega) \in \mathcal{Y}$ be a randomized classifier, where $\omega \sim \nu$ is independent of $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. For any measurable event $\mathcal{A} \subseteq \mathbb{R}^D \times \Omega$, define*

$$P_X(\mathcal{A}) = \mathbb{P}_{\varepsilon,\omega}\left[(X + \varepsilon, \omega) \in \mathcal{A}\right].$$

For every displacement δ,

$$P_{X+\delta}(\mathcal{A}) \geq \Phi\left(\Phi^{-1}\left(P_X(\mathcal{A})\right) - \frac{\|\delta\|_2}{\sigma}\right), \tag{29}$$

and

$$P_{X+\delta}(\mathcal{A}) \leq \Phi\left(\Phi^{-1}\left(P_X(\mathcal{A})\right) + \frac{\|\delta\|_2}{\sigma}\right). \tag{30}$$

Proof. Let Q_X denote the joint distribution of $(X + \varepsilon, \omega)$. Because the internal law ν is unchanged by the input shift, the Radon–Nikodym derivative satisfies

$$\frac{dQ_{X+\delta}}{dQ_X}(z, \omega) = \exp\left(\frac{\langle \delta, z - X \rangle}{\sigma^2} - \frac{\|\delta\|_2^2}{2\sigma^2}\right).$$

The likelihood ratio is therefore independent of ω and coincides with the likelihood ratio between two Gaussian distributions whose means differ by δ. The Neyman–Pearson Gaussian-shift argument consequently applies on the augmented sample space $\mathbb{R}^D \times \Omega$ without modification, yielding Eqs. 29–30. □

C Additional Details of ORaW

This section details the random-projection sampling, LOT retraction, and within-filter allocation in Eq. 16. We first describe the weight construction, then establish the orthogonal geometry, radial invariance, and realization-wise 1-Lipschitz property of the full retracted realization.

C.1 Construction of LOT-Retracted Random-Projection Weights

Base weights and within-filter allocation. For a fixed layer, we suppress the indices (i, ℓ, b). Let C_{in} and C_{out} denote its input- and output-channel counts, k its kernel width, and $n = k^2 C_{\text{in}}$ the number of coefficients in each output filter. The hybrid construction assigns $1 \le r < n$ locations per filter to ORaW and the remaining $n - r$ to SHANW. For output channel $o \in \{1, \dots, C_{\text{out}}\}$, let $\mathcal{I}_o \subset \{1, \dots, n\}$ contain its r allocated ORaW locations. These sets specify the per-filter restriction represented by $\mathcal{I}^{\ell,b}_{\text{oraw},i}$ in the main paper and remain fixed across stochastic evaluations.

The formulation considered here uses trainable base weights $\overline{W}_{\text{oraw}} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k \times k}$. The allocation determines which retracted coefficients enter each final filter; it does not restrict the trainable base weights to r entries per filter.

Random-projection weight generation. At each independently sampled evaluation, draw α_{rp} with the same dimensions as $\overline{W}_{\text{oraw}}$ and independent standard Gaussian entries. The sampled weight tensor is

$$V = \overline{W}_{\text{oraw}} + \frac{\sigma_{\text{rp}}}{\sqrt{r}} \alpha_{\text{rp}}, \qquad \text{vec}(\alpha_{\text{rp}}) \sim \mathcal{N}(0, I_{C_{\text{out}} n}), \tag{31}$$

where $\sigma_{\text{rp}} \ge 0$ controls the projection magnitude, $\text{vec}(\cdot)$ denotes vectorization, and I_d is the $d \times d$ identity matrix. The factor $1/\sqrt{r}$ follows Eq. 16, giving each perturbation coordinate variance σ^2_{rp}/r. The base weights are optimized during training, whereas α_{rp} is sampled rather than optimized. In this formulation, coefficient selection follows the retraction.

Frequency-wise LOT retraction. Following LOT (Xu et al., 2022), embed V on the configured Fourier grid and compute its convolutional frequency response $\widehat{V}(\omega) \in \mathbb{C}^{C_{\text{out}} \times C_{\text{in}}}$, where ω indexes a spatial frequency. At each frequency, form the dimension-compatible Gram matrix

$$A_\omega = \begin{cases} \widehat{V}(\omega)\widehat{V}(\omega)^*, & C_{\text{out}} \le C_{\text{in}}, \\ \widehat{V}(\omega)^*\widehat{V}(\omega), & C_{\text{out}} > C_{\text{in}}. \end{cases} \tag{32}$$

Here, $(\cdot)^*$ denotes conjugate transpose. We require $\widehat{V}(\omega)$ to have full row rank in the first case and full column rank in the second, ensuring that A_ω is positive definite. The retracted response is

$$\widehat{Q}(\omega) = \begin{cases} A_\omega^{-1/2}\widehat{V}(\omega), & C_{\text{out}} \le C_{\text{in}}, \\ \widehat{V}(\omega)A_\omega^{-1/2}, & C_{\text{out}} > C_{\text{in}}, \end{cases} \tag{33}$$

which applies Eq. 24 to the sampled weight tensor. The respective responses satisfy $\widehat{Q}(\omega)\widehat{Q}(\omega)^* = I_{C_{\text{out}}}$ and $\widehat{Q}(\omega)^*\widehat{Q}(\omega) = I_{C_{\text{in}}}$. Thus, the retraction produces orthonormal rows or columns at each frequency rather than normalizing selected coefficients independently.

Differentiable inverse-square-root computation. An explicit rescaled form of the iteration in Eq. 26 is obtained by setting $a_\omega = \|A_\omega\|_{\text{F}} > 0$, $B_\omega = A_\omega/a_\omega$, $Y_0 = B_\omega$, and $Z_0 = I$, where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm and I has the dimensions of A_ω. For each frequency, perform J iterations:

$$\begin{aligned} Y_{j+1} &= \tfrac{1}{2} Y_j(3I - Z_j Y_j), \\ Z_{j+1} &= \tfrac{1}{2}(3I - Z_j Y_j) Z_j, \qquad j = 0, \dots, J-1. \end{aligned} \tag{34}$$

Positive definiteness and this rescaling place every eigenvalue of B_ω in $(0, 1]$. Hence, $\|I - B_\omega\|_2 < 1$ and $Z_j \to B_\omega^{-1/2}$, giving

$$A_\omega^{-1/2} \approx a_\omega^{-1/2} Z_J. \tag{35}$$

Substitution into Eq. 33 yields the finite-iteration response $\widehat{Q}_J(\omega)$. The rescaled iteration approximates the same inverse-square-root operation, with J specifying its approximation depth. Exact

geometric statements apply to the limiting retraction; finite-iteration and arithmetic errors must be accounted for in the bound used by the shared HyWeCS calibration. A singular Gram matrix violates the rank assumption. Any diagonal regularization changes the retraction and must be specified separately.

Spatial reconstruction and coefficient allocation. Using the inverse Fourier convention of Eq. 25, reconstruct the full tensor

$$Q = \mathcal{T}_{\mathrm{LOT}}(V) = \mathcal{F}^{-1}\Big(\{\widehat{Q}(\omega)\}_{\omega}\Big). \tag{36}$$

Here, $\mathcal{F}^{-1}$ denotes inverse Fourier reconstruction, and Q is the full realization $W_{\mathrm{oraw},i}^{\ell,b,\mathrm{LOT}}$ used in Corollary 2. A finite-iteration implementation instead reconstructs Q_J from $\widehat{Q}_J$. For real V, the Fourier responses and their retractions preserve conjugate symmetry, so the exact inverse transform is real.

To specify the restriction in Eq. 16, let $\iota(a)$ map local filter index $a \in \{1, \ldots, n\}$ to its input-channel and spatial-lag coordinates in Q. This fixed map specifies the kernel origin and its $k \times k$ support on the reconstruction grid. Representing each output filter as a vector of length n, the allocated ORaW weights are

$$W_{\mathrm{oraw}}[o, a] = \begin{cases} Q[o, \iota(a)], & a \in \mathcal{I}_o, \\ 0, & a \notin \mathcal{I}_o. \end{cases} \tag{37}$$

The tensor is then reshaped to $C_{\mathrm{out}} \times C_{\mathrm{in}} \times k \times k$. This retains exactly r assigned coefficients per filter and zero-fills the complementary locations for addition with W_{SHANW}. The Fourier-domain retraction need not preserve the original spatial kernel support. Accordingly, ι specifies coefficient extraction rather than following from orthogonality. This operation is distinct from truncating a padded output feature map.

Integration and stochastic execution. The retained weights enter Eq. 3 together with W_{SHANW}, whose nonzero entries occupy the complementary locations. The shared scaling constrains the integrated input-conditioned mapping rather than relying on orthogonality of the assembled kernel.

For a fixed Gaussian draw, gradients with respect to $\overline{W}_{\mathrm{oraw}}$ pass through the deterministic retraction and fixed coefficient selection. New independent evaluations resample the weight tensor while retaining the learned parameters and allocation. Within each stochastic forward pass, the same terminal HyWeCS weights are reused for reciprocal intermediate cross fusion. Certification-compatible sampling and attack-phase reuse follow Algorithms 1–3.

C.2 Geometric Properties of the Full Retracted Realization

We establish the orthogonal geometry, radial invariance, and realization-wise 1-Lipschitz property of the full LOT-retracted tensor constructed in App. C.1. These statements concern the exact retraction before within-filter coefficient selection.

Proposition 1 (Orthogonal geometry and radial invariance of the full ORaW realization)**.** *Let V be the sampled weight tensor in Eq. 31, satisfying the frequency-wise rank conditions in App. C.1. For the LOT retraction defined by Eqs. 24–25, the response at every frequency ω satisfies*

$$\begin{cases} \widehat{Q}(\omega)\widehat{Q}(\omega)^* = I_{C_{\mathrm{out}}}, & C_{\mathrm{out}} \le C_{\mathrm{in}}, \\ \widehat{Q}(\omega)^*\widehat{Q}(\omega) = I_{C_{\mathrm{in}}}, & C_{\mathrm{out}} > C_{\mathrm{in}}. \end{cases} \tag{38}$$

Moreover, for every scalar $a > 0$,

$$\mathcal{T}_{\mathrm{LOT}}(aV) = \mathcal{T}_{\mathrm{LOT}}(V). \tag{39}$$

Proof. Fix a frequency and suppress ω. For $C_{\mathrm{out}} \le C_{\mathrm{in}}$, the retraction is $\widehat{Q} = (\widehat{V}\widehat{V}^*)^{-1/2}\widehat{V}$, giving

$$\widehat{Q}\widehat{Q}^* = (\widehat{V}\widehat{V}^*)^{-1/2}(\widehat{V}\widehat{V}^*)(\widehat{V}\widehat{V}^*)^{-1/2} = I_{C_{\mathrm{out}}}.$$

For $C_{\mathrm{out}} > C_{\mathrm{in}}$, we have $\widehat{Q} = \widehat{V}(\widehat{V}^*\widehat{V})^{-1/2}$, and hence

$$\widehat{Q}^*\widehat{Q} = (\widehat{V}^*\widehat{V})^{-1/2}(\widehat{V}^*\widehat{V})(\widehat{V}^*\widehat{V})^{-1/2} = I_{C_{\mathrm{in}}}.$$

For $a > 0$, the Fourier response of aV is $a\widehat{V}$. In the first case,

$$\begin{aligned}(a^2\widehat{V}\widehat{V}^*)^{-1/2}(a\widehat{V}) &= a^{-1}(\widehat{V}\widehat{V}^*)^{-1/2}(a\widehat{V}) \\ &= (\widehat{V}\widehat{V}^*)^{-1/2}\widehat{V}.\end{aligned}$$

Similarly, in the second case,

$$(a\widehat{V})(a^2\widehat{V}^*\widehat{V})^{-1/2} = \widehat{V}(\widehat{V}^*\widehat{V})^{-1/2}.$$

Thus, positive radial scaling leaves every retracted frequency response unchanged. Inverse Fourier reconstruction then proves Eq. 39. □

Corollary 2 (Uniform 1-Lipschitzness of the full ORaW realization)**.** *Under Proposition 1, let* $W_{\mathrm{oraw},i}^{\ell,b,\mathrm{LOT}} = \mathcal{T}_{\mathrm{LOT}}(V)$ *denote the full retracted tensor. For the zero-padded, output-truncated convolution described in App. B.2, every admissible random-projection draw satisfies*

$$\left\|\mathrm{Conv}\left(x; W_{\mathrm{oraw},i}^{\ell,b,\mathrm{LOT}}\right) - \mathrm{Conv}\left(x'; W_{\mathrm{oraw},i}^{\ell,b,\mathrm{LOT}}\right)\right\|_2 \le \|x - x'\|_2 \tag{40}$$

for all compatible inputs x *and* x'.

Proof. Proposition 1 gives $\|\widehat{Q}(\omega)\|_2 \le 1$ at every frequency. Let $\mathcal{P}$ denote zero padding, $\mathcal{R}$ output truncation, and $\mathcal{F}$ the unitary Fourier transform on the feature space. In the expression below, $\widehat{Q}$ applies $\widehat{Q}(\omega)$ independently at each frequency. Since zero padding is isometric and truncation is non-expansive,

$$\begin{aligned}&\left\|\mathrm{Conv}\left(x - x'; W_{\mathrm{oraw},i}^{\ell,b,\mathrm{LOT}}\right)\right\|_2 \\ &\quad= \left\|\mathcal{R}\mathcal{F}^{-1}\widehat{Q}\mathcal{F}\mathcal{P}(x - x')\right\|_2 \\ &\quad\le \left\|\widehat{Q}\mathcal{F}\mathcal{P}(x - x')\right\|_2 \\ &\quad\le \|\mathcal{F}\mathcal{P}(x - x')\|_2 = \|x - x'\|_2.\end{aligned}$$

Linearity of the convolution for a fixed draw establishes Eq. 40. □

These guarantees apply to the full retracted realization, before retaining the $r_i^{\ell,b}$ coefficients assigned to ORaW. Coefficient selection and integration with SHANW need not preserve orthogonality or the full-realization operator bound; the assembled mapping is therefore constrained by the shared calibration in Eq. 3. Finite-iteration evaluations remain subject to the approximation conditions stated in App. C.1.

D Additional Details of SHANW

This section details the geometry-specific attention and progressive weight construction in Sec. 2.3.2. We first describe the multi-statistic descriptors and aligned Euclidean and hyperbolic responses, then connect them to Gaussian sampling, self-modulation, geometry selection, and recursive weight generation. We subsequently establish their boundedness, input sensitivity, and compatibility with the shared HyWeCS calibration.

For layer ℓ in block b of modality i, consider

$$x_i^{\ell,b} \in \mathbb{R}^{H_\ell \times W_\ell \times C_{\mathrm{in},i}^{\ell,b}}, \qquad C = C_{\mathrm{in},i}^{\ell,b}, \qquad d_i^{\ell,b} = n_i^{\ell,b} - r_i^{\ell,b}.$$

The set $\mathcal{I}_{\mathrm{SHANW},i}^{\ell,b}$ contains the $d_i^{\ell,b}$ complementary weight locations assigned to SHANW within each filter. Their ordering remains fixed across attention alignment, noise sampling, and weight assembly. Throughout this appendix, σ denotes the sigmoid, $\mathrm{vec}(\cdot)$ denotes vectorization, and $\odot$ denotes elementwise multiplication.

D.1 Multi-Statistic Channel Descriptors

Let $\mathcal{P} = \{\mathrm{GAP}, \mathrm{GMP}, \mathrm{GSP}, \mathrm{GMN}\}$ denote global average, maximum, sum, and minimum pooling. For $x = x_i^{\ell,b}$ and channel $c \in \{1, \dots, C\}$,

$$\begin{aligned} \mathrm{GAP}(x)[c] &= \frac{1}{H_\ell W_\ell} \sum_{h=1}^{H_\ell} \sum_{w=1}^{W_\ell} x[h,w,c], \quad \mathrm{GMP}(x)[c] = \max_{h,w} x[h,w,c], \\ \mathrm{GSP}(x)[c] &= \sum_{h=1}^{H_\ell} \sum_{w=1}^{W_\ell} x[h,w,c], \qquad\qquad \mathrm{GMN}(x)[c] = \min_{h,w} x[h,w,c]. \end{aligned} \tag{41}$$

Each descriptor lies in $\mathbb{R}^C$ and supplies channel information to both geometry branches.

D.2 Euclidean-Space Attention

Local Euclidean response. A single 1×1 convolution learns spatially varying channel responses:

$$L_i^{\mathrm{E},\ell,b} = \mathrm{Conv}_{1\times1,i}^{\mathrm{E},\ell,b}\left(x_i^{\ell,b}\right) \in \mathbb{R}^{H_\ell \times W_\ell \times C}. \tag{42}$$

Multi-statistic Euclidean channel response. Each pooled descriptor is processed by a descriptor-specific projection. Their responses are additively fused and mapped through a shared output projection:

$$C_i^{\mathrm{E},\ell,b} = W_{2,i}^{\mathrm{E},\ell,b} \left[\sum_{p\in\mathcal{P}} \phi_{i,p}^{\mathrm{E},\ell,b}\left(W_{1,i,p}^{\mathrm{E},\ell,b} p(x_i^{\ell,b}) + b_{1,i,p}^{\mathrm{E},\ell,b} \right) \right] + b_{2,i}^{\mathrm{E},\ell,b}, \tag{43}$$

where d_c is the channel bottleneck dimension and

$$W_{1,i,p}^{\mathrm{E},\ell,b} \in \mathbb{R}^{d_c \times C}, \qquad W_{2,i}^{\mathrm{E},\ell,b} \in \mathbb{R}^{C \times d_c}.$$

The vectors $b_{1,i,p}^{\mathrm{E},\ell,b} \in \mathbb{R}^{d_c}$ and $b_{2,i}^{\mathrm{E},\ell,b} \in \mathbb{R}^C$ are trainable biases, and $\phi_{i,p}^{\mathrm{E},\ell,b}$ denotes the descriptor-specific activation. The resulting $C_i^{\mathrm{E},\ell,b} \in \mathbb{R}^C$ provides one response per input channel.

Alignment with the `SHANW` weights. Let $\mathrm{APool}_{k_i^{\ell,b}}$ denote adaptive average pooling to the $k_i^{\ell,b} \times k_i^{\ell,b}$ kernel grid. The weight-aligned Euclidean attention is

$$A_{i,\mathrm{E}}^{\ell,b}(x_i^{\ell,b}) = \left[\mathrm{vec}\left\{ \sigma\left(\mathrm{APool}_{k_i^{\ell,b}}(L_i^{\mathrm{E},\ell,b}) + \mathbf{1}_{k_i^{\ell,b}\times k_i^{\ell,b}} \otimes C_i^{\mathrm{E},\ell,b} \right) \right\} \right]_{\mathcal{I}_{\mathrm{SHANW},i}^{\ell,b}}, \tag{44}$$

where $\mathbf{1}$ is the all-ones array and $\otimes$ broadcasts the channel response across kernel positions. The restriction selects the complementary locations in their fixed weight ordering, yielding a vector in $(0,1)^{d_i^{\ell,b}}$.

D.3 Hyperbolic-Space Attention

We use the Lorentz manifold of curvature $-\kappa$, with $\kappa > 0$, origin o_κ, exponential map $\mathrm{Exp}_{o_\kappa}^\kappa$, and geodesic distance d_κ defined in App. B.3. Its intrinsic dimension is d_{H}.

Local hyperbolic response. A single 1×1 projection produces a bounded tangent-space representation:

$$V_i^{\mathrm{H},\ell,b} = \tanh\left(\mathrm{Conv}_{1\times1,i}^{\mathrm{H},\ell,b}(x_i^{\ell,b}) \right) \in \mathbb{R}^{H_\ell \times W_\ell \times d_{\mathrm{H}}}. \tag{45}$$

At spatial position (h, w), its Lorentz embedding is

$$Z_i^{\mathrm{H},\ell,b}[h,w] = \mathrm{Exp}_{o_\kappa}^\kappa\left(0, V_i^{\mathrm{H},\ell,b}[h,w,:]\right).$$

For input channel c, a learnable local prototype is parameterized by

$$P_{i,c}^{\mathrm{H},\ell,b} = \mathrm{Exp}_{o_\kappa}^\kappa\left(0, \tanh(p_{i,c}^{\mathrm{H},\ell,b})\right), \qquad p_{i,c}^{\mathrm{H},\ell,b} \in \mathbb{R}^{d_{\mathrm{H}}}.$$

The corresponding geodesic response is

$$L_i^{\mathrm{H},\ell,b}[h,w,c] = b_{i,c}^{\mathrm{H},\ell,b} - \beta_i^{\mathrm{H},\ell,b} d_\kappa\Big(Z_i^{\mathrm{H},\ell,b}[h,w], P_{i,c}^{\mathrm{H},\ell,b}\Big)^2, \tag{46}$$

where $b_{i,c}^{\mathrm{H},\ell,b}$ is a scalar bias and $\beta_i^{\mathrm{H},\ell,b} > 0$ scales the squared geodesic distance.

Multi-statistic hyperbolic channel response. For each $p \in \mathcal{P}$, embed the pooled descriptor as

$$Z_{i,p}^{\mathrm{H},\ell,b} = \mathrm{Exp}_{o_\kappa}^\kappa\left(0, \tanh\Big(U_{i,p}^{\mathrm{H},\ell,b} p(x_i^{\ell,b}) + a_{i,p}^{\mathrm{H},\ell,b}\Big)\right), \tag{47}$$

where $U_{i,p}^{\mathrm{H},\ell,b} \in \mathbb{R}^{d_{\mathrm{H}} \times C}$ and $a_{i,p}^{\mathrm{H},\ell,b} \in \mathbb{R}^{d_{\mathrm{H}}}$ are the descriptor projection and bias. For channel c, its descriptor-specific prototype is

$$Q_{i,p,c}^{\mathrm{H},\ell,b} = \mathrm{Exp}_{o_\kappa}^\kappa\left(0, \tanh(q_{i,p,c}^{\mathrm{H},\ell,b})\right), \qquad q_{i,p,c}^{\mathrm{H},\ell,b} \in \mathbb{R}^{d_{\mathrm{H}}}.$$

The additively fused channel response is

$$C_i^{\mathrm{H},\ell,b}[c] = \sum_{p\in\mathcal{P}} \left[a_{i,p,c}^{\mathrm{out},\ell,b} - \gamma_{i,p}^{\mathrm{H},\ell,b} d_\kappa\Big(Z_{i,p}^{\mathrm{H},\ell,b}, Q_{i,p,c}^{\mathrm{H},\ell,b}\Big)^2\right], \tag{48}$$

where $a_{i,p,c}^{\mathrm{out},\ell,b}$ is a scalar output bias and $\gamma_{i,p}^{\mathrm{H},\ell,b} > 0$ scales the corresponding squared geodesic distance.

Alignment with the `SHANW` weights. The hyperbolic attention follows the same pooling, broadcasting, and coefficient ordering as the Euclidean branch:

$$A_{i,\mathrm{H}}^{\ell,b}(x_i^{\ell,b}) = \Big[\mathrm{vec}\Big\{\sigma\Big(\mathrm{APool}_{k_i^{\ell,b}}(L_i^{\mathrm{H},\ell,b}) + \mathbf{1}_{k_i^{\ell,b}\times k_i^{\ell,b}} \otimes C_i^{\mathrm{H},\ell,b}\Big)\Big\}\Big]_{\mathcal{I}_{\mathrm{SHANW},i}^{\ell,b}}. \tag{49}$$

Thus, $A_{i,\mathrm{E}}^{\ell,b}(x_i^{\ell,b})$ and $A_{i,\mathrm{H}}^{\ell,b}(x_i^{\ell,b})$ belong to $(0,1)^{d_i^{\ell,b}}$ and refer to the same complementary `SHANW` locations.

D.4 Progressive Construction of SHANW Weights

We now expand Eq. 18 into sampling, geometry selection, recursive accumulation, and within-filter assembly. For one filter, suppress (i,ℓ,b) and write $x = x_i^{\ell,b}$, $n = n_i^{\ell,b}$, $r = r_i^{\ell,b}$, $k = k_i^{\ell,b}$, and $d = d_i^{\ell,b} = n - r$. The intermediate quantities below describe the existing weight-generation rule and introduce no additional trainable module.

Geometry-specific attention and weight alignment. The Euclidean branch combines local convolutional responses with projected channel descriptors, while the hyperbolic branch uses squared Lorentz geodesic distances to learned prototypes. Adaptive average pooling and channel broadcasting align these responses with the kernel grid. The sigmoid and fixed restrictions in Eqs. 44 and 49 yield $A_s(x) \in (0,1)^d$ for $s \in \mathcal{S} = \{\mathrm{E},\mathrm{H}\}$. The attention vectors, base weights, and noise coordinates follow the same ordering of the d allocated locations. Thus, the Euclidean and hyperbolic entries at coordinate a modulate the same weight.

Gaussian sampling and self-modulation. For $t = 1,\dots,T$, where $T \geq 1$ is the recursion depth, independently draw $\eta_s^{(t)} \sim \mathcal{N}(0, I_d)$ across geometries and recursions, with I_d the $d \times d$ identity matrix. The self-modulated noise is

$$\lambda_s^{(t)} = \eta_s^{(t)} \odot \big(\pi_s + \eta_s^{(t)} \odot \pi_s\big), \qquad s \in \mathcal{S}, \tag{50}$$

where π_s is the trainable geometry-specific scale from Sec. 2.3.2. At coordinate a, $\lambda_s^{(t)}[a] = \pi_s[a](\eta_s^{(t)}[a] + \eta_s^{(t)}[a]^2)$, where $\pi_s[a]$ is the scale applied at that location, with broadcasting when a scale is shared. Both branches use standard Gaussian draws in the aligned coordinate space; their geometric distinction enters through $A_s(x)$, not through a separate Gaussian law on the Lorentz manifold. The self-modulated noise need not be centered, as quantified in Lemma 5.

Noise-conditioned geometry selection. At recursion t, form the geometry-selection logits

$$z_s^{(t)} = A_s(x) \odot \lambda_s^{(t)}. \tag{51}$$

For each weight coordinate $a \in \{1, \dots, d\}$, normalize across the two geometries:

$$\omega_s^{(t)}[a] = \frac{\exp(z_s^{(t)}[a])}{\sum_{q \in \mathcal{S}} \exp(z_q^{(t)}[a])}. \tag{52}$$

These coefficients satisfy $\omega_s^{(t)}[a] > 0$ and $\sum_{s \in \mathcal{S}} \omega_s^{(t)}[a] = 1$. The fused attention response is

$$\mu^{(t)} = \sum_{s \in \mathcal{S}} \omega_s^{(t)} \odot A_s(x). \tag{53}$$

The softmax therefore determines the relative Euclidean and hyperbolic contributions at each weight location, rather than normalizing over locations or input channels.

Recursive accumulation and weight generation. Initialize $R^{(0)} = \mathbf{1}_d$, the vector of ones, and accumulate the sigmoid-modulated responses:

$$R^{(t)} = R^{(t-1)} \odot \sigma\Big(\mu^{(t)}\Big), \qquad t = 1, \dots, T. \tag{54}$$

For a fixed input and parameter state, $A_{\mathrm{E}}(x)$ and $A_{\mathrm{H}}(x)$ remain unchanged across recursions; fresh noise changes their mixing coefficients. The recursion accumulates attention factors, not optimizer updates or attention recomputed from successively updated weights. With trainable base weights $\overline{W}_{\mathrm{SHANW}} \in \mathbb{R}^d$, the resulting vector is

$$W_{\mathrm{SHANW}} = \overline{W}_{\mathrm{SHANW}} + \rho R^{(T)}, \qquad \rho = \mathrm{SoftPlus}(\rho_{\mathrm{logit}}) > 0, \tag{55}$$

where ρ_{logit} is trainable. Unrolling Eq. 54 gives $R^{(T)} = \prod_{t=1}^{T} \sigma(\mu^{(t)})$, with multiplication performed coordinate-wise. Substituting Eqs. 51–53 recovers Eq. 18. Thus, T determines the number of sampled attention factors, and ρ scales their accumulated contribution to the base weights.

Within-filter assembly. Fix an ordering $(j_1, \dots, j_d)$ of $\mathcal{I}^{\ell,b}_{\mathrm{SHANW},i}$. The embedding $\mathsf{E}_{\mathcal{I}} : \mathbb{R}^d \to \mathbb{R}^n$ places $v \in \mathbb{R}^d$ at these locations:

$$\big[\mathsf{E}_{\mathcal{I}}(v)\big][j] = \begin{cases} v[a], & j = j_a, \quad a \in \{1, \dots, d\}, \\ 0, & j \notin \mathcal{I}^{\ell,b}_{\mathrm{SHANW},i}. \end{cases} \tag{56}$$

Embedding each filter's vector, stacking the output filters, and reshaping gives the `SHANW` tensor in Eq. 3. The locations reserved for `ORaW` are zero, so addition combines the two mechanisms without overlapping their assigned coefficients. The shared scaling then constrains the integrated input-conditioned mapping. This embedding specifies the vector-to-tensor assembly and adds no trainable parameters.

Learning and stochastic execution. The attention parameters, base weights, geometry-specific scales, and ρ_{logit} are trained through Eq. 15; the Gaussian draws are sampled, not optimized. For a fixed draw, the attention branches, softmax mixing, and recursive product remain part of the weight-generation computation used for backpropagation. New independent evaluations use fresh draws while retaining the learned parameters and fixed allocation. Within each stochastic forward pass, the terminal `HyWeCS` weights are reused for reciprocal intermediate cross fusion. Certification-compatible sampling and attack-phase sample reuse follow Algorithms 1–3.

D.5 Theoretical Properties of SHANW

The following results concern finite inputs, parameters, and sampled noise values, with $T \geq 1$ and $\rho_i^{\ell,b} > 0$ as defined above. Input-sensitivity statements hold on the feature domain specified in their assumptions, with the internal draw fixed.

Lemma 4 (Bounded geometry-specific attention)**.** *For every admissible input, geometry $s \in \{\mathrm{E}, \mathrm{H}\}$, and coordinate $a \in \{1, \dots, d_i^{\ell,b}\}$,*

$$0 < A_{i,s}^{\ell,b}(x_i^{\ell,b})[a] < 1. \tag{57}$$

Proof. The Euclidean local and channel responses are finite for finite features and parameters. In the hyperbolic branch, $\tanh$ bounds the tangent representations and prototype coordinates, giving finite Lorentz geodesic distances. Thus, the logits in Eqs. 44 and 49 are finite. Since $0 < \sigma(z) < 1$ for every finite z, the result follows. □

Lemma 5 (Moments of the self-modulated noise)**.** *For one coordinate in Eq. 50, let $\eta \sim \mathcal{N}(0,1)$ and let π be fixed. Then*

$$\mathbb{E}[\lambda] = \pi, \qquad \mathrm{Var}(\lambda) = 3\pi^2, \qquad \mathbb{E}[\lambda^2] = 4\pi^2. \tag{58}$$

For the corresponding vectors with independent Gaussian coordinates,

$$\mathbb{E}[\lambda] = \pi, \qquad \mathrm{Cov}(\lambda) = 3\,\mathrm{Diag}(\pi^{\odot 2}),$$

where $\pi^{\odot 2}$ denotes coordinate-wise squaring and Diag *forms a diagonal matrix.*

Proof. The self-modulation gives $\lambda = \pi(\eta + \eta^2)$. Using the Gaussian moments

$$\mathbb{E}[\eta] = 0, \qquad \mathbb{E}[\eta^2] = 1, \qquad \mathbb{E}[\eta^3] = 0, \qquad \mathbb{E}[\eta^4] = 3,$$

we obtain

$$\mathbb{E}[\lambda] = \pi(0+1) = \pi, \qquad \mathbb{E}[\lambda^2] = \pi^2\mathbb{E}[\eta^2 + 2\eta^3 + \eta^4] = 4\pi^2.$$

Hence, $\mathrm{Var}(\lambda) = 4\pi^2 - \pi^2 = 3\pi^2$. Coordinate-wise self-modulation preserves independence for fixed scales, so the off-diagonal covariances vanish. □

Theorem 3 (Bounded recursive SHANW update)**.** *Let $\Delta W^{\ell,b}_{\mathrm{SHANW},i} = W^{\ell,b}_{\mathrm{SHANW},i} - \overline{W}^{\ell,b}_{\mathrm{SHANW},i}$. Then*

$$\left\|\Delta W^{\ell,b}_{\mathrm{SHANW},i}\right\|_\infty < \rho_i^{\ell,b}\sigma(1)^T, \qquad \left\|\Delta W^{\ell,b}_{\mathrm{SHANW},i}\right\|_2 < \rho_i^{\ell,b}\sqrt{d_i^{\ell,b}}\sigma(1)^T. \tag{59}$$

Proof. At recursion t and coordinate a, the geometry-wise softmax coefficients satisfy

$$\omega^{(t),\ell,b}_{i,s,a} > 0, \qquad \sum_{s\in\{\mathrm{E},\mathrm{H}\}} \omega^{(t),\ell,b}_{i,s,a} = 1.$$

By Lemma 4, their attention mixture

$$\mu^{(t),\ell,b}_{i,a} = \sum_{s\in\{\mathrm{E},\mathrm{H}\}} \omega^{(t),\ell,b}_{i,s,a} A^{\ell,b}_{i,s}(x_i^{\ell,b})[a]$$

is a convex combination of values in $(0,1)$. Consequently,

$$0 < \mu^{(t),\ell,b}_{i,a} < 1, \qquad \tfrac{1}{2} < \sigma\left(\mu^{(t),\ell,b}_{i,a}\right) < \sigma(1).$$

Multiplying the T factors in Eq. 18 gives

$$\left|\Delta W^{\ell,b}_{\mathrm{SHANW},i}[a]\right| < \rho_i^{\ell,b}\sigma(1)^T.$$

Taking the largest coordinate proves the ℓ_∞ bound. Moreover,

$$\left\|\Delta W^{\ell,b}_{\mathrm{SHANW},i}\right\|_2^2 = \sum_{a=1}^{d_i^{\ell,b}} \left|\Delta W^{\ell,b}_{\mathrm{SHANW},i}[a]\right|^2 < d_i^{\ell,b}(\rho_i^{\ell,b})^2\sigma(1)^{2T},$$

which proves the ℓ_2 bound. □

Proposition 2 (Realization-wise input sensitivity of SHANW)**.** *Fix the self-modulated noise draws and define the concatenated attention vector*

$$A_i^{\ell,b}(x) = \left(A^{\ell,b}_{i,\mathrm{E}}(x), A^{\ell,b}_{i,\mathrm{H}}(x)\right).$$

Assume $\|A_i^{\ell,b}(x) - A_i^{\ell,b}(x')\|_2 \le L^{\ell,b}_{A,i}\|x - x'\|_2$. For recursion t, define

$$\Lambda_i^{(t),\ell,b} = \max_{a,s}\left|\lambda^{(t),\ell,b}_{i,s}[a]\right|.$$

The generated weight vector then satisfies

$$\mathrm{Lip}_2\left(W^{\ell,b}_{\mathrm{SHANW},i}\right) \le \frac{\rho_i^{\ell,b}L^{\ell,b}_{A,i}}{4}\sum_{t=1}^{T}\left(1 + \frac{\Lambda_i^{(t),\ell,b}}{\sqrt{2}}\right). \tag{60}$$

Proof. For one weight location and recursion, suppress the layer indices and write

$$a = (A_{\mathrm{E}}, A_{\mathrm{H}}), \qquad z = a \odot \lambda, \qquad \omega = \mathrm{softmax}(z), \qquad \mu(a) = \omega(a)^{\top} a.$$

The gradient of the fused response is

$$\nabla_a \mu = \omega + \mathrm{Diag}(\lambda) J_{\mathrm{sm}}(z)^{\top} a,$$

where $J_{\mathrm{sm}}(z)$ is the two-dimensional softmax Jacobian. Since $\|J_{\mathrm{sm}}(z)\|_2 \le 1/2$, $\|\omega\|_2 \le 1$, $\|a\|_2 < \sqrt{2}$, and $\|\mathrm{Diag}(\lambda)\|_2 \le \Lambda^{(t)}$,

$$\|\nabla_a \mu\|_2 \le 1 + \frac{\Lambda^{(t)}}{\sqrt{2}}.$$

Using $\sup_z |\sigma'(z)| = 1/4$, each sigmoid-modulated factor has Lipschitz constant at most $\frac{1}{4}(1 + \Lambda^{(t)}/\sqrt{2})$ with respect to its attention pair. These operations act separately on the attention pair at each weight location, so their block-diagonal Jacobian gives the same bound for the vector of factors. For scalar factors $m_t, m'_t \in (0, 1)$, telescoping gives

$$\left| \prod_{t=1}^{T} m_t - \prod_{t=1}^{T} m'_t \right| \le \sum_{t=1}^{T} |m_t - m'_t|.$$

Applying this inequality coordinate-wise and the triangle inequality in the Euclidean norm bounds the product's sensitivity by the sum of the factor bounds. Composition with $A_i^{\ell,b}$ and multiplication by $\rho_i^{\ell,b}$ establish Eq. 60. □

Corollary 3 (Certificate compatibility of the SHANW-integrated layer)**.** *Fix an admissible internal draw and let*

$$\widetilde{W}_i^{\ell,b}(x) = W_{\mathrm{oraw},i}^{\ell,b} + W_{\mathrm{SHANW},i}^{\ell,b}(x)$$

denote the uncalibrated tensor after within-filter assembly. On a bounded feature domain, assume

$$\sup_x \left\| \mathrm{Conv}\Big(\cdot; \widetilde{W}_i^{\ell,b}(x)\Big) \right\|_{2\to 2} \le S_i^{\ell,b}, \qquad \sup_x \|x\|_2 \le M_i^{\ell,b}.$$

Let $K_{\mathrm{SHANW},i}^{\ell,b}$ satisfy the operator-level weight-variation bound

$$\left\| \mathrm{Conv}\Big(\cdot; W_{\mathrm{SHANW},i}^{\ell,b}(x) - W_{\mathrm{SHANW},i}^{\ell,b}(x')\Big) \right\|_{2\to 2} \le K_{\mathrm{SHANW},i}^{\ell,b} \|x - x'\|_2. \tag{61}$$

If the input-independent upper bound $\widehat{L}_i^{\ell,b} > 0$ in Eq. 3 satisfies

$$\widehat{L}_i^{\ell,b} \ge S_i^{\ell,b} + M_i^{\ell,b} K_{\mathrm{SHANW},i}^{\ell,b}, \tag{62}$$

then, on the stated domain,

$$\mathrm{Lip}_2(\mathcal{G}_i^{\ell,b}) \le \tau_i^{\ell,b} \le 1. \tag{63}$$

Proof. For a fixed draw, $W_{\mathrm{oraw},i}^{\ell,b}$ is independent of the input, so only SHANW contributes to the input dependence of $\widetilde{W}_i^{\ell,b}(x)$. The decomposition

$$\begin{aligned} &\mathrm{Conv}\Big(x; \widetilde{W}_i^{\ell,b}(x)\Big) - \mathrm{Conv}\Big(x'; \widetilde{W}_i^{\ell,b}(x')\Big) \\ &\quad = \mathrm{Conv}\Big(x - x'; \widetilde{W}_i^{\ell,b}(x)\Big) + \mathrm{Conv}\Big(x'; \widetilde{W}_i^{\ell,b}(x) - \widetilde{W}_i^{\ell,b}(x')\Big) \end{aligned}$$

and the stated bounds give

$$\mathrm{Lip}_2\Big(x \mapsto \mathrm{Conv}\Big(x; \widetilde{W}_i^{\ell,b}(x)\Big)\Big) \le S_i^{\ell,b} + M_i^{\ell,b} K_{\mathrm{SHANW},i}^{\ell,b}.$$

This is the input-conditioned convolution bound in Lemma 2. By Eq. 62, the right-hand side is at most $\widehat{L}_i^{\ell,b}$. The shared factor $c_i^{\ell,b} := \min\{1, \tau_i^{\ell,b}/\widehat{L}_i^{\ell,b}\}$ is input-independent for this draw and satisfies

$$c_i^{\ell,b} \widehat{L}_i^{\ell,b} = \min\{\widehat{L}_i^{\ell,b}, \tau_i^{\ell,b}\} \le \tau_i^{\ell,b}.$$

Therefore, $\|\mathcal{G}_i^{\ell,b}(x) - \mathcal{G}_i^{\ell,b}(x')\|_2 \le \tau_i^{\ell,b} \|x - x'\|_2$, which proves Eq. 63. □

From coefficient sensitivity to operator sensitivity. Proposition 2 bounds a generated weight vector in the Euclidean norm. To obtain the operator-level constant in Eq. 61, these bounds must be combined across output filters and converted to the induced convolution-operator norm. For a standard zero-padded $k \times k$ convolution, let K_o bound the Euclidean input sensitivity of the generated coefficients in output filter $o \in \{1, \dots, C_{\text{out}}\}$, where C_{out} is the output-channel count and $k = k_i^{\ell,b}$. A sufficient choice is

$$K_{\text{SHANW},i}^{\ell,b} = k \left(\sum_{o=1}^{C_{\text{out}}} K_o^2 \right)^{1/2}. \tag{64}$$

Indeed, placing each coefficient vector on its fixed support preserves its Euclidean norm. For the assembled kernel difference $U = W_{\text{SHANW},i}^{\ell,b}(x) - W_{\text{SHANW},i}^{\ell,b}(x')$, stacking the output filters therefore gives

$$\|U\|_{\text{F}} \le \left(\sum_{o=1}^{C_{\text{out}}} K_o^2 \right)^{1/2} \|x - x'\|_2,$$

where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm. Let $U_{u,v}$ be the channel matrix at kernel lag (u, v). Each zero-padded spatial shift is non-expansive, so

$$\|\operatorname{Conv}(\cdot; U)\|_{2\to 2} \le \sum_{u,v} \|U_{u,v}\|_2 \le \sum_{u,v} \|U_{u,v}\|_{\text{F}} \le k\|U\|_{\text{F}},$$

where the final inequality applies Cauchy–Schwarz to the k^2 kernel positions. Combining the two inequalities proves Eq. 61 with Eq. 64. Other padding or convolution conventions require their corresponding operator-norm conversion.

E Proofs of the HyWeCS Robustness Guarantees

This section provides the supporting condition for reciprocal intermediate cross fusion and the detailed proofs of Theorem 1, Corollary 1, and Theorem 2. All feature tensors are understood in their vectorized forms when evaluating Euclidean and operator norms.

E.1 Sufficient Condition for the Reciprocal Cross-Fusion Bound

Theorem 1 requires the complete reciprocal cross-fusion mapping, including the feature dependence of its reused terminal weights, to satisfy

$$\operatorname{Lip}_{\Pi}\left(\mathcal{C}^b\right) \le 1.$$

We next derive a sufficient condition for this bound by accounting for both the operator magnitude of the terminal HyWeCS weights and their variation with the corresponding modality representations.

Proposition 3 (Reciprocal cross-fusion sensitivity bound)**.** *Let u and v denote the paired modality representations entering reciprocal intermediate cross fusion at stage b. Let $W_i^{T,b}(u)$ and $W_{i+1}^{T,b}(v)$ denote the induced channel-wise linear operators obtained from the feature-conditioned terminal HyWeCS weights. Define*

$$\mathcal{C}^b(u, v) = \left(W_i^{T,b}(u)v \right) \| \left(W_{i+1}^{T,b}(v)u \right), \tag{65}$$

where $\|$ denotes channel-wise concatenation.

Assume that the terminal operators satisfy

$$\sup_u \left\| W_i^{T,b}(u) \right\|_{2\to 2} \le a_i^b, \qquad \sup_v \left\| W_{i+1}^{T,b}(v) \right\|_{2\to 2} \le a_{i+1}^b, \tag{66}$$

and that their feature dependence obeys

$$\begin{aligned} \left\| W_i^{T,b}(u) - W_i^{T,b}(u') \right\|_{2\to 2} &\le k_i^b \|u - u'\|_2, \\ \left\| W_{i+1}^{T,b}(v) - W_{i+1}^{T,b}(v') \right\|_{2\to 2} &\le k_{i+1}^b \|v - v'\|_2. \end{aligned} \tag{67}$$

Suppose further that $\|u\|_2 \le M_i^b$ *and* $\|v\|_2 \le M_{i+1}^b$. *Define*

$$B_{\mathrm{c}}^b = \begin{bmatrix} M_{i+1}^b k_i^b & a_i^b \\ a_{i+1}^b & M_i^b k_{i+1}^b \end{bmatrix}. \tag{68}$$

Then

$$\mathrm{Lip}_{\Pi}\left(\mathcal{C}^b\right) \le \left\|B_{\mathrm{c}}^b\right\|_2. \tag{69}$$

Consequently, $\|B_{\mathrm{c}}^b\|_2 \le 1$ *is sufficient to establish the reciprocal cross-fusion condition required by Theorem 1.*

Proof. Consider two paired representations (u, v) and (u', v'). For the first reciprocal transformation, adding and subtracting $W_i^{T,b}(u)v'$ gives

$$\begin{aligned} &\left\|W_i^{T,b}(u)v - W_i^{T,b}(u')v'\right\|_2 \\ &\le \left\|W_i^{T,b}(u)\left(v - v'\right)\right\|_2 + \left\|\left[W_i^{T,b}(u) - W_i^{T,b}(u')\right]v'\right\|_2 \\ &\le a_i^b\|v - v'\|_2 + M_{i+1}^b k_i^b\|u - u'\|_2. \end{aligned}$$

The same decomposition for the second reciprocal transformation yields

$$\begin{aligned} &\left\|W_{i+1}^{T,b}(v)u - W_{i+1}^{T,b}(v')u'\right\|_2 \\ &\le a_{i+1}^b\|u - u'\|_2 + M_i^b k_{i+1}^b\|v - v'\|_2. \end{aligned}$$

Define the input-displacement and output-displacement vectors

$$d = \begin{bmatrix} \|u - u'\|_2 \\ \|v - v'\|_2 \end{bmatrix}, \qquad e = \begin{bmatrix} \left\|W_i^{T,b}(u)v - W_i^{T,b}(u')v'\right\|_2 \\ \left\|W_{i+1}^{T,b}(v)u - W_{i+1}^{T,b}(v')u'\right\|_2 \end{bmatrix}.$$

The preceding inequalities imply

$$e \preceq B_{\mathrm{c}}^b d,$$

where $\preceq$ denotes element-wise comparison. Since all entries of e, d, and B_{c}^b are nonnegative,

$$\|e\|_2 \le \left\|B_{\mathrm{c}}^b d\right\|_2 \le \left\|B_{\mathrm{c}}^b\right\|_2 \|d\|_2.$$

Moreover, channel-wise concatenation gives

$$\|a\|b\|_2^2 = \|a\|_2^2 + \|b\|_2^2.$$

Therefore,

$$\begin{aligned} &\left\|\mathcal{C}^b(u, v) - \mathcal{C}^b(u', v')\right\|_2 \\ &= \|e\|_2 \\ &\le \left\|B_{\mathrm{c}}^b\right\|_2 \left(\|u - u'\|_2^2 + \|v - v'\|_2^2\right)^{1/2} \\ &= \left\|B_{\mathrm{c}}^b\right\|_2 \|(u, v) - (u', v')\|_{2,\Pi}. \end{aligned}$$

Taking the supremum over all distinct paired representations proves Eq. 69. In particular, $\|B_{\mathrm{c}}^b\|_2 \le 1$ implies $\mathrm{Lip}_{\Pi}(\mathcal{C}^b) \le 1$. □

E.2 Proof of Theorem 1

Proof. Fix an admissible internal stochastic realization. The residual pathway in Eq. 5 is the composition

$$\mathcal{R}_i^b = \mathcal{G}_i^{2,b} \circ \varsigma \circ \mathcal{G}_i^{1,b}.$$

By the composition rule and the assumed layer-wise bounds,

$$\mathrm{Lip}_2\left(\mathcal{R}_i^b\right) \le \tau_i^{2,b} L_\varsigma \tau_i^{1,b}.$$

The residual-block mapping can be written as

$$\mathcal{B}_i^b = I + \mathcal{R}_i^b.$$

Hence, for arbitrary representations x and x',

$$\begin{aligned}\left\|\mathcal{B}_i^b(x) - \mathcal{B}_i^b(x')\right\|_2 &\le \|x - x'\|_2 + \left\|\mathcal{R}_i^b(x) - \mathcal{R}_i^b(x')\right\|_2 \\ &\le \left(1 + \tau_i^{2,b}\tau_i^{1,b}L_\varsigma\right)\|x - x'\|_2.\end{aligned}$$

Consequently,

$$\mathrm{Lip}_2\left(\mathcal{B}_i^b\right) \le 1 + \tau_i^{2,b}\tau_i^{1,b}L_\varsigma \le 2,$$

which proves the residual-block claim in Theorem 1.

For the paired modality representations, define

$$\mathcal{B}_{\mathrm{pair}}^b\left(x_i, x_{i+1}\right) = \left(\mathcal{B}_i^b(x_i), \mathcal{B}_{i+1}^b(x_{i+1})\right).$$

For $X = (x_i, x_{i+1})$ and $X' = (x_i', x_{i+1}')$, the paired product norm gives

$$\begin{aligned}&\left\|\mathcal{B}_{\mathrm{pair}}^b(X) - \mathcal{B}_{\mathrm{pair}}^b(X')\right\|_{2,\Pi}^2 \\ &= \left\|\mathcal{B}_i^b(x_i) - \mathcal{B}_i^b(x_i')\right\|_2^2 + \left\|\mathcal{B}_{i+1}^b(x_{i+1}) - \mathcal{B}_{i+1}^b(x_{i+1}')\right\|_2^2 \\ &\le 4\left(\|x_i - x_i'\|_2^2 + \|x_{i+1} - x_{i+1}'\|_2^2\right).\end{aligned}$$

Thus,

$$\mathrm{Lip}_\Pi\left(\mathcal{B}_{\mathrm{pair}}^b\right) \le 2.$$

The reciprocal cross-fusion condition in Theorem 1 gives

$$\mathrm{Lip}_\Pi\left(\mathcal{C}^b\right) \le 1.$$

The composition rule therefore yields

$$\begin{aligned}\mathrm{Lip}_\Pi\left(\mathcal{C}^b \circ \mathcal{B}_{\mathrm{pair}}^b\right) &\le \mathrm{Lip}_\Pi\left(\mathcal{C}^b\right)\mathrm{Lip}_\Pi\left(\mathcal{B}_{\mathrm{pair}}^b\right) \\ &\le 2.\end{aligned}$$

Proposition 3 provides a sufficient architectural condition for this reciprocal cross-fusion bound.

We next consider the complete multimodal score mapping. By the definition of the verified uniform bound $\widehat{L}_\Theta$, every admissible internal realization satisfies

$$\begin{aligned}&\|\widetilde{\boldsymbol{s}}_\Theta\left(X; \Gamma_{\mathrm{det}}, \Omega_{\mathrm{H}}\right) - \widetilde{\boldsymbol{s}}_\Theta\left(X'; \Gamma_{\mathrm{det}}, \Omega_{\mathrm{H}}\right)\|_2 \\ &\le \widehat{L}_\Theta\|X - X'\|_{2,\Pi}.\end{aligned}$$

Let

$$\chi_\Theta = \min\left\{1,\ \frac{L_\star}{\widehat{L}_\Theta}\right\}$$

denote the positive network-wide scaling factor used by the joint deterministic–stochastic classifier. Since

$$\boldsymbol{s}_\Theta = \chi_\Theta\widetilde{\boldsymbol{s}}_\Theta,$$

we obtain

$$\begin{aligned}&\|\boldsymbol{s}_\Theta\left(X; \Gamma_{\mathrm{det}}, \Omega_{\mathrm{H}}\right) - \boldsymbol{s}_\Theta\left(X'; \Gamma_{\mathrm{det}}, \Omega_{\mathrm{H}}\right)\|_2 \\ &\le \chi_\Theta\widehat{L}_\Theta\|X - X'\|_{2,\Pi} \\ &= \min\left\{\widehat{L}_\Theta, L_\star\right\}\|X - X'\|_{2,\Pi} \\ &\le L_\star\|X - X'\|_{2,\Pi}.\end{aligned}$$

Because this inequality holds for every admissible Ω_{H},

$$\sup_{\Omega_{\mathrm{H}}} \mathrm{Lip}_\Pi\left(\boldsymbol{s}_\Theta\left(\cdot; \Gamma_{\mathrm{det}}, \Omega_{\mathrm{H}}\right)\right) \le L_\star.$$

Finally, consider the expected score maps at X and X' using the same Gaussian perturbation and internal realization. Jensen's inequality gives

$$\begin{aligned}
&\left\|\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}(X) - \overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}(X')\right\|_2 \\
&= \left\|\mathbb{E}_{\boldsymbol{\varepsilon},\Omega_{\mathrm{H}}}\left[\boldsymbol{s}_{\Theta}\left(X+\boldsymbol{\varepsilon};\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}}\right) - \boldsymbol{s}_{\Theta}\left(X'+\boldsymbol{\varepsilon};\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}}\right)\right]\right\|_2 \\
&\leq \mathbb{E}_{\boldsymbol{\varepsilon},\Omega_{\mathrm{H}}}\left[\left\|\boldsymbol{s}_{\Theta}\left(X+\boldsymbol{\varepsilon};\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}}\right) - \boldsymbol{s}_{\Theta}\left(X'+\boldsymbol{\varepsilon};\Gamma_{\mathrm{det}},\Omega_{\mathrm{H}}\right)\right\|_2\right] \\
&\leq L_\star \left\|X - X'\right\|_{2,\Pi}.
\end{aligned}$$

Therefore,

$$\mathrm{Lip}_{\Pi}\left(\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\right) \leq L_\star.$$

Together, these results establish Eq. 10. □

E.3 Proof of Corollary 1

Proof. Let $c = c_{\mathrm{e}}(X)$ denote the class selected by the joint expected-logit classifier. For every $j \neq c$, Theorem 1 gives

$$\left|\left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X+\delta\right)\right]_c - \left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X\right)\right]_c\right| \leq L_\star \|\delta\|_{2,\Pi}$$

and

$$\left|\left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X+\delta\right)\right]_j - \left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X\right)\right]_j\right| \leq L_\star \|\delta\|_{2,\Pi}.$$

Consequently,

$$\begin{aligned}
&\left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X+\delta\right)\right]_c - \left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X+\delta\right)\right]_j \\
&\geq \left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X\right)\right]_c - \left[\overline{\boldsymbol{s}}^{\mathrm{H}}_{\Theta}\left(X\right)\right]_j - 2L_\star\|\delta\|_{2,\Pi} \\
&\geq \Delta^{\mathrm{H}}_{\Theta}(X) - 2L_\star\|\delta\|_{2,\Pi}.
\end{aligned}$$

Every competing margin remains positive whenever

$$\|\delta\|_{2,\Pi} < \frac{\Delta^{\mathrm{H}}_{\Theta}(X)}{2L_\star}.$$

Therefore,

$$\widehat{y}^{\mathrm{H}}_{\Theta}\left(X+\delta\right) = c_{\mathrm{e}}(X),$$

which proves Eq. 11. For $L_\star = 2$, the corresponding expected-logit radius is

$$r_{\exp}(X) = \frac{\Delta^{\mathrm{H}}_{\Theta}(X)}{4}.$$

For the probability-based classifier, define the event

$$\mathcal{A}_c = \left\{(z,\omega) : \arg\max_{j\in\mathcal{Y}} \boldsymbol{s}_{\Theta,j}\left(z;\Gamma_{\mathrm{det}},\omega\right) = c\right\}.$$

The class probability can then be written as

$$p^{\mathrm{H}}_c(X) = \mathbb{P}_{\boldsymbol{\varepsilon},\Omega_{\mathrm{H}}}\left[(X+\boldsymbol{\varepsilon},\Omega_{\mathrm{H}}) \in \mathcal{A}_c\right].$$

Applying the Gaussian-shift result with auxiliary internal randomness from Lemma 3 to the selected class gives

$$p^{\mathrm{H}}_{c_{\mathrm{s}}(X)}\left(X+\delta\right) \geq \Phi\left(\Phi^{-1}(p_A) - \frac{\|\delta\|_{2,\Pi}}{\sigma}\right).$$

For every competing class $j \neq c_{\mathrm{s}}(X)$,

$$\begin{aligned}
p^{\mathrm{H}}_j\left(X+\delta\right) &\leq \Phi\left(\Phi^{-1}\left(p^{\mathrm{H}}_j(X)\right) + \frac{\|\delta\|_{2,\Pi}}{\sigma}\right) \\
&\leq \Phi\left(\Phi^{-1}(p_B) + \frac{\|\delta\|_{2,\Pi}}{\sigma}\right).
\end{aligned}$$

Because $\Phi(\cdot)$ is strictly increasing, the selected-class lower bound exceeds every competing-class upper bound whenever

$$\Phi^{-1}(p_A) - \frac{\|\delta\|_{2,\Pi}}{\sigma} > \Phi^{-1}(p_B) + \frac{\|\delta\|_{2,\Pi}}{\sigma}.$$

Rearranging gives

$$\|\delta\|_{2,\Pi} < \frac{\sigma}{2}\left[\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right].$$

Thus,

$$g_\Theta^{\mathrm{H}}(X+\delta) = c_{\mathrm{s}}(X),$$

which proves Eq. 12.

When $c_{\mathrm{e}}(X) = c_{\mathrm{s}}(X)$, the shared class satisfies both guarantees and is therefore certified within

$$\min\left\{r_{\exp}(X), r_{\mathrm{RS}}(X)\right\}.$$

□

E.4 Proof of Theorem 2

Proof. Let $\delta = (\delta_1, \ldots, \delta_m)$ satisfy

$$\|\delta\|_{\infty,\Pi} \le \epsilon_\infty.$$

For modality i, every coordinate of δ_i has magnitude at most ϵ_∞, and therefore

$$\|\delta_i\|_2^2 \le \dim(x_i)\epsilon_\infty^2.$$

Since

$$D = \sum_{i=1}^m \dim(x_i),$$

we obtain

$$\begin{aligned}\|\delta\|_{2,\Pi}^2 &= \sum_{i=1}^m \|\delta_i\|_2^2 \\ &\le \sum_{i=1}^m \dim(x_i)\epsilon_\infty^2 = D\epsilon_\infty^2.\end{aligned} \tag{70}$$

Consequently,

$$\|\delta\|_{2,\Pi} \le \sqrt{D}\,\epsilon_\infty. \tag{71}$$

For every admissible internal realization, the calibrated score map is $L_\star$-Lipschitz by Theorem 1. Thus, for arbitrary multimodal inputs Z and Z',

$$\begin{aligned}&\|\boldsymbol{s}_\Theta(Z;\Gamma_{\det},\Omega_{\mathrm{H}}) - \boldsymbol{s}_\Theta(Z';\Gamma_{\det},\Omega_{\mathrm{H}})\|_2 \\ &\le L_\star\|Z - Z'\|_{2,\Pi}.\end{aligned}$$

Since $\mathcal{L}_{\mathrm{MML}}(\cdot, y)$ is L_ℓ-Lipschitz with respect to its score-vector argument, composition gives

$$\begin{aligned}&|\mathcal{L}_{\mathrm{MML}}(\boldsymbol{s}_\Theta(Z), y) - \mathcal{L}_{\mathrm{MML}}(\boldsymbol{s}_\Theta(Z'), y)| \\ &\le L_\ell\,\|\boldsymbol{s}_\Theta(Z) - \boldsymbol{s}_\Theta(Z')\|_2 \\ &\le L_\ell L_\star\|Z - Z'\|_{2,\Pi},\end{aligned}$$

where the fixed deterministic configuration and stochastic realization are suppressed inside the last two displays.

Taking the expectation over $\boldsymbol{\varepsilon}$ and Ω_{H} does not enlarge this constant. Hence, $\overline{\mathcal{L}}_{X,y}(\cdot)$ is $L_\ell L_\star$-Lipschitz in the multimodal perturbation. Because the theorem assumes that this function is smooth, it is differentiable and satisfies

$$\left\|\nabla_\delta \overline{\mathcal{L}}_{X,y}(0)\right\|_2 \le L_\ell L_\star.$$

By L_s-smoothness,

$$\begin{aligned}\overline{\mathcal{L}}_{X,y}(\delta) &\le \overline{\mathcal{L}}_{X,y}(0) + \left\langle \nabla_\delta \overline{\mathcal{L}}_{X,y}(0), \delta \right\rangle + \frac{L_s}{2}\|\delta\|_{2,\Pi}^2 \\ &\le \overline{\mathcal{L}}_{X,y}(0) + \left\|\nabla_\delta \overline{\mathcal{L}}_{X,y}(0)\right\|_2 \|\delta\|_{2,\Pi} + \frac{L_s}{2}\|\delta\|_{2,\Pi}^2 \\ &\le \overline{\mathcal{L}}_{X,y}(0) + L_\ell L_\star \|\delta\|_{2,\Pi} + \frac{L_s}{2}\|\delta\|_{2,\Pi}^2.\end{aligned}$$

The second inequality follows from the Cauchy–Schwarz inequality. Applying Eq. 71 gives

$$\overline{\mathcal{L}}_{X,y}(\delta) \le \overline{\mathcal{L}}_{X,y}(0) + \epsilon_\infty \sqrt{D}\, L_\ell L_\star + \frac{L_s D}{2}\epsilon_\infty^2.$$

The right-hand side is independent of the particular admissible δ. Taking the supremum over

$$\{\delta : \|\delta\|_{\infty,\Pi} \le \epsilon_\infty\}$$

therefore yields

$$\overline{\mathcal{L}}_{X,y}^{\mathrm{adv}} \le \overline{\mathcal{L}}_{X,y}(0) + \epsilon_\infty \sqrt{D}\, L_\ell L_\star + \frac{L_s D}{2}\epsilon_\infty^2,$$

which proves Eq. 14. □

F Certification-Compatible Training and Inference

This section presents the certification-compatible training and inference-time certification procedures of the HyWeCS-integrated multimodal fusion learner. For a fixed deterministic configuration Γ_{det}, the internal weight-space randomness is

$$\Omega_{\mathrm{H}} = (\boldsymbol{\alpha}_{\mathrm{rp}}, \boldsymbol{\eta}) \sim \nu_{\mathrm{H}},$$

where $\boldsymbol{\alpha}_{\mathrm{rp}}$ collects all ORaW random-projection draws and $\boldsymbol{\eta}$ collects the Euclidean and hyperbolic SHANW noise draws.

Throughout Algorithms 1–3, we write

$$\mathsf{S}_\Theta(Z; \Omega_{\mathrm{H}}) := \boldsymbol{s}_\Theta(Z; \Gamma_{\mathrm{det}}, \Omega_{\mathrm{H}})$$

for one calibrated stochastic score realization. Each evaluation of S_Θ constructs the complementary ORaW and SHANW weights, applies their shared layer-wise calibration, evaluates the modality-specific residual mappings, reuses the terminal weights for reciprocal intermediate cross fusion, and scales the resulting score vector by $\min\{1, L_\star/\widehat{L}_\Theta\}$.

F.1 Certification-Compatible Training

Certification-compatible training follows Eq. 15 with $\lambda_{\mathrm{adv}} = 0$. Thus, every parameter update averages the multimodal learning loss over Gaussian input noise and the joint ORaW–SHANW realizations used by the certified classifiers. For Q_{tr} stochastic evaluations, we use

$$\widehat{\mathcal{L}}_{\mathrm{cert}} = \frac{1}{Q_{\mathrm{tr}}} \sum_{q=1}^{Q_{\mathrm{tr}}} \mathcal{L}_{\mathrm{MML}}\left(\mathsf{S}_\Theta\left(X + \boldsymbol{\varepsilon}^{(q)}; \Omega_{\mathrm{H}}^{(q)}\right), y\right), \tag{72}$$

where $\boldsymbol{\varepsilon}^{(q)} \sim \mathcal{N}(0, \sigma^2 I)$ and $\Omega_{\mathrm{H}}^{(q)} \sim \nu_{\mathrm{H}}$.

The choice $Q_{\mathrm{tr}} = 1$ gives a standard stochastic update, whereas larger values reduce the Monte Carlo variation of the training objective (see Algorithm 1).

F.2 Inference-Time Certification

At inference, Θ and Γ_{det} remain fixed, while Gaussian input noise and the internal HyWeCS randomness are independently resampled across evaluations. The expected-logit and probability-based population guarantees follow Corollary 1.

Algorithm 1 Certification-compatible training

Require: Training set $\mathcal{D}$, parameters Θ, epochs E, batch size B, smoothing scale σ, stochastic samples Q_{tr}, optimizer $\mathcal{O}$
Ensure: Trained parameters Θ

for $e = 1$ to E **do**
  for mini-batch $(X, y) \sim \mathcal{D}$ of size B **do**
    $\widehat{\mathcal{L}}_{\mathrm{cert}} \leftarrow 0$
    for $q = 1$ to Q_{tr} **do**
      Sample $\boldsymbol{\varepsilon}^{(q)} \sim \mathcal{N}(0, \sigma^2 I)$
      Sample $\Omega_{\mathrm{H}}^{(q)} = (\boldsymbol{\alpha}_{\mathrm{rp}}^{(q)}, \boldsymbol{\eta}^{(q)}) \sim \nu_{\mathrm{H}}$
      $\boldsymbol{s}^{(q)} \leftarrow \mathsf{S}_\Theta(X + \boldsymbol{\varepsilon}^{(q)}; \Omega_{\mathrm{H}}^{(q)})$
      $\widehat{\mathcal{L}}_{\mathrm{cert}} \leftarrow \widehat{\mathcal{L}}_{\mathrm{cert}} + Q_{\mathrm{tr}}^{-1} \mathcal{L}_{\mathrm{MML}}(\boldsymbol{s}^{(q)}, y)$
    end for
    $\Theta \leftarrow \mathcal{O}(\Theta, \nabla_\Theta \widehat{\mathcal{L}}_{\mathrm{cert}})$
  end for
end for
return Θ

For finite-sample probability certification, we follow the one-sided binomial confidence construction of Clopper and Pearson Clopper & Pearson (1934) and the randomized-smoothing certification protocol of Cohen et al. (Cohen et al., 2019). Given k occurrences of a candidate class among n independent evaluations, define

$$\mathrm{LCB}_{\mathrm{CP}}(k, n, \alpha) = \begin{cases} 0, & k = 0, \\ F^{-1}_{\mathrm{Beta}(k, n-k+1)}(\alpha), & 1 \le k \le n, \end{cases} \tag{73}$$

where $F^{-1}_{\mathrm{Beta}(a,b)}$ denotes the quantile function of a $\mathrm{Beta}(a, b)$ distribution. This is a one-sided $(1 - \alpha)$ lower confidence bound for the candidate-class probability.

Algorithm 2 first uses an independent pilot sample to select the candidate probability-based class. A second sample set then estimates its probability and the expected score vector. The probability-based radius is a formal finite-sample certificate; the expected-logit radius returned by the algorithm is a Monte Carlo estimate of the population quantity in Corollary 1.

The probability radius follows from the conservative choice $\overline{p}_B = 1 - \underline{p}_A$:

$$\frac{\sigma}{2}\left[\Phi^{-1}(\underline{p}_A) - \Phi^{-1}(1 - \underline{p}_A)\right] = \sigma \Phi^{-1}(\underline{p}_A).$$

Therefore, whenever $\underline{p}_A > 1/2$, Algorithm 2 returns a valid finite-sample ℓ_2 certificate with confidence at least $1 - \alpha$.

In contrast, $\widehat{r}^{\mathrm{MC}}_{\mathrm{exp}}$ is an estimate of the population expected-logit radius. A formal finite-sample expected-logit certificate additionally requires valid simultaneous confidence bounds for the expected class scores. When such bounds $[\underline{s}_c, \overline{s}_c]$ are available, the certified radius is

$$\underline{r}_{\mathrm{exp}}(X) = \frac{\left[\underline{s}_{c_{\mathrm{e}}} - \max_{j \neq c_{\mathrm{e}}} \overline{s}_j\right]_+}{2L_\star}.$$

G Adversarial Training for Empirical Robustness

This section implements the adversarial component of Eq. 15. The inner maximization accounts for Gaussian input smoothing and the complete internal randomness Ω_{H}.

Let

$$\mathcal{B}_{\infty,\Pi}(\epsilon_\infty) = \{\delta : \|\delta\|_{\infty,\Pi} \le \epsilon_\infty\}$$

denote the multimodal perturbation set. At the beginning of each attack trajectory, we sample

$$\mathcal{U}_A = \left\{\left(\boldsymbol{\varepsilon}^{[A,q]}, \Omega_{\mathrm{H}}^{[A,q]}\right)\right\}_{q=1}^{Q_A}$$

Algorithm 2 Inference-time certification

Require: Trained parameters Θ, input X, smoothing scale σ, global constant $L_\star$, pilot samples n_0, certification samples n, failure probability α
Ensure: Expected-logit class $\widehat{c}_{\mathrm{e}}$ and radius estimate $\widehat{r}_{\mathrm{exp}}^{\mathrm{MC}}$; smoothed class $\widehat{c}_{\mathrm{s}}$ and certified radius $\widehat{r}_{\mathrm{RS}}$

Stage I: Candidate-class selection
$N_c^{(0)} \leftarrow 0$ for every $c \in \mathcal{Y}$
for $q = 1$ to n_0 **do**
  Sample $\boldsymbol{\varepsilon}^{(q)} \sim \mathcal{N}(0, \sigma^2 I)$ and $\Omega_{\mathrm{H}}^{(q)} \sim \nu_{\mathrm{H}}$
  $\boldsymbol{s}^{(q)} \leftarrow \mathsf{S}_\Theta(X + \boldsymbol{\varepsilon}^{(q)}; \Omega_{\mathrm{H}}^{(q)})$
  $c^{(q)} \leftarrow \arg\max_{c \in \mathcal{Y}} s_c^{(q)}$
  $N_{c^{(q)}}^{(0)} \leftarrow N_{c^{(q)}}^{(0)} + 1$
end for
$A \leftarrow \arg\max_{c \in \mathcal{Y}} N_c^{(0)}$
Stage II: Independent certification sampling
$N_A \leftarrow 0$ and $\boldsymbol{u} \leftarrow \mathbf{0} \in \mathbb{R}^{|\mathcal{Y}|}$
for $q = 1$ to n **do**
  Independently sample $\boldsymbol{\varepsilon}^{(q)} \sim \mathcal{N}(0, \sigma^2 I)$ and $\Omega_{\mathrm{H}}^{(q)} \sim \nu_{\mathrm{H}}$
  $\boldsymbol{s}^{(q)} \leftarrow \mathsf{S}_\Theta(X + \boldsymbol{\varepsilon}^{(q)}; \Omega_{\mathrm{H}}^{(q)})$
  $N_A \leftarrow N_A + \mathbf{1}\{\arg\max_c s_c^{(q)} = A\}$
  $\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{s}^{(q)}$
end for
Stage III: Expected-logit estimate
$\widehat{\boldsymbol{\mu}} \leftarrow n^{-1}\boldsymbol{u}$
$\widehat{c}_{\mathrm{e}} \leftarrow \arg\max_{c \in \mathcal{Y}} \widehat{\mu}_c$
$\widehat{\Delta} \leftarrow \widehat{\mu}_{\widehat{c}_{\mathrm{e}}} - \max_{j \neq \widehat{c}_{\mathrm{e}}} \widehat{\mu}_j$
$\widehat{r}_{\mathrm{exp}}^{\mathrm{MC}} \leftarrow [\widehat{\Delta}]_+ / (2L_\star)$
Stage IV: Finite-sample probability certificate
$\underline{p}_A \leftarrow \mathrm{LCB}_{\mathrm{CP}}(N_A, n, \alpha)$
if $\underline{p}_A > 1/2$ **then**
  $\widehat{c}_{\mathrm{s}} \leftarrow A$
  $\widehat{r}_{\mathrm{RS}} \leftarrow \sigma \Phi^{-1}(\underline{p}_A)$
else
  $\widehat{c}_{\mathrm{s}} \leftarrow$ ABSTAIN
  $\widehat{r}_{\mathrm{RS}} \leftarrow 0$
end if
return $(\widehat{c}_{\mathrm{e}}, \widehat{r}_{\mathrm{exp}}^{\mathrm{MC}}, \widehat{c}_{\mathrm{s}}, \widehat{r}_{\mathrm{RS}})$

and retain these realizations throughout the projected-gradient steps. The sample-average attack objective is

$$\widehat{\mathcal{L}}_A\left(\delta; \mathcal{U}_A\right) = \frac{1}{Q_A} \sum_{q=1}^{Q_A} \mathcal{L}_{\mathrm{MML}}\left(\mathsf{S}_\Theta\left(X + \delta + \boldsymbol{\varepsilon}^{[A,q]}; \Omega_{\mathrm{H}}^{[A,q]}\right), y\right). \tag{74}$$

After obtaining $\delta^\star$, the parameter-update phase independently samples

$$\mathcal{U}_U = \left\{\left(\boldsymbol{\varepsilon}^{[U,q]}, \Omega_{\mathrm{H}}^{[U,q]}\right)\right\}_{q=1}^{Q_U}.$$

Using the same update-phase realization for the clean and adversarial terms, we estimate

$$\begin{aligned}\widehat{\mathcal{L}}_U\left(\Theta, \delta^\star; \mathcal{U}_U\right) = \frac{1}{Q_U} \sum_{q=1}^{Q_U} \Big[& \mathcal{L}_{\mathrm{MML}}\left(\mathsf{S}_\Theta\left(X + \boldsymbol{\varepsilon}^{[U,q]}; \Omega_{\mathrm{H}}^{[U,q]}\right), y\right) \\ & + \lambda_{\mathrm{adv}} \mathcal{L}_{\mathrm{MML}}\left(\mathsf{S}_\Theta\left(X + \delta^\star + \boldsymbol{\varepsilon}^{[U,q]}; \Omega_{\mathrm{H}}^{[U,q]}\right), y\right)\Big].\end{aligned} \tag{75}$$

The choice $Q_A = 1$ trains against one sampled stochastic realization, whereas $Q_A > 1$ provides an expectation-over-transformation-aware estimate of the expected attack gradient. Increasing Q_U reduces the Monte Carlo variation of the parameter update. The attack samples remain fixed within each projected-gradient trajectory, while the update uses an independent set of Gaussian, `ORaW`, and `SHANW` realizations (see Algorithm 3).

Algorithm 3 Adversarial training for empirical robustness

Require: Training set $\mathcal{D}$, parameters Θ, epochs E, batch size B, smoothing scale σ, adversarial budget ϵ_∞, step size ζ_{adv}, attack steps K, attack samples Q_A, update samples Q_U, trade-off λ_{adv}, optimizer $\mathcal{O}$
Ensure: Trained parameters Θ

for $e = 1$ to E **do**
  for mini-batch $(X, y) \sim \mathcal{D}$ of size B **do**
    Stage I: Sample the attack realization
    Draw $\mathcal{U}_A = \{(\boldsymbol{\varepsilon}^{[A,q]}, \Omega_{\text{H}}^{[A,q]})\}_{q=1}^{Q_A}$
    Initialize $\delta^{(0)} \in \mathcal{B}_{\infty,\Pi}(\epsilon_\infty)$
    Stage II: Projected inner maximization
    for $t = 0$ to $K - 1$ **do**
      $g^{(t)} \leftarrow \nabla_\delta \widehat{\mathcal{L}}_A(\delta^{(t)}; \mathcal{U}_A)$
      $\delta^{(t+1)} \leftarrow \Pi_{\mathcal{B}_{\infty,\Pi}(\epsilon_\infty)} \left(\delta^{(t)} + \zeta_{\text{adv}} \operatorname{sign}(g^{(t)})\right)$
    end for
    $\delta^\star \leftarrow \operatorname{stopgrad}(\delta^{(K)})$
    Stage III: Independent parameter update
    Draw fresh $\mathcal{U}_U = \{(\boldsymbol{\varepsilon}^{[U,q]}, \Omega_{\text{H}}^{[U,q]})\}_{q=1}^{Q_U}$
    $\widehat{\mathcal{L}} \leftarrow \widehat{\mathcal{L}}_U(\Theta, \delta^\star; \mathcal{U}_U)$
    $\Theta \leftarrow \mathcal{O}(\Theta, \nabla_\Theta \widehat{\mathcal{L}})$
  end for
end for
return Θ

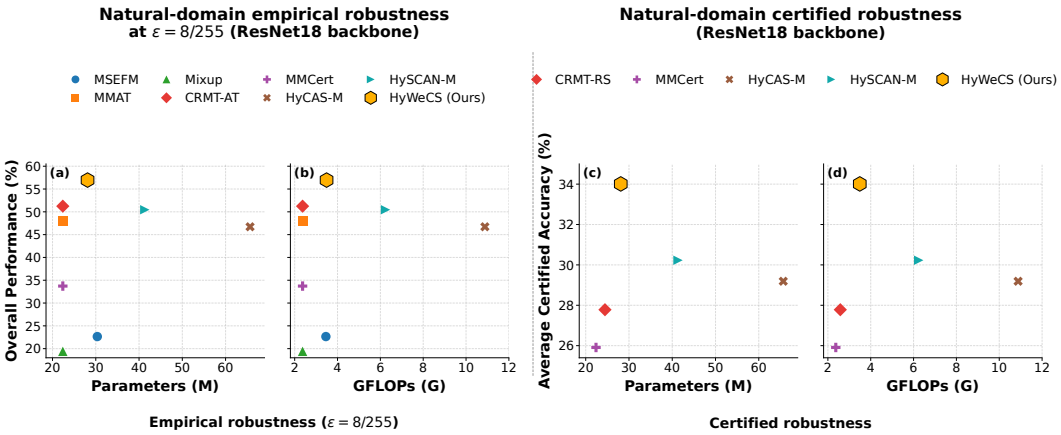


Figure 7: **Robustness–efficiency trade-off on natural-domain data with `ResNet18` backbones.** Panels (a)–(b) report *Overall Performance* on average across `KS`, `UCF101`, and `VGGS` against `AutoAttack`, `PGD`, `BPDA`, and `BPDA-EOT` at $\epsilon = 8/255$, as a function of parameter count and `GFLOPs`, respectively. Panels (c)–(d) report certified accuracy on average across `RAVDESS` and `UCF101` over all evaluated smoothing-noise and ℓ_2-radius settings, as a function of the same complexity measures. Points closer to the upper-left corner indicate a more favorable robustness–efficiency trade-off.

H Experimental Setup

H.1 Datasets

We evaluate the `HyWeCS`-integrated multimodal fusion learner on eight benchmarks spanning natural and medical domains. The natural-domain benchmarks comprise `RAVDESS` (Livingstone & Russo, 2018), `UCF101` (Soomro et al., 2012), `Kinetics-Sounds` (Arandjelovic & Zisserman, 2017) (`KS`), and `VGGSound` (Chen et al., 2020) (`VGGS`). `UCF101` uses RGB and optical-flow modalities, whereas `RAVDESS`, `KS`, and `VGGS` use audio–visual modalities.

The medical-domain benchmarks comprise `BRCA` (Lingle et al., 2016) and `BLCA` (Network et al., 2014), which combine imaging and multi-omics data, together with `HAM10000` (Tschandl et al., 2018) and `SIPaKMeD` (Plissiti et al., 2018), which are evaluated in image–tabular multimodal set-

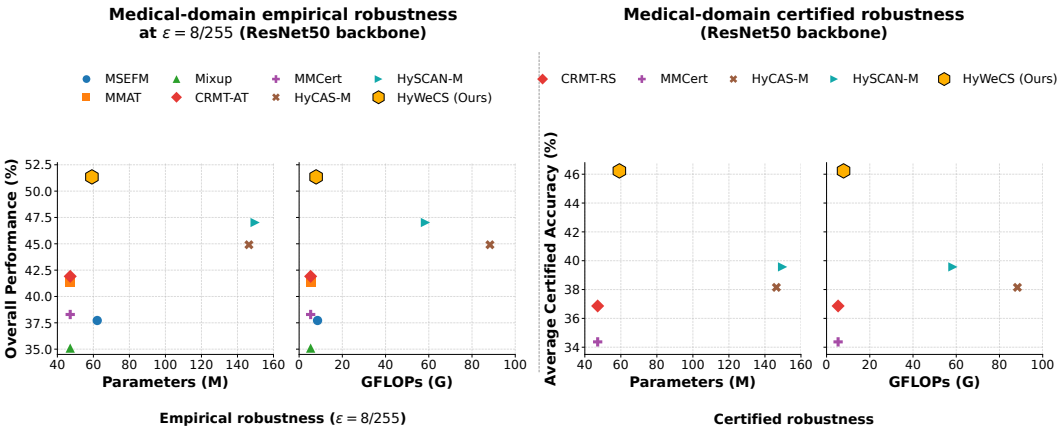


Figure 8: **Robustness–efficiency trade-off on medical-domain data with `ResNet50` backbones.** The left pair reports *Overall Performance* on average across `BRCA`, `BLCA`, `HAM10000`, and `SIPaKMeD` against `EOTPGD` and `BPDA-EOT` at $\epsilon = 8/255$, as a function of parameter count and `GFLOPs`. The right pair reports certified accuracy on average across `BRCA`, `BLCA`, and `HAM10000` over all evaluated smoothing-noise and ℓ_2-radius settings, as a function of the same complexity measures. Points closer to the upper-left corner indicate a more favorable robustness–efficiency trade-off.

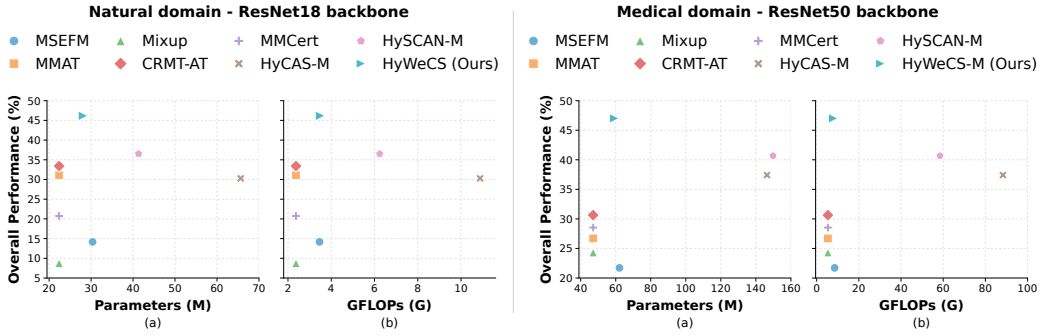


Figure 9: **Robustness–efficiency trade-off on natural- and medical-domain data under stronger adaptive ℓ_∞ attacks at $\epsilon = 16/255$.** The left pair reports *Overall Performance* on average across `RAVDESS`, `KS`, `UCF101`, and `VGGS` against `PGD-EOT` and `BPDA-EOT` with `ResNet18` backbones. The right pair reports *Overall Performance* on average across `BRCA`, `BLCA`, `HAM10000`, and `SIPaKMeD` against the same attacks with `ResNet50` backbones. Points closer to the upper-left corner indicate stronger robustness at lower computational cost.

tings following prior studies (Restrepo et al., 2024; Das et al., 2025; Adebiyi et al., 2024; Suzana & Islam, 2025).

For the natural-domain benchmarks, we use the train, validation, and test splits and corresponding preprocessing protocol of `CRMT` (Yang et al., 2024), whereas the medical-domain benchmarks follow the `CURE` protocol (Dhar et al., 2026c). We assess robustness under two complementary criteria: **(i)** certified accuracy under multimodal ℓ_2 perturbations and **(ii)** empirical accuracy under white-box multimodal ℓ_∞ attacks.

H.2 Network Settings, Execution, and Implementation Details

Baselines. We compare the `HyWeCS`-integrated learner with certified, empirical, and randomized defenses under matched backbone, data-processing, and evaluation protocols.

For certified robustness, we consider `RS-M`, which applies standard Gaussian randomized smoothing (Cohen et al., 2019) to the multimodal learner; `CRMT-RS`, which applies the same smoothing protocol to `CRMT` (Yang et al., 2024); `MMCert` (Wang et al., 2024); `HyCAS-M` (Dhar et al., 2026d); and `HySCAN-M` (Dhar et al., 2026a). For empirical robustness, we compare with `MSEFM` (Tian & Xu, 2021), `MMAT` (Li et al., 2022), `Mixup` (Li et al., 2022), `CRMT-AT` (Yang et al., 2024),

HyCAS−M (Dhar et al., 2026d), and HySCAN−M (Dhar et al., 2026a). Here, CRMT−AT denotes the CRMT backbone trained with standard adversarial training (Madry et al., 2017).

Because HyCAS (Dhar et al., 2026d) and HySCAN (Dhar et al., 2026a) were originally formulated for unimodal learning, we evaluate both their original and multimodal adaptations. We denote the original unimodal realizations by HyCAS−U and HySCAN−U. For HyCAS−M and HySCAN−M, each modality is processed by a separate defended encoder, and the resulting modality-specific representations are fused only after the final encoder stage. These controlled adaptations therefore retain late fusion and do not reuse modality-specific terminal stochastic weights for reciprocal intermediate cross fusion, as performed by HyWeCS in Sec. 2.3 and Eq. 6. For modality-specific attack analysis, we additionally compare with the unimodal defenses RPF (Dong & Xu, 2023), CTRW (Ma et al., 2023), and HyCAS−U (Dhar et al., 2026d).

Backbone and modality protocol. Following the CRMT configuration (Yang et al., 2024), we use a separate ResNet18 encoder for each modality on the natural-domain benchmarks. For the medical-domain benchmarks, we use modality-specific ResNet50 encoders to evaluate the transferability of HyWeCS beyond the ResNet18 setting. All compared methods use the same dataset-specific encoders, fusion head, input preprocessing, and data partitions.

When a non-imaging modality is provided as a D_i-dimensional vector, such as multi-omics, EHR, or tabular data, it is reshaped into a pseudo-image tensor in $\mathbb{R}^{H_i \times W_i \times C_i}$ before entering its modality-specific convolutional encoder. This provides a uniform tensor interface across heterogeneous modalities while preserving the original vector entries.

HyWeCS integration. Within each modality-specific residual block, standard convolutions are replaced by HyWeCS layers. The $r_i^{\ell,b}$ weight locations assigned to ORaW and the complementary $n_i^{\ell,b} - r_i^{\ell,b}$ locations assigned to SHANW are jointly calibrated through Eq. 3. At the end of each residual-block stage, the terminal calibrated weights are reused to transform the paired modality representations through the reciprocal intermediate cross-fusion operation in Eq. 6. Thus, HyWeCS governs both modality-specific representation learning and intermediate cross-modal interaction without an independently parameterized fusion-weight tensor or a separate feature-space stochastic module.

Stochastic execution. The internal realization is denoted by $\Omega_{\mathrm{H}} = (\boldsymbol{\alpha}_{\mathrm{rp}}, \boldsymbol{\eta})$, where $\boldsymbol{\alpha}_{\mathrm{rp}}$ collects the ORaW random-projection draws and $\boldsymbol{\eta}$ collects the Euclidean and hyperbolic SHANW noise draws. During certification-compatible training, Gaussian input noise and Ω_{H} are resampled for every stochastic evaluation, as described in Algorithm 1.

During adversarial training, the attack-phase stochastic sample set is retained throughout one projected-gradient trajectory, while an independent realization is drawn for the parameter-update phase, following Algorithm 3. At inference, every stochastic forward pass uses fresh Gaussian input noise and a fresh internal realization. For expectation-over-transformation attacks, gradients are averaged over the prescribed number of independently sampled realizations.

Certified evaluation. We certify the multimodal predictor under the product norm $\|\cdot\|_{2,\Pi}$ using the expected-logit and probability-based guarantees in Corollary 1. Following HyCAS (Dhar et al., 2026d) and HySCAN (Dhar et al., 2026a), we use a pilot set of $n_0 = 100$ samples to select the candidate class, followed by $n = 100{,}000$ independent samples to estimate its probability. A one-sided Clopper–Pearson confidence bound (Clopper & Pearson, 1934) with failure probability $\alpha = 0.001$ is used following the standard randomized-smoothing certification protocol (Cohen et al., 2019). The expected-logit radius is evaluated separately from the finite-sample probability certificate.

We evaluate $\sigma \in \{0.25, 0.50, 1.0\}$. The dataset-specific radius schedules are:

- For RAVDESS, UCF101, KS, and VGGS, we report certified accuracy at
$$r \in \{0, 0.25, 0.50, 0.75, 1.0, 1.25, 1.50, 2.0\}.$$
- For BRCA, BLCA, HAM10000, and SIPaKMeD, we report certified accuracy at
$$r \in \{0, 0.50, 1.0\}.$$

Empirical evaluation. We evaluate empirical robustness under white-box multimodal ℓ_∞ perturbations, with the internal ORaW–SHANW randomness active throughout evaluation. For stochasticity-aware attacks, each gradient estimate accounts for the sampled internal realizations rather than attacking a deterministic surrogate.

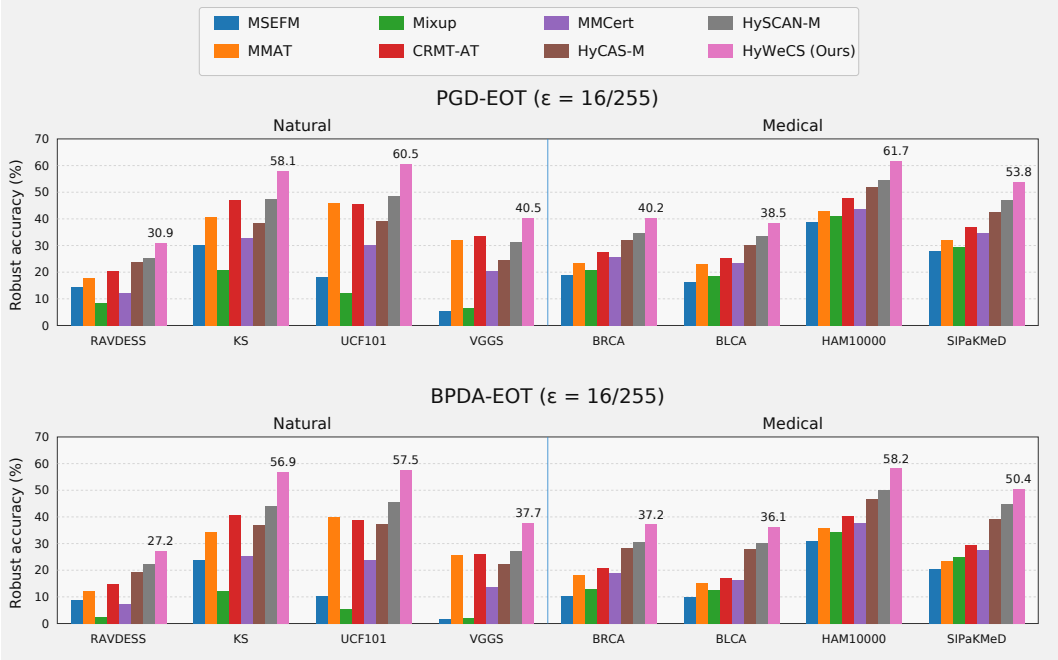


Figure 10: **Robustness under stronger adaptive ℓ_∞ attacks.** Robust accuracy (%) at $\epsilon = 16/255$ under **(top)** `PGD-EOT` and **(bottom)** `BPDA-EOT` across all four natural- and four medical-domain multimodal benchmarks.

The dataset-specific attack settings are:

- On `KS`, `UCF101`, and `VGGS`, we report clean and robust accuracy under `AutoAttack` (Croce & Hein, 2020), `PGD` (Madry et al., 2017), `EOTPGD` (Athalye et al., 2018; Tramer et al., 2020), `BPDA` (Athalye et al., 2018; Tramer et al., 2020), and `BPDA-EOT` (Athalye et al., 2018; Tramer et al., 2020) at $\epsilon \in \{8/255, 16/255\}$.
- On `RAVDESS`, we report robust accuracy under `EOTPGD` and `BPDA-EOT` at $\epsilon \in \{8/255, 16/255\}$.
- On `BRCA`, `BLCA`, `HAM10000`, and `SIPaKMeD`, we report robust accuracy under `EOTPGD` and `BPDA-EOT` at $\epsilon \in \{8/255, 16/255\}$.

All iterative attacks use 20 optimization steps. The EOT-based attacks average over 10 stochastic realizations per gradient estimate. Unless stated otherwise, every experiment is repeated with five random seeds, and we report the mean and standard deviation where applicable.

Training protocol. Certification-compatible training follows Algorithm 1, equivalently setting $\lambda_{\text{adv}} = 0$ in the joint training objective. Each input is perturbed with i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ during training, using the same σ as in certification. When empirical robustness is jointly optimized, we activate the adversarial term and follow Algorithm 3.

All models are trained end-to-end for 200 epochs. For the natural-domain benchmarks, we use `AdamW` with batch size 64, initial learning rate 10^{-2}, weight decay 10^{-5}, and a step learning-rate schedule with step size 30 and decay factor $\gamma = 0.1$. For `BRCA`, `BLCA`, and `HAM10000`, we use `SGD` with batch size 64, learning rate 5×10^{-2}, and a step learning-rate schedule with step size 30 and decay factor $\gamma = 0.8$. Optimization uses the task-specific multimodal learning loss, such as categorical cross-entropy or negative log-likelihood.

I Additional Results

I.1 Robustness under Stronger Adaptive ℓ_∞ Attacks

Figure 10 extends the empirical evaluation to the larger perturbation budget $\epsilon = 16/255$. Under `PGD-EOT`, `HyWeCS` achieves the highest robust accuracy on all eight benchmarks. Its improvement over the strongest competing methods (e.g., `HySCAN-M`, `HyCAS-M`, `CRMT-AT`) ranges from 4.9 to

11.8 percentage points, with an average per-benchmark margin of 7.4 points. The largest gains occur on UCF101 and KS, where HyWeCS improves robust accuracy by 11.8 and 10.6 points, respectively. The advantage also persists across the medical benchmarks, ranging from 4.9 points on BLCA to 7.2 points on HAM10000.

The improvement becomes more pronounced under BPDA−EOT. HyWeCS exceeds the strongest baseline by 4.7–12.8 percentage points across the eight benchmarks, yielding an average margin of 8.3 points. In particular, it improves robust accuracy by 12.8 points on KS, 12.0 points on UCF101, and 10.4 points on VGGS. On the medical benchmarks, the gains remain consistently positive, ranging from 5.7 points on SIPaKMeD to 8.1 points on HAM10000.

On average across all benchmarks, HyWeCS achieves robust accuracies of 48.0% under PGD−EOT and 45.2% under BPDA−EOT. The sustained improvements under attacks that account for stochastic realizations and approximate gradients through defended operations provide evidence that the observed robustness is not restricted to attacks that overlook the internal randomness. Instead, the results support the effectiveness of coordinating ORaW and SHANW within each filter's weight space and carrying the same stochastic realization into reciprocal cross-modal interaction.

I.2 Robustness under Modality-Specific Attacks

Table 6 evaluates a complementary threat model on KS, where only the visual modality (#v) or only the audio modality (#a) is perturbed while the paired modality remains unmodified. This setting determines whether a defense can preserve reliable multimodal prediction when one modality becomes adversarially corrupted.

Clean and visual-modality performance. HyWeCS achieves a clean accuracy of 78.7%, exceeding the strongest competing method, CRMT−AT, by 2.5 percentage points. When only the visual modality is attacked, HyWeCS reaches 72.1% under AutoAttack and 73.8% under PGD. These results improve upon the strongest baselines by 5.8 and 4.0 percentage points, respectively. Thus, the gain is retained across both clean and visual-attack evaluations rather than arising from a trade-off between standard and adversarial accuracy.

Audio-modality robustness. All competing methods experience a larger degradation when the audio modality is perturbed, revealing that this stream constitutes the more vulnerable modality in the evaluated setting. Nevertheless, HyWeCS obtains 54.9% under AutoAttack #a and 58.3% under PGD #a. These values exceed the strongest baselines by 9.0 and 11.9 percentage points, respectively. The larger gains under audio attacks indicate that the proposed stochastic weight-space construction is particularly effective when the less robust modality is corrupted.

Table 6: Performance comparison under modality-specific attacks on KS. We compare unimodal defenses evaluated with their corresponding unimodal learners (RPF (Dong & Xu, 2023), CTRW (Ma et al., 2023), HyCAS−U (Dhar et al., 2026d), and HySCAN−U (Dhar et al., 2026a)) and multimodal defenses integrated with the same fusion learner used by HyWeCS (MSEFM, MMAT, Mixup, CRMT−AT, HyCAS−M, and HySCAN−M). #v and #a indicate attacks on the visual and audio modalities, respectively; w/o denotes evaluation without attack.

Method	w/o	AutoAttack		PGD	
	Clean	#v	#a	#v	#a
Unimodal defenses					
RPF	70.2	61.9	38.7	64.1	39.5
CTRW	74.7	62.3	40.5	65.9	41.8
HyCAS−U	70.9	63.5	41.7	67.3	42.5
HySCAN−U	72.1	64.4	43.6	68.2	44.7
Multimodal defenses					
MSEFM	62.7	54.2	29.7	57.6	31.4
MMAT	65.6	61.4	35.1	63.2	36.9
Mixup	66.9	58.8	13.2	62.6	14.7
CRMT−AT	76.2	66.3	44.5	69.8	46.4
HyCAS−M	71.5	64.1	42.8	68.1	43.9
HySCAN−M	73.4	65.2	45.9	69.5	46.3
HyWeCS (Ours)	**78.7**	**72.1**	**54.9**	**73.8**	**58.3**

Comparison with unimodal and multimodal defenses. Relative to the strongest unimodal defense in each attacked setting, HyWeCS improves robust accuracy by 7.7 points under AutoAttack #v, 11.3 points under AutoAttack #a, 5.6 points under PGD #v, and 13.6 points under PGD #a. Against the strongest multimodal baselines, the corresponding gains are 5.8, 9.0, 4.0, and 11.9 percentage points. Hence, the improvement cannot be attributed solely to the use of a multimodal learner; it remains substantial against both unimodal stochastic defenses and multimodal adversarially trained alternatives.

The visual–audio robustness gap of HyWeCS is 17.2 points under AutoAttack and 15.5 points under PGD. In comparison, the corresponding gaps are 19.3/23.2 points for HySCAN−M and 21.8/23.4

points for CRMT−AT. This reduced disparity indicates a more balanced use of the two modalities. It is also consistent with the intended role of reciprocal intermediate cross fusion in Eq. 6, where the terminal stochastic weights learned from each modality regulate the paired modality representation. Consequently, HyWeCS preserves stronger multimodal predictions even when one input stream is independently compromised.

I.3 Optimal Robustness–Efficiency Trade-off

Figures 7–9 examine whether the robustness improvements of HyWeCS are obtained merely by increasing model capacity. Each method is compared in terms of robustness, parameter count, and GFLOPs; therefore, points closer to the upper-left region represent higher robustness at lower computational cost. Across all evaluated settings, HyWeCS consistently lies on the upper performance envelope while remaining substantially lighter than the strongest hybrid randomized baselines.

On the natural-domain benchmarks with ResNet18 backbones (Fig. 7), HyWeCS achieves an *Overall Performance* of 56.96% on average across KS, UCF101, and VGGS against AutoAttack, PGD, BPDA, and BPDA−EOT at $\epsilon = 8/255$. It also obtains an average certified accuracy of 34.01% across RAVDESS and UCF101 over the evaluated smoothing-noise and ℓ_2-radius settings, while requiring only 28.10M parameters and 3.489 GFLOPs. Compared with HySCAN−M, HyWeCS improves empirical and certified robustness by 6.50 and 3.78 percentage points, respectively, while reducing parameters by 32.0% and GFLOPs by 44.1%. Compared with HyCAS−M, the corresponding robustness gains increase to 10.24 and 4.82 points, with reductions of 57.2% in parameters and 67.9% in GFLOPs.

The advantage becomes more pronounced on the medical-domain benchmarks with ResNet50 backbones (Fig. 8). HyWeCS achieves 51.35% empirical robustness on average across BRCA, BLCA, HAM10000, and SIPaKMeD against EOTPGD and BPDA−EOT at $\epsilon = 8/255$, together with an average certified accuracy of 46.23% across the reported medical-domain certification settings. These results are obtained with 59.176M parameters and 7.862 GFLOPs. Relative to HySCAN−M, HyWeCS improves empirical and certified robustness by 4.33 and 6.66 points while using 60.5% fewer parameters and 86.6% fewer GFLOPs. Relative to HyCAS−M, it improves the two robustness measures by 6.44 and 8.08 points while reducing parameters and GFLOPs by 59.6% and 91.1%, respectively.

This behavior remains consistent under the larger perturbation budget $\epsilon = 16/255$ (Fig. 9). Against PGD−EOT and BPDA−EOT, HyWeCS achieves an *Overall Performance* of 46.16% on average across all four natural-domain benchmarks and 47.01% on average across all four medical-domain benchmarks. These results exceed the strongest competing method (HySCAN−M) by 9.64 and 6.34 percentage points, respectively, without increasing the parameter count or GFLOPs used at $\epsilon = 8/255$.

The observed gains therefore cannot be attributed to a larger backbone or an indiscriminate increase in stochastic capacity. Instead, HyWeCS coordinates ORaW and SHANW at complementary weight locations within every filter and reuses the terminal stochastic weights for reciprocal intermediate cross fusion. This avoids an independently parameterized feature-space stochastic path while retaining both modality-specific and cross-modal randomization. Consequently, HyWeCS Pareto-dominates the more expensive hybrid baselines and constitutes an empirically optimal robustness–efficiency trade-off among the evaluated defenses.

J Ablation Studies

We conduct two complementary studies to determine whether the robustness improvements arise from the proposed stochastic mechanisms rather than from increased capacity or generic architectural randomization. First, we isolate the individual and joint contributions of ORaW and SHANW, followed by the random-projection and LOT principles underlying ORaW. Second, we replace each proposed component with representative stochastic alternatives while preserving the same standard-convolution backbone, within-filter allocation, multimodal fusion topology, calibration, optimization, and evaluation protocols.

Components		RAVDESS		UCF101	
ORaW	SHANW	Cert.	EOTPGD	Cert.	EOTPGD
✗	✗	21.2	14.3	19.8	34.5
✗	✓	34.7	30.1	32.3	53.7
✓	✗	30.4	27.5	29.2	48.8
✓	✓	**42.8**	**38.7**	**39.3**	**62.1**

(a) Core HyWeCS components

ORaW Components		RAVDESS		UCF101	
RP	LOT	Cert.	EOTPGD	Cert.	EOTPGD
✗	✗	34.7	30.1	32.3	53.7
✓	✗	39.6	35.2	36.4	58.5
✗	✓	37.9	33.8	35.5	57.2
✓	✓	**42.8**	**38.7**	**39.3**	**62.1**

(b) Random projection and LOT in ORaW

Table 7: **Ablation of the principal HyWeCS components.** Results on RAVDESS and UCF101 report certified accuracy (Cert.) at $(r, \sigma) = (0.75, 0.25)$ and EOTPGD robustness at $\epsilon = 8/255$ using 20 steps and 10 stochastic evaluations per gradient. **(a)** ORaW–SHANW complementarity with capacity-matched deterministic replacements. **(b)** Contributions of random projection (RP) and LOT within ORaW, with SHANW fixed.

J.1 Component-Wise Ablation of HyWeCS

Table 7 examines whether ORaW and SHANW contribute independently and whether their coordination provides additional robustness. Whenever one component is disabled, its allocated locations are replaced by ordinary trainable weights, thereby preserving the filter dimensions and trainable capacity. Panel (b) further decomposes ORaW while retaining SHANW and all remaining architectural components.

Complementarity of ORaW and SHANW. Panel (a) demonstrates that both stochastic principles independently improve certified and empirical robustness. On RAVDESS, SHANW alone improves certified and EOTPGD accuracy over the deterministic replacement by 13.5 and 15.8 percentage points, respectively, while ORaW alone yields gains of 9.2 and 13.2 points. On UCF101, the corresponding improvements are $12.5/19.2$ points for SHANW and $9.4/14.3$ points for ORaW. Thus, each component contributes a distinct robustness benefit rather than merely compensating for the absence of the other.

Their within-filter coordination produces a further consistent gain. Relative to SHANW alone, the complete formulation improves certified and EOTPGD accuracy by $8.1/8.6$ points on RAVDESS and $7.0/8.4$ points on UCF101. Relative to ORaW alone, the corresponding gains are $12.4/11.2$ and $10.1/13.3$ points. Overall, the complete HyWeCS realization outperforms the deterministic replacement by $21.6/24.4$ points on RAVDESS and $19.5/27.6$ points on UCF101. These results substantiate the central design principle of HyWeCS: orthogonally retracted projection diversity and progressive hybrid-space attention noise provide complementary robustness when jointly realized within each filter's weight space.

Contributions of random projection and LOT. Panel (b) further identifies the roles of the two principles underlying ORaW. Adding random-projection resampling to the SHANW-only reference improves certified and EOTPGD accuracy by $4.9/5.1$ points on RAVDESS and $4.1/4.8$ points on UCF101. Introducing LOT without projection resampling provides improvements of $3.2/3.7$ and $3.2/3.5$ points, respectively.

Their combination remains consistently stronger than either principle in isolation. The complete ORaW construction exceeds the random-projection-only variant by $3.2/3.5$ points on RAVDESS and $2.9/3.6$ points on UCF101, while surpassing the LOT-only variant by $4.9/4.9$ and $3.8/4.9$ points. This pattern reflects their complementary roles: random projection diversifies the realized weights, whereas LOT places those realizations within a common orthogonal convolutional geometry (Xu et al., 2022). Their coordination therefore retains stochastic projection diversity while supplying the structured geometry subsequently governed by the shared HyWeCS calibration.

J.2 Controlled Replacement of Stochastic Mechanisms

We next investigate whether existing stochastic mechanisms can recover the same certified–empirical robustness. Panel (a) of Table 8 keeps SHANW fixed and replaces ORaW. The comparison includes filter-space random projection through RPF (Dong & Xu, 2023) and SN-RPF (Dhar et al., 2026d), weight-space randomization through CTRW, and the proposed LOT-retracted weight-space realization.

Panel (b) keeps ORaW fixed and replaces SHANW only at the same complementary $n - r$ weight locations. The RANI-W control reconfigures the Euclidean attention-noise response of RANI into

Replacement	RAVDESS		UCF101		#P	#F
	Cert.	EOTPGD	Cert.	EOTPGD		
RPF	36.9	32.4	34.1	55.7	**23.05**	**3.39**
CTRW	38.5	34.8	36.2	57.9	**23.05**	3.42
SN-RPF	37.2	33.5	34.9	56.8	39.80	5.57
ORaW (Ours)	**42.8**	**38.7**	**39.3**	**62.1**	28.08	4.82

(a) Non-smoothing-aware replacement; SHANW fixed

Replacement	RAVDESS		UCF101		#P	#F
	Cert.	EOTPGD	Cert.	EOTPGD		
RANI-W	39.1	34.8	35.5	57.4	**27.60**	**4.37**
RWAN	41.5	37.3	37.9	60.7	35.24	5.34
SHANW (Ours)	**42.8**	**38.7**	**39.3**	**62.1**	28.08	4.82

(b) Smoothing-aware replacement; ORaW fixed

Table 8: **Controlled replacement of stochastic mechanisms on RAVDESS and UCF101.** The certified and EOTPGD settings follow Table 7; #P and #F denote parameters (M) and FLOPs (G), respectively. **(a)** With SHANW fixed, ORaW is compared with RPF (Dong & Xu, 2023), CTRW (Ma et al., 2023), and SN-RPF, our controlled spectrally normalized random-projection variant derived from RPFAN (Dhar et al., 2026d). **(b)** With ORaW fixed, SHANW is compared with RANI-W, our weight-aligned control derived from RANI (Dhar et al., 2026d), and RWAN (Dhar et al., 2026a), restricted to the same complementary $n - r$ weight locations. All variants retain the same backbone, allocation, fusion, calibration, smoothing, and training settings.

weight-aligned perturbations, with its original feature-space injection removed. The RWAN control retains its attention-gated Gaussian weight-perturbation rule but applies it only to the same $n - r$ locations rather than to all coefficients of every filter. Consequently, all panel-(b) variants use the same weight-space capacity and the same standard-convolution backbone.

Non-smoothing-aware stochastic mechanisms. Panel (a) shows that ORaW consistently achieves the strongest certified and adaptive-attack robustness. Relative to RPF, it improves certified and EOTPGD accuracy by $5.9/6.3$ percentage points on RAVDESS and $5.2/6.4$ points on UCF101. The corresponding improvements over CTRW are $4.3/3.9$ and $3.1/4.2$ points, respectively. Compared with the SN-RPF control, ORaW improves the two robustness measures by $5.6/5.2$ points on RAVDESS and $4.4/5.3$ points on UCF101.

These improvements arise from a distinction in both stochastic locus and geometric control. RPF and SN-RPF organize projection randomness over complete filters, whereas CTRW randomizes convolutional weights without retracting each sampled tensor into a common orthogonal convolutional family. In contrast, ORaW introduces projection diversity at selected locations within every filter and applies the LOT retraction before integrating those weights with SHANW.

This construction is particularly relevant to certificate-compatible geometry. Spectral normalization of convolutional operators commonly relies on an estimated or relaxed operator bound, while unrolling a convolution into a linear matrix can lead to loose Lipschitz estimates (Xu et al., 2022). LOT instead parameterizes orthogonal convolutions directly through a Fourier-domain inverse-square-root transformation. Its parameterization is invariant to radial rescaling, avoids the norm-dependent identity bias identified for exponential orthogonal parameterizations, represents a broader orthogonal family, and directly handles unequal input and output channel dimensions (Xu et al., 2022). Accordingly, ORaW combines fine-grained weight-space projection diversity with deterministic orthogonal geometry, after which the shared HyWeCS calibration constrains the complete integrated mapping.

The resulting robustness does not require the larger computation of the filter-space SN-RPF control. ORaW uses 28.08M parameters and 4.82G FLOPs, reducing the SN-RPF configuration by 11.72M parameters and 0.75G FLOPs, corresponding to reductions of approximately 29.4% and 13.5%. Simultaneously, it improves all four reported robustness measurements. This comparison indicates that LOT-retracted weight-space projection can provide stronger certified–empirical robustness with a more compact realization than the spectrally normalized filter-space alternative.

Smoothing-aware weight-space stochastic mechanisms. Panel (b) compares the attention-noise alternatives at exactly the same complementary $n - r$ weight locations. Replacing SHANW with RANI-W reduces certified and EOTPGD accuracy by $3.7/3.9$ points on RAVDESS and $3.8/4.7$ points on UCF101. The stronger RWAN control narrows this gap; nevertheless, SHANW remains ahead by $1.3/1.4$ points on RAVDESS and $1.4/1.4$ points on UCF101.

The comparison is location matched while preserving the defining stochastic rule of each mechanism. The original RWAN formulation samples Gaussian noise with the dimensions of the convolutional kernel and modulates it through an input-conditioned Euclidean attention response (Dhar et al., 2026a). It ordinarily permits attention-gated perturbations over all trainable coefficients of

every filter. In this controlled study, the same rule is restricted to the $n - r$ locations allocated to the smoothing-aware component, ensuring identical weight-space capacity across RWAN and SHANW. The RWAN perturbation rule itself does not impose an orthogonal retraction, spectral normalization, or an explicit realization-wise 1- or 2-Lipschitz constraint. Its certified accuracy in this table therefore follows from the common layer-wise calibration and Gaussian smoothing used for all variants.

The advantage of SHANW is consequently attributable to its stochastic construction rather than to a different perturbation locus. Both RANI−W and RWAN derive their attention-noise responses in Euclidean space. In contrast, SHANW independently samples Euclidean and hyperbolic noise, self-modulates both realizations, and adaptively combines their noise-conditioned attention responses during progressive weight generation. The Euclidean component captures locally measured coordinate variation, whereas the hyperbolic component introduces curvature-aware geodesic variation. Their adaptive fusion therefore diversifies the stochastic neighborhoods explored by the complementary weights instead of merely increasing the variance of a single Euclidean noise source.

The consistent gains over both Euclidean controls support the benefit of this hybrid-space formulation on the evaluated multimodal benchmarks. This result does not imply that one geometry is universally preferable; rather, the learned geometry weighting permits the model to emphasize the stochastic response that is more informative for the current representation.

SHANW also yields a favorable computational profile relative to RWAN. It uses 7.16M fewer parameters and 0.52G fewer FLOPs, corresponding to reductions of approximately 20.3% and 9.7%, while improving every reported robustness value. Relative to the lighter RANI−W control, SHANW adds only 0.48M parameters and 0.45G FLOPs, yet improves certified accuracy by up to 3.8 points and EOTPGD robustness by up to 4.7 points.

Together, the factorial and controlled-replacement studies show that the performance of HyWeCS is not explained by generic randomization. ORaW benefits from coordinating fine-grained projection resampling with LOT-induced orthogonal geometry, whereas SHANW benefits from progressively combining Euclidean and hyperbolic attention-noise responses at the complementary weight locations. Their integration within every filter, followed by shared Lipschitz calibration, provides the strongest certified and empirical robustness while maintaining a favorable computational profile.

K Discussion

Why and how HyWeCS improves certified and empirical robustness. The central design principle of HyWeCS is to coordinate heterogeneous stochastic mechanisms directly within the weight space of every convolutional filter while preserving certificate-compatible sensitivity control. Specifically, ORaW introduces non-smoothing-aware random-projection diversity under LOT-retracted geometry, whereas SHANW introduces smoothing-aware, feature-conditioned stochasticity through self-modulated Euclidean–hyperbolic attention responses. Their complementary allocation ensures that the two mechanisms jointly exploit distinct stochastic principles without modifying the LOT-retracted ORaW coefficients through overlapping perturbations. The resulting hybrid weights are subsequently governed by the shared calibration in Eq. 3, thereby coupling stochastic diversity with a common layer-wise robustness constraint.

This construction directly supports certified robustness. For every admissible stochastic realization, the calibrated layer satisfies $\mathrm{Lip}_2(\mathcal{G}_i^{\ell,b}) \le \tau_i^{\ell,b} \le 1$. Together with the residual and reciprocal cross-fusion structure, Theorem 1 establishes a uniform end-to-end 2-Lipschitz score bound. Corollary 1 further connects this controlled geometry with formal ℓ_2 certification through margin-based and joint-randomness randomized-smoothing guarantees. Importantly, the latter jointly marginalizes Gaussian input perturbations and the internal ORaW–SHANW realizations, aligning the certificate with the stochastic learner used during inference.

The same controlled stochastic geometry also supports empirical robustness under ℓ_∞ perturbations. Theorem 2 establishes theoretical control over the effect of bounded ℓ_∞ perturbations on the multimodal learning objective, while the empirical evaluation directly examines the resulting learner under strong white-box attacks. At $\epsilon = 8/255$, the improvements persist across PGD, AutoAttack, BPDA, EOTPGD, and BPDA−EOT. At the stronger $\epsilon = 16/255$ setting, HyWeCS improves over the strongest competing defense by 4.9–11.8 points under PGD−EOT and 4.7–12.8 points under BPDA−EOT. Thus, the same stochastic architecture supports both formal ℓ_2 certification and strong empirical ℓ_∞ robustness without requiring separate defense pipelines.

Why both within-filter stochastic mechanisms are required. ORaW and SHANW play complementary rather than redundant roles. Assigning entire filters to one mechanism would leave every filter governed by only one stochastic principle, postponing their interaction until downstream feature aggregation. Conversely, applying both mechanisms to the same coefficients would alter the values obtained through LOT retraction. HyWeCS therefore assigns $r_i^{\ell,b}$ weight locations to ORaW and the complementary $n_i^{\ell,b} - r_i^{\ell,b}$ locations to SHANW. This construction allows every filter to jointly exploit projection diversity and smoothing-aware stochasticity while preserving their respective weight-generation processes.

Their stochastic mechanisms are also fundamentally different. For ORaW, a random projection is resampled before LOT retraction, introducing variability within a structured orthogonal convolutional geometry. For SHANW, independently sampled Gaussian variables self-modulate Euclidean- and hyperbolic-derived attention responses. Noise-conditioned geometry selection then determines their relative contributions at each complementary weight location, while recursive accumulation progressively constructs the resulting stochastic weights. Consequently, ORaW contributes projection-driven stochastic diversity under geometric control, whereas SHANW contributes representation-dependent stochasticity that adapts according to complementary Euclidean and non-Euclidean evidence.

The component-wise ablations strongly support this complementarity. Relative to deterministic replacements, SHANW alone improves certified/EOTPGD accuracy by $13.5/15.8$ points on RAVDESS and $12.5/19.2$ points on UCF101. ORaW alone provides corresponding gains of $9.2/13.2$ and $9.4/14.3$ points. Their joint realization increases the gains to $21.6/24.4$ points on RAVDESS and $19.5/27.6$ points on UCF101 relative to the deterministic replacement (Table 7). The controlled-replacement experiments provide complementary evidence: ORaW consistently exceeds RPF, CTRW, and SN−RPF, whereas SHANW remains stronger than the location-matched RANI−W and RWAN alternatives (Table 8). Together, these results show that the improvement arises from the specific coordination of the two stochastic mechanisms rather than from simply increasing randomness.

Role of reciprocal intermediate cross fusion. A second distinguishing property of HyWeCS is that its stochastic weight geometry is not confined to modality-specific feature extraction. At each residual-block stage, the terminal stochastic weights learned from one modality are reciprocally reused to transform the paired modality representation through Eq. 6. The same HyWeCS realization therefore participates in both modality-specific encoding and intermediate cross-modal interaction.

This reuse is particularly important for efficiency and multimodal robustness. Rather than introducing a separately parameterized stochastic fusion module, reciprocal cross fusion directly reuses the already generated terminal weights. Accordingly, HyWeCS extends stochasticity into the fusion process without a separate feature-space stochastic learner or independently generated fusion weights. It also allows the stochastic characteristics learned from each modality to influence the representation of its paired modality, creating a direct connection between modality-specific robustness and cross-modal interaction.

The modality-specific attack results in Table 6 are consistent with this behavior. On KS, HyWeCS achieves 72.1% and 73.8% robust accuracy when only the visual modality is attacked using AutoAttack and PGD, respectively. When only the audio modality is perturbed, it maintains 54.9% and 58.3% robust accuracy. These results exceed the strongest multimodal baselines across both attack settings and reduce the visual–audio robustness disparity relative to HySCAN−M and CRMT−AT. This behavior supports reciprocal intermediate cross fusion as an effective mechanism for carrying stochastic robustness across interacting modalities.

Robustness across natural and medical domains. An important property of HyWeCS is that its robustness mechanism does not depend on domain-specific priors. It does not require lesion masks, organ-specific constraints, scanner-specific corrections, pathology rules, or handcrafted clinical features. Instead, robustness is introduced through weight-space stochasticity, Lipschitz calibration, and reciprocal cross-modal interaction. This architecture-level formulation allows the same principle to operate across substantially different modality combinations and data distributions.

The experimental results demonstrate this transfer across both natural- and medical-domain benchmarks. On the natural-domain suite, HyWeCS improves certified robustness on RAVDESS and UCF101 and empirical robustness across KS, UCF101, and VGGS. On the medical-domain suite, it improves certified ℓ_2 robustness by up to 12.4% over the strongest competing defense (Table 2). At $\epsilon = 8/255$, the gains further extend to adaptive ℓ_∞ attacks on BRCA, BLCA, HAM10000, and

SIPaKMeD (Table 4). Under the stronger $\epsilon = 16/255$ setting, positive improvements persist under both PGD−EOT and BPDA−EOT (Fig. 10). The consistency across audio–visual, RGB–optical-flow, imaging–multi-omics, and image–tabular settings indicates that the proposed stochastic principle is not tied to a particular modality pair or data domain.

What the experiments show and imply. The empirical evidence provides several complementary observations. First, the certified improvements persist across multiple Gaussian smoothing scales and certified radii (Tables 1–2), indicating that the advantage is not restricted to one certification operating point. Second, the empirical improvements extend from conventional attacks to adaptive stochasticity-aware attacks (Tables 3–4). In particular, the gains under EOTPGD and BPDA−EOT remain when the attacker explicitly accounts for stochastic realizations.

Third, Fig. 6 shows that HyWeCS degrades gradually as the perturbation strength, number of PGD steps, and number of EOT evaluations increase. Its robustness advantage is therefore sustained across progressively stronger attack configurations rather than being associated with one specific optimization setting. The larger-budget results in Fig. 10 reinforce this observation at $\epsilon = 16/255$.

Fourth, the factorial ablations identify complementary contributions from ORaW and SHANW, while the decomposition of ORaW shows that random-projection resampling and LOT retraction each contribute to the final robustness (Table 7). Finally, the controlled replacements demonstrate that representative filter-space randomization, unconstrained random-weight mechanisms, and Euclidean attention-noise alternatives do not recover the same certified–empirical robustness under matched experimental conditions (Table 8). Collectively, these findings support the central hypothesis of HyWeCS: robust multimodal fusion benefits from coordinating heterogeneous stochastic principles inside every filter and extending the same stochastic realization into cross-modal interaction.

Robustness–efficiency trade-off. The robustness improvements of HyWeCS are achieved without increasing model capacity relative to the strongest hybrid randomized baselines. On the natural-domain benchmarks with ResNet18, HyWeCS uses 28.10M parameters and 3.489 GFLOPs. Relative to HySCAN−M, it reduces parameters and GFLOPs by 32.0% and 44.1%, respectively, while improving empirical and certified robustness by 6.50 and 3.78 points. Relative to HyCAS−M, the corresponding parameter and GFLOP reductions increase to 57.2% and 67.9% (Fig. 7).

The efficiency advantage becomes more pronounced on the medical-domain benchmarks with ResNet50. HyWeCS uses 59.176M parameters and 7.862 GFLOPs. Compared with HySCAN−M, it reduces parameters by 60.5% and GFLOPs by 86.6%, while improving empirical and certified robustness by 4.33 and 6.66 points. Compared with HyCAS−M, the corresponding reductions are 59.6% and 91.1% (Fig. 8). The same model configurations preserve their robustness advantage at $\epsilon = 16/255$ (Fig. 9).

This favorable robustness–efficiency behavior follows directly from the architectural organization of HyWeCS. Complementary stochasticity is realized inside the existing convolutional weight space, while terminal stochastic weights are reused during reciprocal cross fusion. The design therefore avoids a separately parameterized feature-space stochastic pathway while retaining both modality-specific and cross-modal randomization. The resulting learner achieves a stronger robustness–computation trade-off than the evaluated hybrid randomized defenses across both backbone families.

Asymptotic cost and computational considerations. We characterize the computational cost of HyWeCS relative to a standard multimodal convolutional backbone. For modality i and layer ℓ, let $S_i^\ell = H_i^\ell W_i^\ell$ denote the spatial size, $C_{\mathrm{in},i}^\ell$ and $C_{\mathrm{out},i}^\ell$ the input and output channel dimensions, and k_i^ℓ the kernel width. The underlying convolution requires

$$T_{\mathrm{conv},i}^\ell = \Theta\big(S_i^\ell (k_i^\ell)^2 C_{\mathrm{in},i}^\ell C_{\mathrm{out},i}^\ell\big)\,.$$

For ORaW, let G_i^ℓ denote the number of Fourier frequencies used by the LOT realization, $d_i^\ell = \min\{C_{\mathrm{in},i}^\ell, C_{\mathrm{out},i}^\ell\}$, $\bar{C}_i^\ell = \max\{C_{\mathrm{in},i}^\ell, C_{\mathrm{out},i}^\ell\}$, and J the number of inverse-square-root iterations. A dense realization has upper-order cost

$$T_{\mathrm{ORaW},i}^\ell = \mathcal{O}\Big(G_i^\ell C_{\mathrm{in},i}^\ell C_{\mathrm{out},i}^\ell \log G_i^\ell + G_i^\ell (d_i^\ell)^2 \bar{C}_i^\ell + G_i^\ell J (d_i^\ell)^3 + (k_i^\ell)^2 C_{\mathrm{in},i}^\ell C_{\mathrm{out},i}^\ell\Big),$$

where the terms correspond to Fourier transformation/reconstruction, frequency-wise Gram construction and application, iterative inverse-square root, and random-projection weight generation, respectively.

For SHANW, let $C = C^{\ell}_{\mathrm{in},i}$, d_{H} denote the hyperbolic dimension, d_c the Euclidean channel bottleneck, $\mathcal{P}$ the set of pooled statistics, and $d_{\mathrm{sh}} = n^{\ell}_i - r^{\ell}_i$ the number of complementary SHANW locations in each output filter. For recursion depth T, its principal additional cost is

$$T^{\ell}_{\mathrm{SHANW},i} = \mathcal{O}\Big(S^{\ell}_i C(C + d_{\mathrm{H}}) + |\mathcal{P}|S^{\ell}_i C + |\mathcal{P}|C(d_c + d_{\mathrm{H}}) + T\, C^{\ell}_{\mathrm{out},i} d_{\mathrm{sh}}\Big),$$

covering the local Euclidean and hyperbolic projections, multi-statistic descriptors, geometry-specific channel responses, and recursive location-wise weight generation. Within-filter support embedding and shared layer-wise calibration scale linearly with the number of generated coefficients.

The resulting stochastic forward-pass cost can therefore be written as

$$T^{\mathtt{HyWeCS}}_{\mathrm{fwd}} = T_{\mathrm{conv}} + \sum_{i,\ell} \big(T^{\ell}_{\mathrm{ORaW},i} + T^{\ell}_{\mathrm{SHANW},i}\big) + T_{\mathrm{cross}},$$

where T_{cross} denotes the reciprocal application of terminal HyWeCS weights to the paired modality representations. Because these terminal weights are directly reused, cross-modal interaction does not require an additional stochastic fusion-weight generator.

Training introduces multiplicative costs associated with stochastic evaluation and adversarial optimization. Let $T^{\mathtt{HyWeCS}}_{\mathrm{fb}}$ denote the forward–backward cost of one stochastic realization. For mini-batch size B and Q_{tr} stochastic samples, certification-compatible training requires

$$T_{\mathrm{train,cert}} = \mathcal{O}\big(BQ_{\mathrm{tr}} T^{\mathtt{HyWeCS}}_{\mathrm{fb}}\big)\,.$$

For adversarial training with K projected-gradient steps, Q_A attack-phase stochastic samples, Q_U parameter-update samples, and total multimodal input dimension D, Algorithm 3 requires

$$T_{\mathrm{train,adv}} = \mathcal{O}\big(B(KQ_A + Q_U) T^{\mathtt{HyWeCS}}_{\mathrm{fb}} + BKD\big)\,,$$

up to constant factors associated with the clean and adversarial update terms.

Inference-time certification follows the standard Monte Carlo scaling of randomized smoothing. With pilot size n_0 and certification size n, Algorithm 2 requires

$$T_{\mathrm{cert}}(X) = \Theta\big((n_0 + n) T^{\mathtt{HyWeCS}}_{\mathrm{fwd}}\big)$$

up to class-counting and score-accumulation operations. For the reported evaluation, $n_0 = 100$ and $n = 100{,}000$. The samples can be processed sequentially, allowing class counts and score statistics to be accumulated online without storing all stochastic realizations simultaneously. Hence, the certification memory footprint is primarily governed by the model, one forward-pass activation footprint, and class-level accumulators.

Social impact implications of HyWeCS. The broader implication of HyWeCS is its potential to strengthen the reliability of multimodal systems in applications where decisions depend on several heterogeneous information sources. This is particularly relevant when corruption of a single modality can propagate through cross-modal interaction and influence the final prediction. By improving both certified and empirical robustness across natural- and medical-domain benchmarks, HyWeCS provides a general architecture-level approach rather than relying on task-specific robustness interventions.

Its application to medical-domain multimodal learning is especially relevant for robust clinical decision-support research, where imaging, omics, and tabular measurements can exhibit substantially different noise and acquisition characteristics. The results indicate that certificate-compatible stochastic weight-space learning can improve robustness across such heterogeneous sources without introducing handcrafted medical priors. More broadly, the same principle may benefit other multimodal systems that integrate visual, acoustic, sensor, or structured information.

Reproducible use of stochastic defenses also benefits from transparent reporting of smoothing scales, stochastic sample counts, confidence levels, attack configurations, and EOT settings. For high-impact applications, these robustness mechanisms can complement standard practices such as independent validation, uncertainty-aware prediction, and human-in-the-loop decision support.

Future directions. The current study focuses on convolutional multimodal fusion. A natural extension is to investigate within-weight-space stochastic complementarity in transformer-based vision–language models and multimodal foundation models, where stochastic mechanisms must interact with self-attention, token-level alignment, and substantially larger parameter spaces. Further work can explore scalable approximations to LOT retraction and randomized-smoothing certification, tighter reciprocal cross-fusion sensitivity analysis, structured modality corruption, and missing-modality settings.

Another promising direction is to investigate whether ORaW–SHANW complementarity can be generalized beyond convolutional weights to attention projections, adapter parameters, and cross-modal routing layers. Such extensions would examine whether the central principle of HyWeCS—jointly exploiting heterogeneous stochastic mechanisms within a shared parameter space under certificate-compatible control—can provide a general foundation for robust multimodal learning at larger scales.