# Modality-Aware SAM: Sharpness-Aware-Minimization Driven Gradient Modulation for Harmonized Multimodal Learning

NeurIPS 2025

Hossein Rajoli Nowdeh, Jie Je, Xiaolong Ma, Fatemeh Afghah

# The Challenge of Harmonized Learning in Multimodal Models

Human perception is inherently multimodal, we integrate sounds, visuals, and language seamlessly. Deep learning aims to mimic this through multimodal models, **but**:

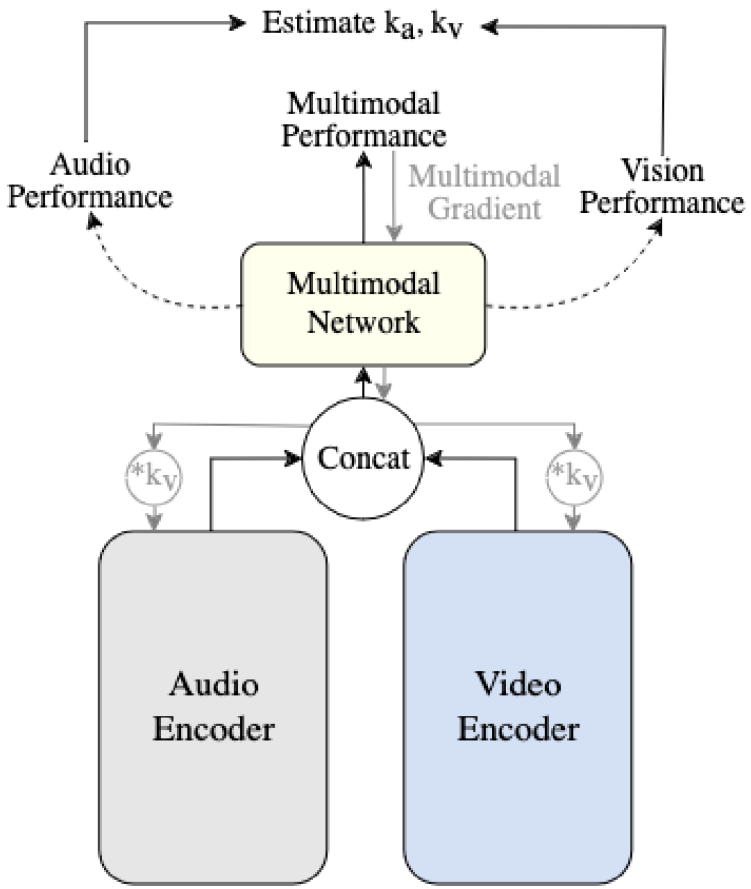
**Modality Dominance**

One strong modality often overshadows weaker ones, limiting the model's ability to learn diverse, complementary features.

**Uncoordinated Convergence**

Each modality converges at a different speed. Encoders and shared networks become misaligned.

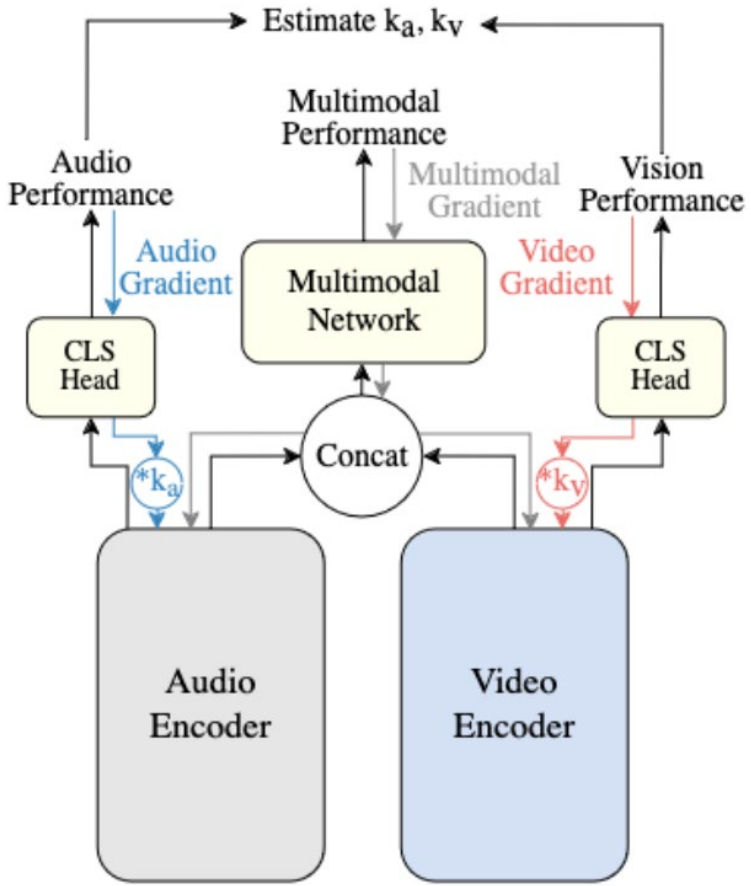
# Typical methods (Gradient Balancing Techniques)




**Key Strategy:** Adjust Unimodal Encoder Learning Rates

- Modulating gradient flow for each modality during training.
- Adjustment is done dynamically on learning rates or through intermittent learning
- Decision-level fusion Simple fusion
- They tune each modality neural path independently
- Neural paths are separated through the architecture
- They all use unimodal performance (signal) so they support limited very late fusion
- Validation Delta or Gradient Norm

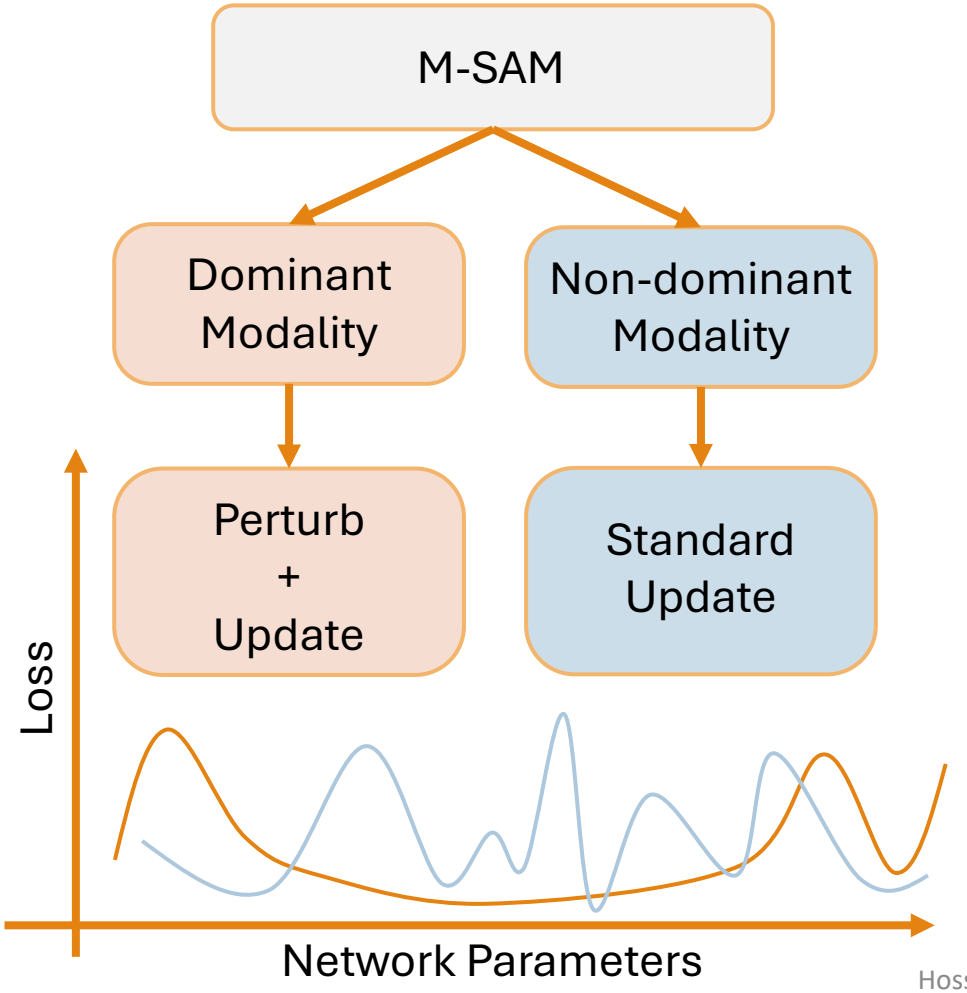
# Typical methods (Multi-task Learning)




**Key Strategy:** Use Dedicated Unimodal Losses to Improve Multimodal Encoders

- In a bi-modality system it has 3 terms of loss:
  $$L_T = L_{mm} + L_{m1} + L_{m2}$$
- This approach tends to extract most discriminative features
- Not necessarily gradients are aligned
- Teacher-student supervisory signal can be adopted from stronger modality
- DLM losses can be applied to align representations
- Limited to simple fusion strategies

# M-SAM: Gradient Modulation for Balanced Learning




**M-SAM selectively applies SAM optimization**

- SAM is in favor of dominant modality
- SAM would Not be beneficial for other modalities

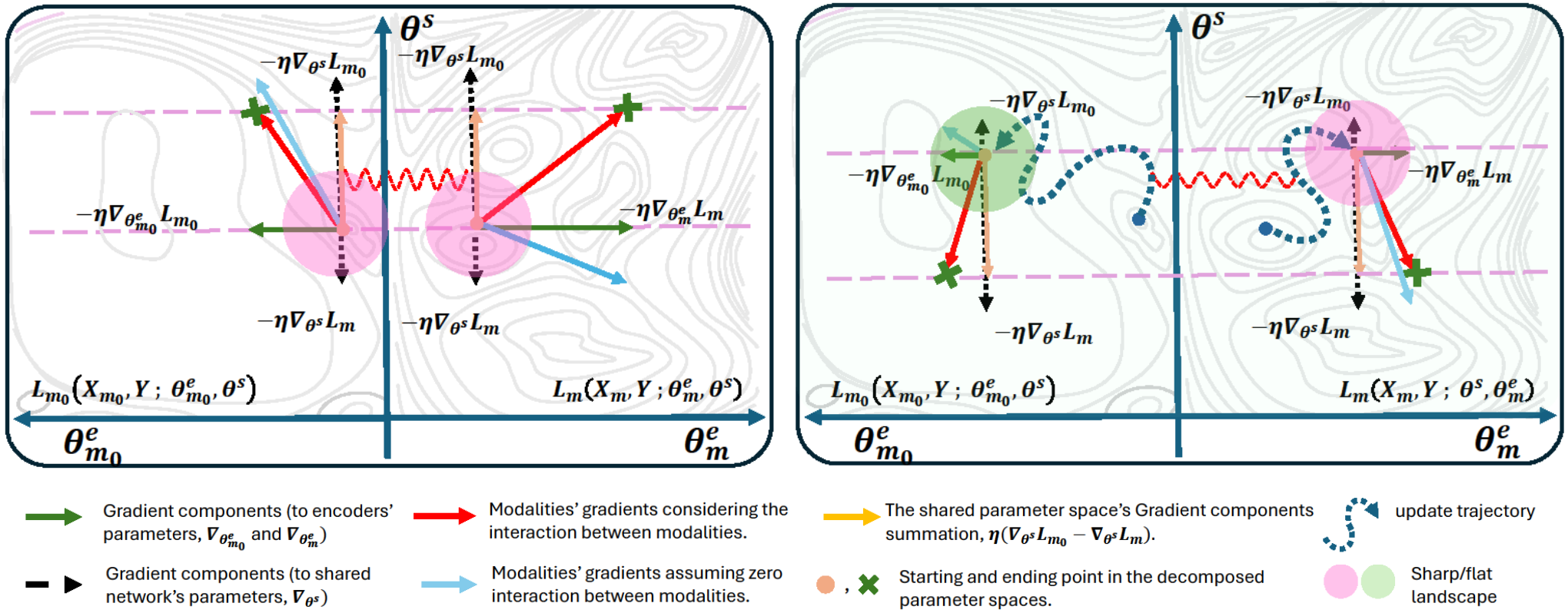
**Shapley based contribution analysis**

- Tracks which modality drives performance
- Indirectly help in loss surface decomposition

**What we seek for**

- The dominant modality lands in the flat minima
- Non-dominant modality explore more freely (Reacher feature synergy)

# M-SAM's Core Workflow and Key Operation



# Modality Contribution in M-SAM

Inspired by cooperative game theory, the Shapley value answers:

*"How much does modality $m$ contribute to the total prediction if it joins all possible combinations of other modalities?"*

- Let $\Phi(x)$ be the model performance using all modalities.

$$V_m(S, \Phi) = \Phi(S \cup \{m\}) - \Phi(S)$$

- Repeat it across all subsets $S$ of modality $m$.
- Shapley value $\Phi_m(x)$ is the weighted average of all $V_m(S, \Phi)$ terms

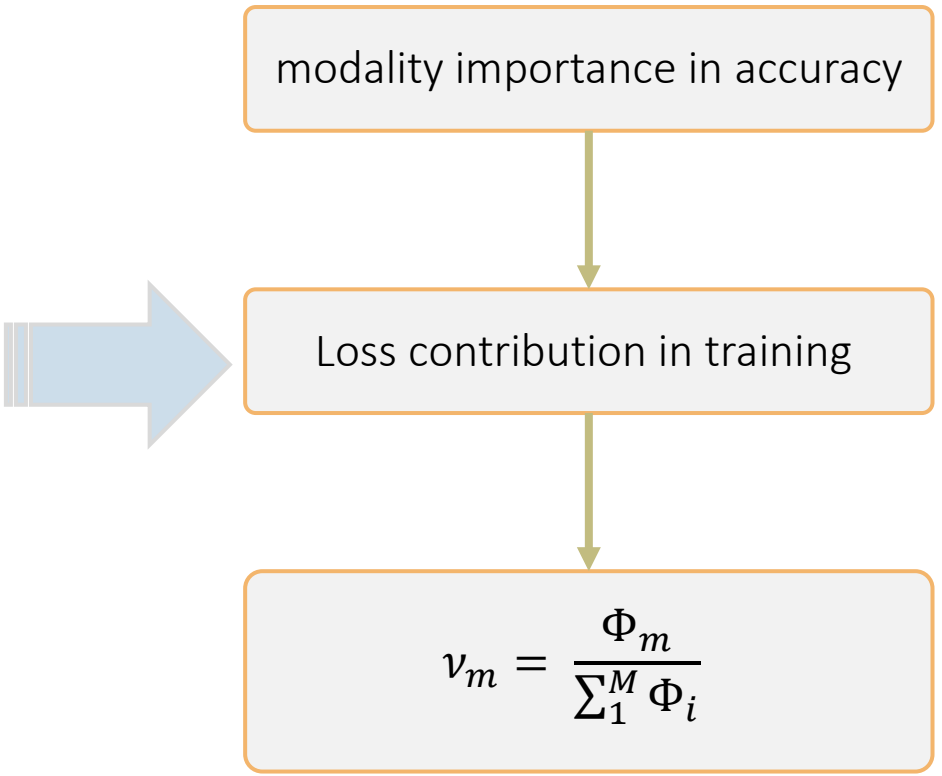
$$\Phi_m(x) = \sum_{S \subseteq \mathcal{M} \setminus \mathrm{m}} \frac{|S!|\,(M-|S|-1)!}{m!} V_m(S, \Phi)$$

- For two modalities:

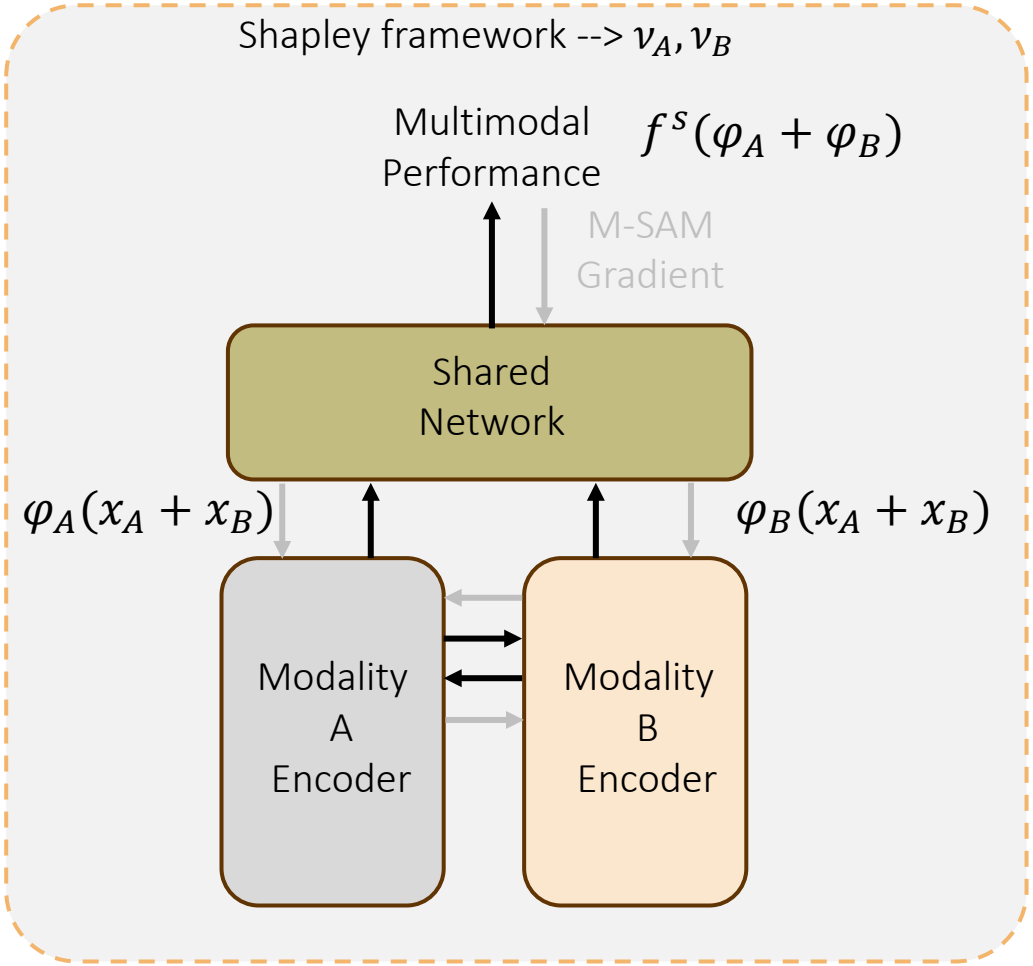
$$\Phi_{m_1}(x) = \frac{1}{2}\left[\Phi(x_1, x_2) - \Phi(0_1, x_2)\right] + \Phi(x_1, 0_2)$$

- For three modalities:

$$\Phi_{m_1}(x) = \frac{1}{3}\left[\Phi(x_1, x_2, x_3) - \Phi(0_1, x_2, x_3)\right] + \frac{1}{6}\left[\Phi(x_1, 0_2, x_3) - \Phi(0_1, 0_2, x_3)\right] + \frac{1}{6}\left[\Phi(x_1, x_2, 0_3) - \Phi(0_1, x_2, 0_3)\right] + \frac{1}{3}\left[\Phi(x_1, 0_2, 0_3) - \Phi(0_1, 0_2, 0_3)\right]$$



# M-SAM's Core Workflow and Key Operation




$$L = \frac{1}{N}\sum_{1}^{N} L\left(f^s(\varphi_A + \varphi_B), y\right)$$

$$\nu_m = \frac{\Phi_m}{\sum_1^M \Phi_i}$$

$$L = (\nu_1 + \nu_2 + \ \ldots + \nu_M)\mathrm{L}$$

# M-SAM's Core Workflow and Key Operation

**Algorithm 1** M-SAM Algorithm

**Require:** Training dataset $\mathcal{D} = \bigcup_{i=1}^{N}\{\bigcup_{m=1}^{M}(x_m^i, y^i)\}\}$, neural network $f(\cdot)$ with parameters $\theta$, loss function $\mathcal{L}$, mini-batch size $b$, learning rate $\eta$, neighborhood size $\rho$, weight decay coefficient$\lambda$,

**Ensure:** Trained parameters $\theta^*$

Initialize parameters $\theta_0$, $t = 0$

**while** not converged **do**

  Sample minibatch $\mathcal{B} = \{((x_1^1, \ldots, x_M^1), y^1), \ldots, ((x_1^b, \ldots, x_M^b), y^b)\}$

  $\mathcal{L}(\theta_t) = \sum_M \mathcal{L}(f(x_m^i; \theta_t), y^i)$ ▷ Eq. 5

  $= \sum_M v_m \mathcal{L}(f(x_1^i, \ldots, x_M^i; \theta_t), y^i)$

  $\mathcal{L}_{\mathrm{d}}(\theta_t) = v_d \mathcal{L}(f(x_1^i, \ldots, x_M^i; \theta_t), y^i) \mid m_{\mathrm{d}}$

  $= \arg \max_{m \in \{1,\ldots,M\}} v_m$

  $\mathcal{L}_{\mathrm{s}}(\theta_t) = \sum_m v_m \mathcal{L}(f(x_1^i, \ldots, x_M^i; \theta_t), y^i)$ , $\forall m \in \mathcal{M} = \{m \in \{1, \ldots, M\}, m \neq m_d\}$

  $\nabla \mathcal{L}_{\mathrm{d}}(\theta_t) = \mathrm{Backward}(\mathcal{L}_{\mathrm{d}}, f(\cdot))$

  $\nabla \mathcal{L}_{\mathrm{s}}(\theta_t) = \mathrm{Backward}(\mathcal{L}_{\mathrm{s}}, f(\cdot))$

  $\epsilon_t^{\mathrm{d}} = \rho \frac{\nabla \mathcal{L}_{\mathrm{d}}(\theta_t)}{\|\nabla \mathcal{L}_{\mathrm{d}}(\theta_t)\|_2}$

  $\theta_t \leftarrow \theta_t - \eta_t \left[\nabla \mathcal{L}_{\mathrm{d}}(\theta_t + \epsilon_t^{\mathrm{d}}) + \nabla \mathcal{L}_{\mathrm{s}}(\theta_t) + \lambda \theta_t\right]$

  $t \leftarrow t + 1$

**end while**

$$\varepsilon = \rho \frac{\overrightarrow{\nabla L}}{|\nabla L|} = \rho \frac{\nabla\big((v_1 + v_2 + \ \ldots + v_M)\big)L}{|\nabla L|}$$

$$\varepsilon = \hat{\rho} \frac{\nabla(v_1 L)}{|\nabla(v_1 L)|} + \hat{\rho} \frac{\nabla(v_2 L)}{|\nabla(v_2 L)|} + \ldots + \hat{\rho} \frac{\nabla(v_M L)}{|\nabla(v_M L)|} = \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_M$$

$$\min_{\theta} \max_{\|\varepsilon\| \leq \rho} L(\theta + \varepsilon) + \frac{\lambda}{2}\|\theta\|^2$$

$$\min_{\theta} \max_{\|\varepsilon\| \leq \rho} [L_1(\theta + \varepsilon) + \ \ldots + L_M(\theta + \varepsilon)] + \frac{\lambda}{2}\|\theta\|^2$$

$$\min_{\theta} \max_{\|\varepsilon\| \leq \rho} \left[L_{m_d}\left(\theta + \varepsilon_{m_d}\right) - L_{m_d}(\theta)\right] + L_{m_d}(\theta)$$

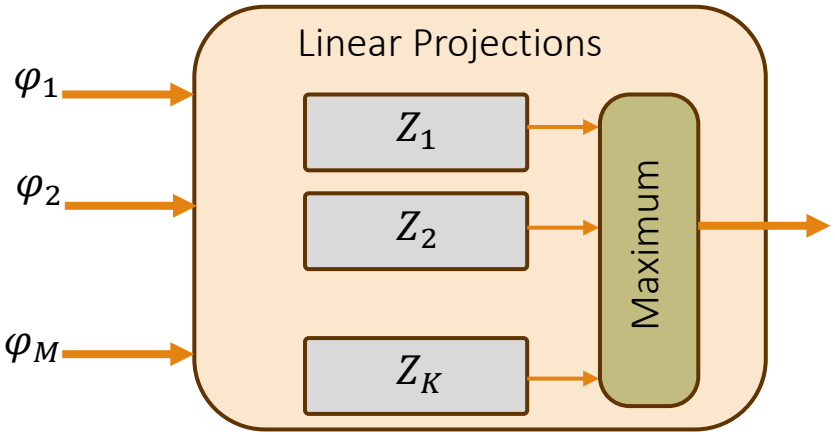
$$+[L_1(\theta) + L_2(\theta) + \cdots + L_M(\theta)] + \frac{\lambda}{2}\|\theta\|^2$$

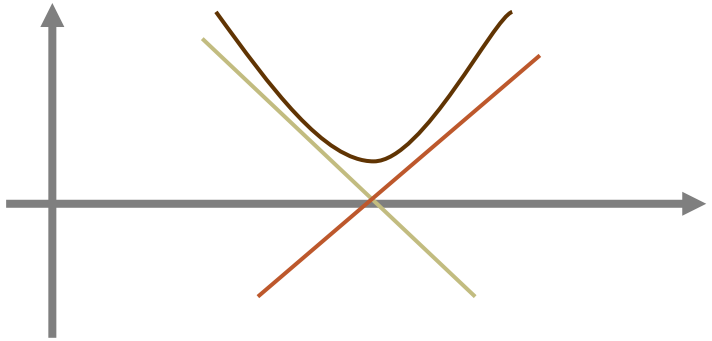
# Experimental Setting

| Field of Research | Size | Dataset | Modality | Samples | content |
|---|---|---|---|---|---|
| Affective Computing | L | UR-Funny(Hasan et al., 2019) | {a, v, t} | 16,514 | humor |
| | M | CREMA-D (Cao et al., 2014) | {a, v} | 7,442 | emotion |
| Multimedia | M | AV-MNIST(Vielzeuf et al., 2018) | {a, v} | 70,000 | digit |
| | S | AVE(Tian et al., 2018) | {a, v} | 4,143 | event detection |

**Fusion Strategy:**

- *Early fusion:* MAXOUT
- *Late fusion:* Independent modality encoders, fused at the decision level



**CREMA-D:**

- Audio: ReasNet18 (spectrogram)
- Video: 1 frame per minute

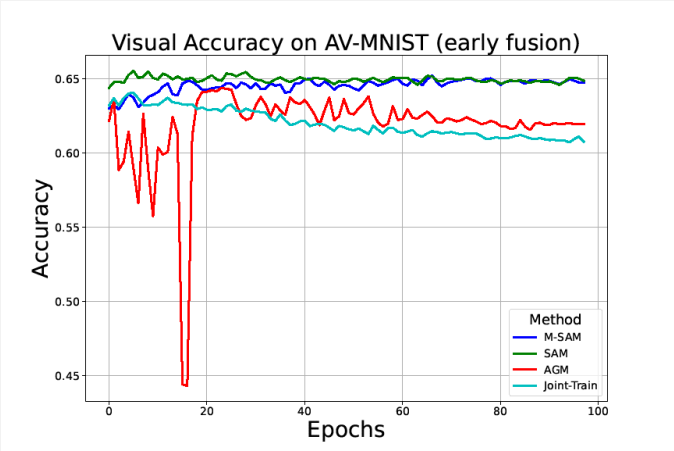
**AV-MNIST:**

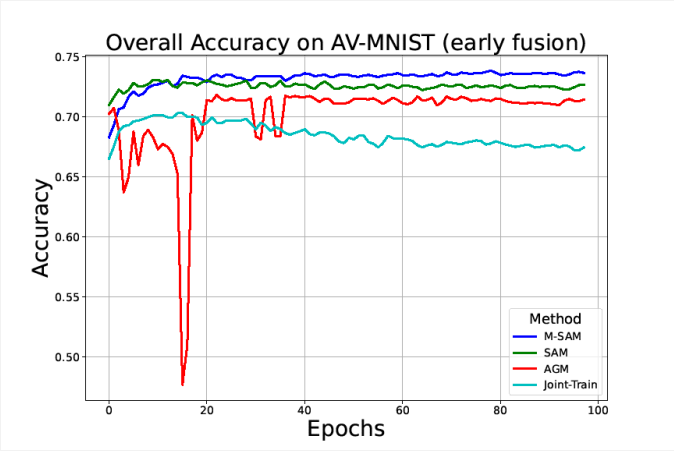
- Audio: ReasNet18 (spectrogram)
- Video: Image

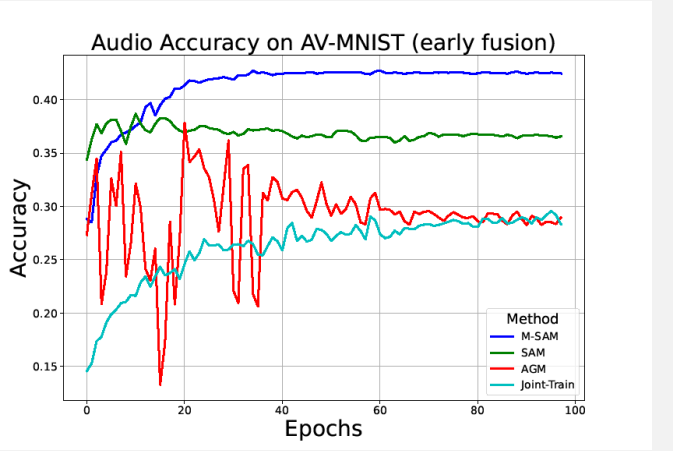
**UR-Funny:**

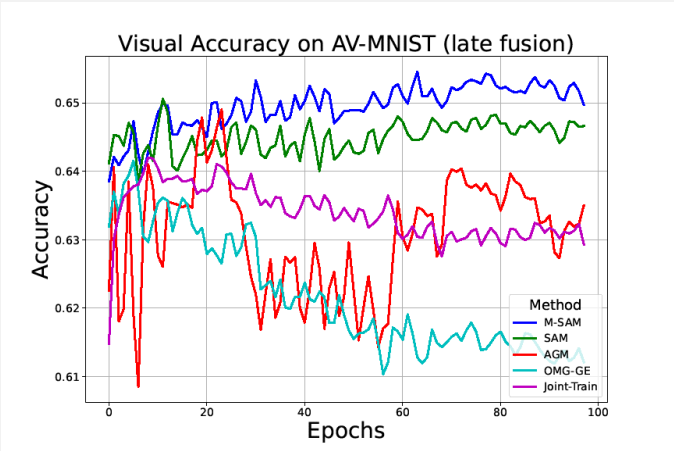
- Audio, Frame, Text: Transformer

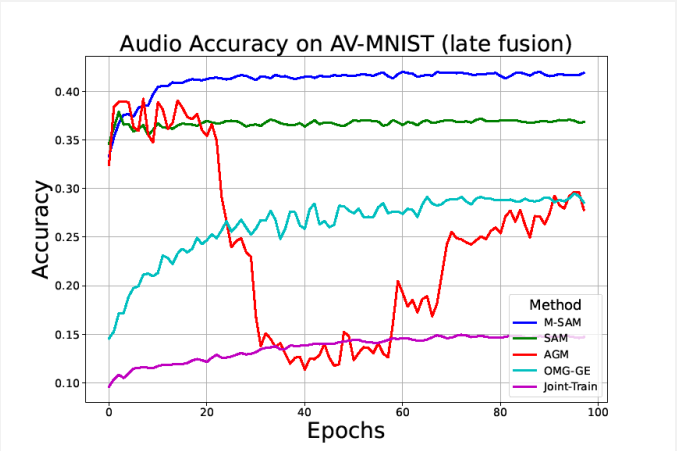
# Experimental Results (AV-MNIST curves)

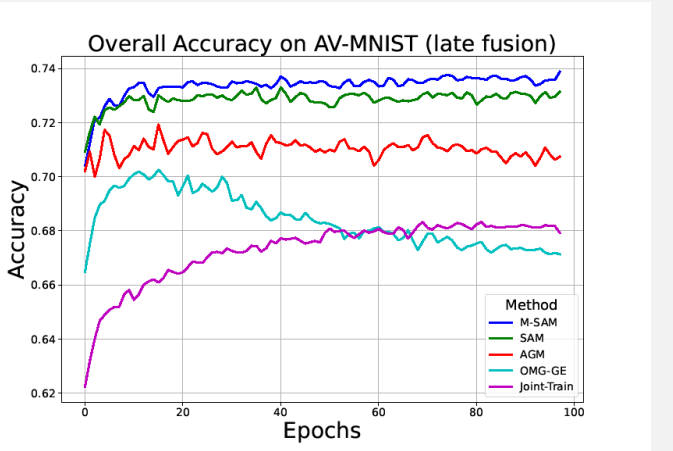


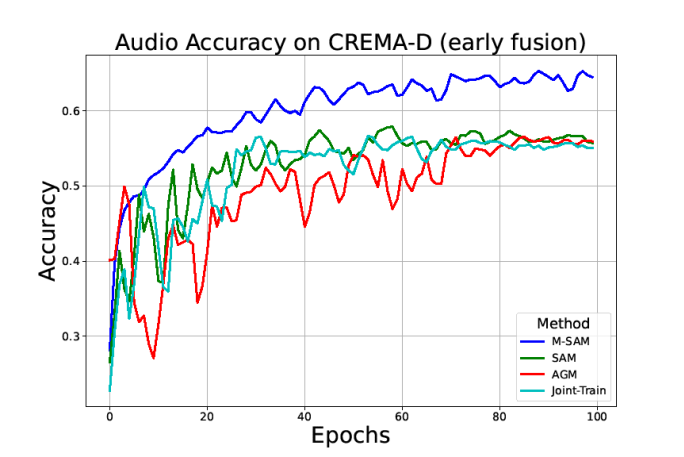



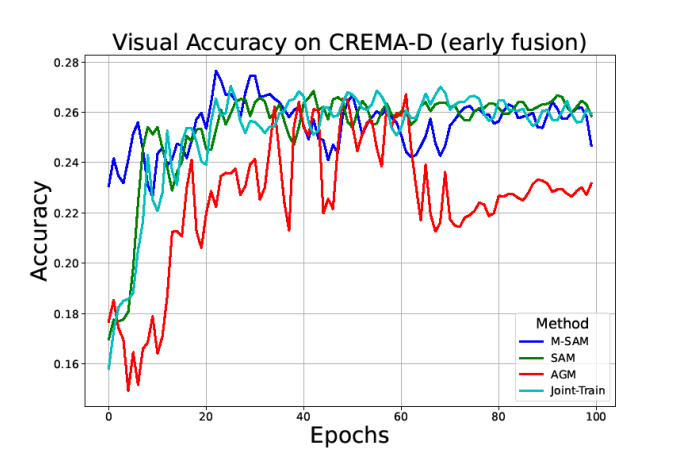



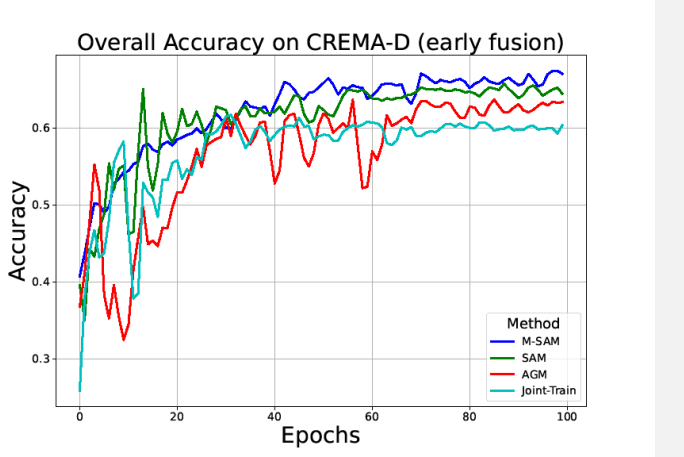



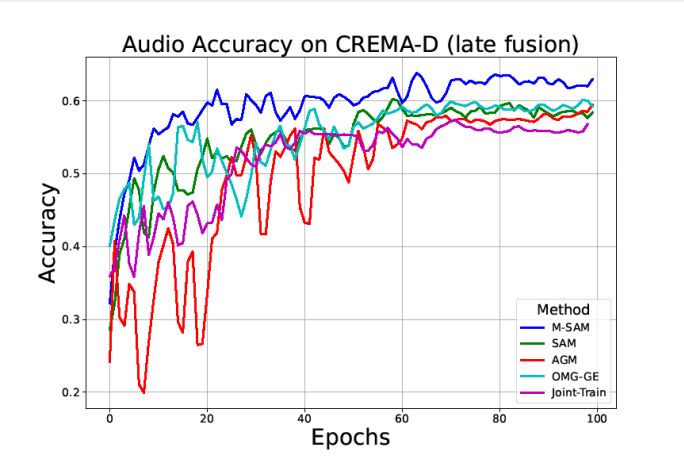



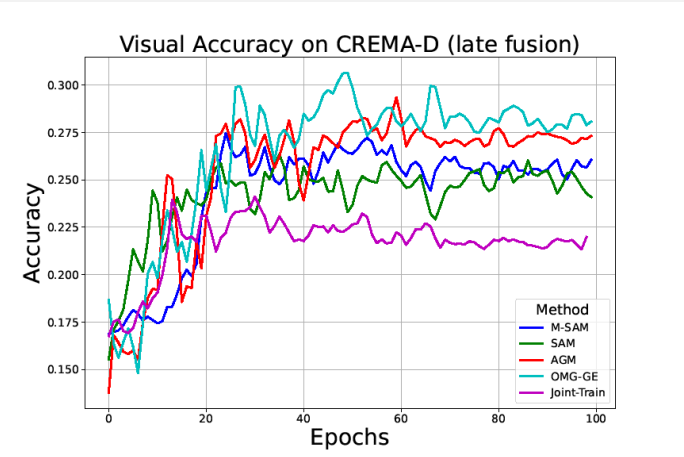
# Experimental Results (CREMA-D curves)

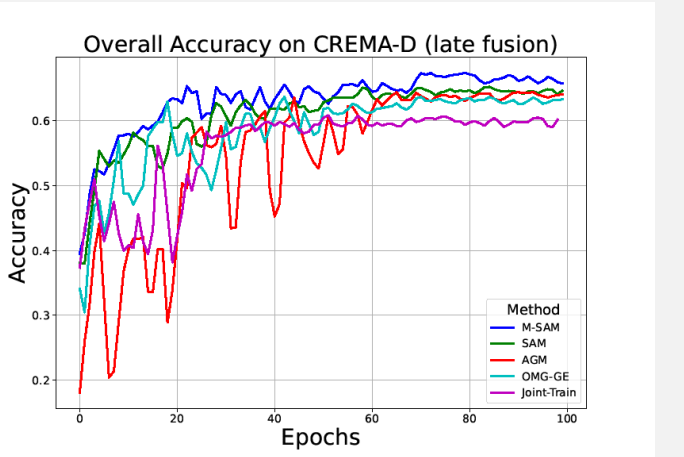


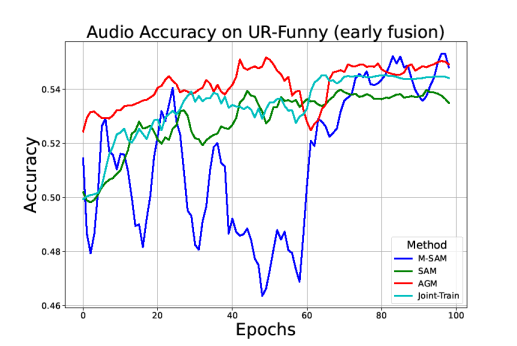



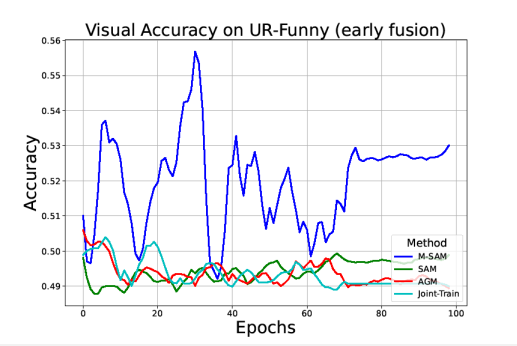



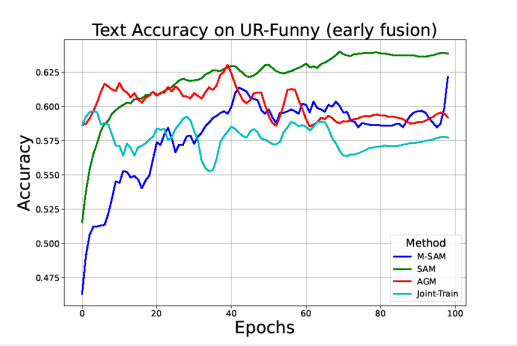



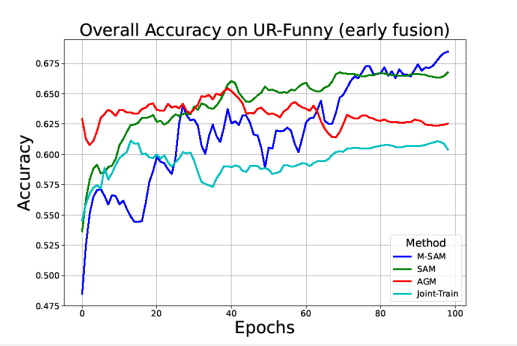



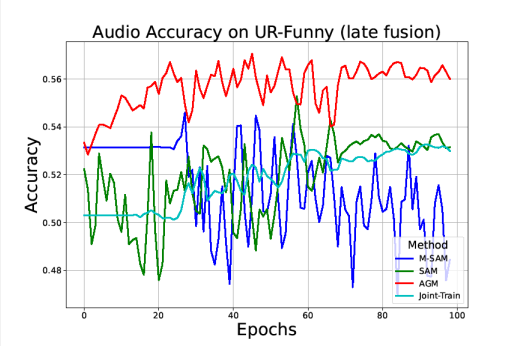
# Experimental Results (UR-Funny curves)

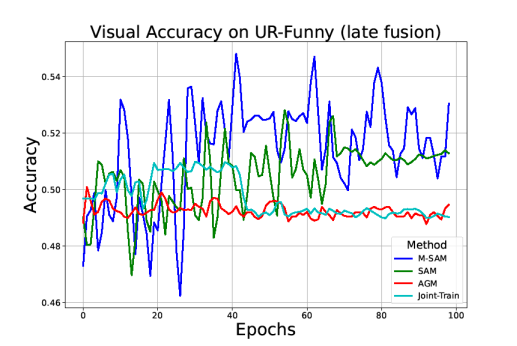


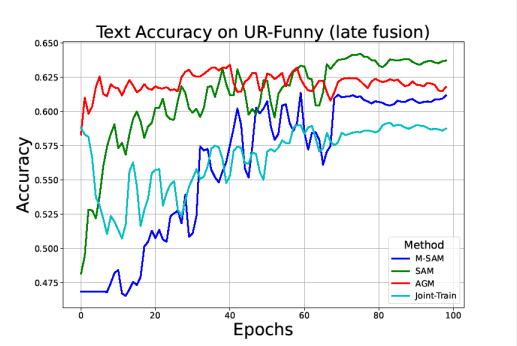



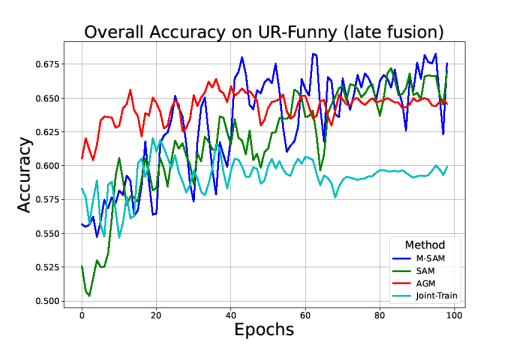



# Experimental Results (Late Fusion)

Table 1: Accuracy ($Acc.$), and single-modal accuracy ($Acc_a, Acc_v, Acc_t$) on the AV-MNIST, CREMA-D, UR-Funny, and AVE datasets using ***late fusion*** architecture. Please note that the OGM-GE method could not extend to more than two modality cases in their original shape.

| Model | AV-MNIST [33] | | | CREMA-D [2] | | | UR-Funny [11] | | | | AVE [31] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_{mm}}$ | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_{mm}}$ | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_t}$ | $\mathbf{Acc_{mm}}$ | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_{mm}}$ |
| Single-audio | 39.61 | .. | .. | 52.12 | .. | .. | 59.23 | .. | .. | .. | 65.43 | .. | .. |
| Single video | .. | 65.14 | .. | .. | 60.37 | .. | .. | 53.16 | .. | .. | .. | 64.58 | .. |
| Single-text | .. | .. | .. | .. | .. | .. | .. | .. | 63.46 | .. | .. | .. | .. |
| Joint-Train | 14.59 | 62.85 | 68.41 | 56.08 | 44.32 | 61.19 | 50.31 | 53.51 | 49.78 | 64.50 | 59.10 | 63.72 | 73.19 |
| MSES [8] | 27.50 | 63.34 | 70.68 | 55.31 | 45.72 | 64.13 | 55.31 | 49.69 | 57.87 | 64.23 | 65.84 | 71.93 | 76.47 |
| MSLR [38] | 22.72 | 62.92 | 70.62 | 55.75 | 47.84 | 62.93 | 53.14 | **53.59** | 46.93 | 65.52 | 70.39 | 69.41 | 75.22 |
| OGM-GE [28] | 24.53 | 55.85 | 71.08 | 58.15 | **58.90** | 64.42 | .. | .. | .. | .. | 67.91 | 71.09 | 75.53 |
| AGM [23] | 38.90 | 63.65 | 72.14 | 56.35 | 54.12 | 64.72 | 54.87 | 49.36 | 62.22 | 65.97 | 70.68 | 72.34 | 77.11 |
| MM-Pareto [35] | 42.17 | 64.31 | 73.22 | 61.85 | 56.94 | 66.63 | 54.59 | 52.12 | 62.37 | 67.04 | 70.31 | **73.88** | 77.68 |
| CGGM [10] | 39.53 | 64.13 | 73.42 | 57.14 | 55.07 | 67.03 | **55.21** | 49.83 | 62.74 | 67.43 | 70.91 | 73.17 | 77.83 |
| Recon-Boost [13] | 40.12 | 64.18 | 73.59 | 57.08 | 54.83 | 67.47 | 55.13 | 50.08 | **62.88** | 67.61 | **71.13** | 73.48 | 78.02 |
| SAM | 36.31 | 64.67 | 73.17 | 60.69 | 51.43 | 66.32 | 53.67 | 51.37 | 61.48 | 65.95 | 66.13 | 72.83 | 77.66 |
| M-SAM | **41.93** | **64.97** | **74.08** | **62.78** | 53.22 | **68.56** | 51.56 | 52.67 | 60.17 | **68.31** | 68.27 | 72.57 | **79.67** |

# Experimental Results (Early Fusion)

Table 2: Accuracy ($Acc.$), and single-modal accuracy ($Acc_a, Acc_v, Acc_t$) on the AV-MNIST, CREMA-D, UR-Funny, and AVE datasets using ***early fusion*** architecture.

| Model | AV-MNIST [33] | | | CREMA-D [2] | | | UR-Funny [11] | | | | AVE [31] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_{mm}}$ | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_{mm}}$ | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_t}$ | $\mathbf{Acc_{mm}}$ | $\mathbf{Acc_a}$ | $\mathbf{Acc_v}$ | $\mathbf{Acc_{mm}}$ |
| Joint-Train | 24.28 | 60.14 | 71.15 | 55.31 | 51.72 | 62.13 | 54.87 | 50.86 | 54.14 | 65.15 | 67.40 | 71.85 | 76.29 |
| AGM [23] | **47.79** | **68.48** | 72.26 | 51.42 | 47.54 | 64.09 | **64.88** | **55.20** | 63.36 | 66.07 | 68.85 | 72.46 | 77.08 |
| MM-Pareto [35] | 39.82 | 66.15 | 72.74 | 56.90 | 52.83 | 65.30 | 58.32 | 53.08 | 60.45 | 65.92 | 68.52 | 72.18 | 76.83 |
| CGGM [10] | 41.30 | 67.27 | **73.11** | **57.84** | 53.67 | 66.90 | 60.42 | 53.40 | 62.53 | 66.98 | 69.01 | **72.81** | 77.32 |
| Recon-Boost [13] | 40.94 | 67.01 | 72.97 | 57.56 | **54.01** | 66.74 | 59.88 | 53.26 | 62.18 | 66.83 | 68.91 | 72.63 | 77.18 |
| SAM | 38.56 | 64.81 | 73.22 | 56.35 | 53.52 | 66.22 | 54.08 | 49.77 | 63.86 | 66.87 | 68.37 | 72.02 | 76.76 |
| M-SAM (Ours) | 45.63 | 67.72 | **74.48** | 56.83 | 53.71 | **68.43** | 63.20 | 54.77 | **65.24** | **67.92** | **70.11** | 72.46 | **78.23** |

# Quantifying Generalization of Models

To rank models by their real-world reliability, we need a metric that reflects how well they generalize, not just memorize.

***Hessian eigenvalues*** captures curvature but requires second-order derivatives.

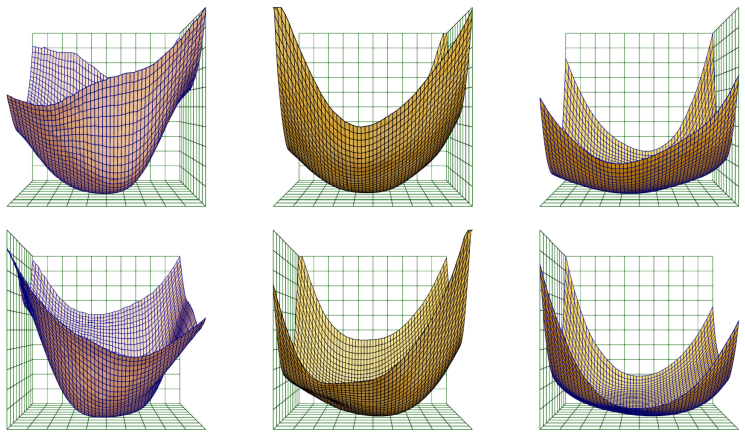
***Sharpness Metric***

$$Q_{x,f}(\varepsilon, A) = \frac{max_{y \in C_\varepsilon} f(x + Ay) - f(x)}{1 + f(x)} \times 100$$

***Normalized Generalization Gap***

- $\frac{Acc_{train} - Acc_{test}}{Acc_{train}} \times 100$
- Cheap: No extra computation
- Larger Gap --> Sharper minima --> less generalization

# Quantifying Generalization of Models



(a) AGM (b) SAM (c) M-SAM

Figure 1: Loss landscape visualization of CREMA-D (late fusion) for AGM, SAM, and M-SAM from two different viewpoints.