COUNTERFACTUAL SENSITIVITY-GUIDED INTERVENTION FOR MULTI-MODAL DOMIAN GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal domain generalization (MMDG) aims to learn models from multiple source domains that generalize to unseen target domains without access to target data during training. A key challenge in this setting is that the predictive benefits of individual modalities may not persist across domains. Our analysis shows that a modality that supports fused predictions in source domains can become less useful or even detrimental in a target domain, highlighting the need to draw on alternative evidence when modality contributions change. To develop this capability using source data alone, we propose Counterfactual Sensitivity-Guided Intervention (CSI). CSI uses online, sample-specific logit gaps between complete and modality-removed inputs to identify modalities on which the current prediction strongly depends. Preferentially removing these modalities creates training views that encourage prediction from the remaining evidence. An annealed objective progressively increases the contribution of these views to training, while an exponential moving average teacher guides prediction from reduced evidence using stable distillation targets derived from complete inputs. Experiments on two public MMDG benchmarks demonstrate improved unseen-domain generalization and robustness to modality-specific perturbations, with additional loss-landscape analyses indicating flatter minima.

1 INTRODUCTION

Multimodal learning (Radford et al., 2021; Tsai et al., 2019) integrates complementary information from heterogeneous modalities to improve predictive performance. However, distribution shifts (Zhou et al., 2022; Wang et al., 2026b) arising from environmental changes and modality-specific variations can substantially impair its generalization to new domains. Multimodal domain generalization (Dong et al., 2026) (MMDG) addresses this challenge by learning from multiple multimodal source domains to generalize to unseen target domains without accessing target data during training. In this setting, distribution shifts can affect both the evidence available within individual modalities and how that evidence contributes to the fused prediction.

We examine this problem by analyzing how the contribution of individual modalities to fused predictions changes across domains. For a fixed source-trained fusion model, we compare predictions from complete inputs with those obtained after removing individual modality features. We define a modality's net contribution as complete-input accuracy minus accuracy after its removal. In the EPIC-Kitchens $\mathrm{D2} + \mathrm{D3} \rightarrow \mathrm{D1}$ setting, Fig. 1(a) shows that the class-averaged contribution of audio decreases from 3.27 percentage points in the source domains to 0.14 in the target domain. More critically, its contribution can reverse within the same action class: for Class 1, audio improves fused accuracy by 8.45 percentage points in the source domains, yet reduces it by 5.95 percentage points in the target domain. This reversal exposes a challenge for generalization: a modality that benefits source-domain predictions can become detrimental to target-domain predictions. These observations motivate models that can draw on alternative evidence when the predictive benefit of a relied-upon modality diminishes or reverses.

We therefore introduce source-domain training conditions that change the evidence available to the model. Modality removal provides such a condition by explicitly withholding one modality's feature input while preserving the remaining features, encouraging the model to predict from alternative evidence. To examine whether this training strategy improves generalization, we compare modality

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.

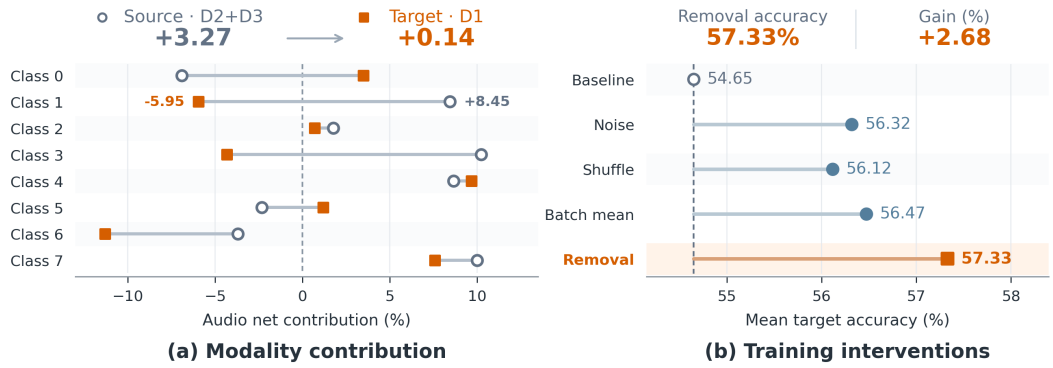


Figure 1: Cross-domain modality contributions and training interventions on EPIC-Kitchens (D2 + D3 → D1). (a) Audio net contributions of a fixed fusion model on source and target domains; the top values summarize class averages. (b) Mean accuracy on complete, unperturbed target inputs after different training interventions. Dashed lines indicate zero contribution in (a) and baseline accuracy in (b).

removal with feature noise, feature shuffling, and batch-mean replacement, alongside a baseline without intervention. All interventions follow the same sample and modality selection procedure, with other training settings held fixed. We evaluate all models on complete, unperturbed target-domain inputs to assess their generalization under natural domain shifts. As shown in Fig. 1(b), modality removal achieves the highest mean target accuracy among the evaluated interventions, reaching 57.33% compared with 54.65% for the baseline. This comparison supports our choice of modality removal as a source-domain training intervention.

While modality removal creates opportunities to learn from alternative evidence, removing different modalities can affect the current prediction to different degrees. For a given sample, removing one modality may substantially alter the fused prediction, whereas removing another may leave it largely unchanged. This motivates selecting interventions according to their impact on the model's current prediction. A larger prediction gap indicates stronger predictive dependence on the removed modality. Prioritizing these removals exposes the model to training conditions in which influential evidence is withheld, encouraging it to use the remaining modalities for prediction. This selection criterion reflects the model's current dependence on multimodal evidence without assuming which modality will become unreliable in an unseen target domain.

Based on this reasoning, we propose Counterfactual Sensitivity-Guided Intervention (CSI) for MMDG. During source-domain training, CSI computes removal-induced logit gaps online and converts them into a sensitivity distribution for each sample, preferentially selecting modalities with larger gaps for removal. The interventions thus adapt to the model's evolving dependence on multimodal evidence. To progressively develop the ability to predict from remaining evidence, CSI employs an annealed objective that gradually increases the contribution of masked-view learning while retaining supervision on complete inputs. An exponential moving average teacher further guides this process by providing predictions from complete inputs as distillation targets. This supervision encourages the student to approximate predictions informed by all modalities using the remaining evidence, while parameter averaging stabilizes the distillation targets throughout training. We evaluate CSI on EPIC-Kitchens and HAC, examining generalization to unseen domains and robustness under modality-specific perturbations. Overall, our main contributions are as follows:

- We analyze cross-domain changes in modality contributions to fused predictions. Source-domain predictive benefits can diminish or reverse in target domains, highlighting the need to learn alternative evidence use under domain shifts.
- We propose CSI, which combines online sensitivity-guided modality removal with an annealed objective and teacher guidance. By preferentially removing influential modalities for each sample, CSI encourages prediction from the remaining evidence.
- Comprehensive experiments on two public MMDG benchmarks demonstrate that CSI achieves flatter minima, improved unseen-domain generalization and greater fusion robustness.

2 Related Work

2.1 Domain Generalization

Domain generalization (DG) aims to learn models that generalize to unseen domains (Gulrajani & Lopez-Paz, 2020). Existing research explores several major directions, including domain alignment, data augmentation, representation learning, and optimization-based methods. Domain alignment methods (Li et al., 2018b; 2025a; Ballas & Diou, 2025; Han et al., 2026) reduce distribution discrepancies to learn invariant features. Data augmentation techniques (Zhang et al., 2017; Volpi et al., 2018; Qiao et al., 2020; Cho et al., 2025; Chen et al., 2026) introduce synthetic variations to approximate potential domain shifts. Representation learning approaches (Lin et al., 2021; Qu et al., 2023; Xu et al., 2025; Wen et al., 2025; Zhao et al., 2026) disentangle domain-invariant and domain-specific components or encourage stable semantic structures. Optimization-based approaches seek flatter solutions associated with improved robustness (Izmailov et al., 2018; Foret et al., 2020; Li et al., 2025a; Peng et al., 2026), using strategies such as sharpness-aware minimization (SAM) (Foret et al., 2020; Wang et al., 2023; Zhou et al., 2024) and weight averaging (Izmailov et al., 2018; Cha et al., 2021; Arpit et al., 2022; Morales-Brotons et al., 2024; Li et al., 2024; Ajroldi et al., 2025; Yang et al., 2026). Building on this body of work, we investigate DG in multimodal settings, where domain shifts can alter how individual modalities contribute to fused predictions.

2.2 Multi-Modal Domain Generalization.

Compared with the extensive progress in uni-modal DG, research on MMDG remains relatively limited. Existing works explore this problem from various perspectives, including cross-modal alignment (Planamente et al., 2022), representation decomposition (Dong et al., 2023), self-supervised learning (Dong et al., 2024), cross-modal consistency and knowledge transfer (Fan et al., 2024), unified representation learning (Huang et al., 2025), and adversarial alignment with dynamic weighting (Li et al., 2025b). Other approaches address modality imbalance through adaptive dropout and collaborative distillation (Wang et al., 2026a), or mitigate fusion overfitting through feature entropy regularization (Yarici & AlRegib, 2026). Beyond the standard supervised setting, MMDG has also been extended to semi-supervised learning by leveraging unlabeled data through pseudo-labeling and cross-modal alignment (Li et al., 2026). Our work examines cross-domain changes in modality contributions and uses sample-specific responses to modality removal to guide training interventions, encouraging prediction from the remaining evidence.

3 Methodology

3.1 Problem Definition

We consider the standard MMDG setting (Dong et al., 2023). Let $\mathcal{S} = \{\mathcal{D}_j\}_{j=1}^{K}$ denote K source domains, where $\mathcal{D}_j = \{(x_i^{(j)}, y_i^{(j)})\}_{i=1}^{N_j}$ contains labeled samples drawn from a domain-specific distribution P_{XY}^{j}. Each input $x = (x_1, \ldots, x_M)$ comprises $M \geq 2$ modalities, and $y \in \{1, \ldots, C\}$ denotes its class label. The objective is to learn from $\mathcal{S}$ a model that generalizes to unseen target domains without accessing their data during training. For an unseen target distribution $P_{\mathcal{T}}$, the desired risk is

$$\mathcal{R}_{\mathcal{T}}(f) = \mathbb{E}_{(x,y)\sim P_{\mathcal{T}}}\big[\ell(f(x), y)\big], \tag{1}$$

where ℓ is the cross-entropy loss. Each modality is processed by an encoder h_m, including the feature normalization used before fusion:

$$z_m = h_m(x_m) \in \mathbb{R}^{d_m}, \qquad z = [z_1, \ldots, z_M]. \tag{2}$$

A fusion classifier g maps the concatenated features to class logits:

$$f_\theta(x) = g(z) \in \mathbb{R}^C, \tag{3}$$

where θ denotes the encoder and fusion-classifier parameters. We omit domain and sample indices below unless required.

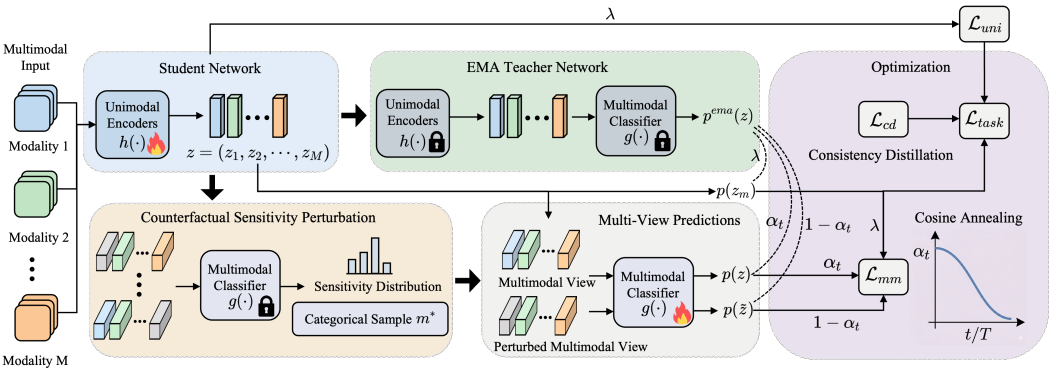


Figure 2: Overview of CSI. Counterfactual sensitivity guides sample-specific modality removal, while an EMA teacher provides complete-input predictions as distillation targets.

3.2 Counterfactual Sensitivity-Guided Intervention

We propose Counterfactual Sensitivity-Guided Intervention (CSI) to prepare models for cross-domain changes in modality contributions. CSI constructs training views in which one modality is removed while the remaining features are preserved. For each sample, it first measures the response of the current fused prediction to individual modality removal, then uses these responses to select a modality for removal. An annealed objective combines complete-input and masked-input supervision, while an exponential moving average (EMA) teacher supplies complete-input predictions to guide learning from the remaining evidence. Figure 2 illustrates the overall framework.

3.2.1 Counterfactual Sensitivity Estimation

Removing different modalities can affect the current fused prediction to different degrees. To characterize this dependence for each sample, we compare complete-input logits with those obtained after removing one modality's features. For modality m, the counterfactual representation is

$$z_{\setminus m} = [z_1, \ldots, z_{m-1}, \mathbf{0}, z_{m+1}, \ldots, z_M], \tag{4}$$

where the zero vector has dimension d_m and all other modality features remain unchanged. We define counterfactual sensitivity as the mean squared difference between the two logit vectors:

$$\Delta_m(z) = \frac{1}{C} \left\| g(z) - g(z_{\setminus m}) \right\|_2^2. \tag{5}$$

A larger gap identifies a removal that more strongly affects the current fused logits. This response measures current predictive dependence without assuming which modality will become unreliable in a target domain. We convert the responses into a sensitivity distribution over modalities:

$$s_m(z) = \frac{\exp\left(\Delta_m(z)/\tau\right)}{\sum_{j=1}^{M} \exp\left(\Delta_j(z)/\tau\right)}, \tag{6}$$

where $\tau > 0$ controls the concentration of the distribution. Sensitivities are recomputed online using the current student and serve only as intervention-selection signals; gradients are not propagated through their computation. The same encoded features are reused for all removal evaluations.

3.2.2 Sensitivity-Guided Modality Removal

The sensitivity distribution determines which evidence is withheld during training. For each sample, we draw one modality according to

$$m^* \sim \text{Categorical}\left(s_1(z), \ldots, s_M(z)\right). \tag{7}$$

We then construct the masked representation as

$$r_m = \begin{cases} 0, & m = m^*, \\ 1, & \text{otherwise}, \end{cases} \qquad \tilde{z} = [r_1 z_1, \ldots, r_M z_M]. \tag{8}$$

When activated, this intervention removes one modality's feature input while preserving the others. Modalities with larger removal responses are sampled more frequently, so training more often withholds evidence that strongly influences the current prediction. Recomputing these responses online allows modality selection to adapt to the student's changing dependence on multimodal evidence. The resulting training views encourage prediction from the remaining modalities, helping the model use alternative evidence when modality contributions change across domains.

3.2.3 Learning from Remaining Evidence

Annealed supervision. Complete-input supervision supports the development of task-relevant modality representations during early training. As training progresses, we gradually increase the contribution of masked-input learning, encouraging the model to predict from the remaining evidence. We combine these two forms of supervision with auxiliary unimodal supervision to support feature learning within each modality, defining the corresponding losses as

$$\mathcal{L}_{\text{full}} = \ell(g(z), y), \quad \mathcal{L}_{\text{mask}} = \ell(g(\tilde{z}), y), \quad \mathcal{L}_{\text{uni}} = \sum_{m=1}^{M} \ell(g_m(z_m), y), \tag{9}$$

where g_m is a modality-specific classifier. The supervised objective is

$$\mathcal{L}_{\text{task}} = \alpha_t \mathcal{L}_{\text{full}} + (1 - \alpha_t)\mathcal{L}_{\text{mask}} + \lambda \mathcal{L}_{\text{uni}}. \tag{10}$$

Here, $\lambda \geq 0$ controls auxiliary supervision. Over $T > 1$ training epochs, we decrease α_t using cosine annealing:

$$\alpha_t = \alpha_{\text{end}} + \frac{\alpha_{\text{start}} - \alpha_{\text{end}}}{2} \left[1 + \cos\left(\frac{\pi(t-1)}{T-1}\right)\right], \tag{11}$$

where $t \in \{1, \ldots, T\}$ and $0 \leq \alpha_{\text{end}} \leq \alpha_{\text{start}} \leq 1$. As α_t decreases, the objective progressively shifts training emphasis from complete-input supervision toward prediction from the remaining evidence.

Complete-input teacher guidance. Hard labels do not encode the sample-specific class relations captured by complete-input predictions.We therefore complement label supervision with soft predictions from a teacher observing complete inputs. To stabilize the targets as the student and its interventions evolve, we maintain an EMA teacher whose encoder and fusion-classifier parameters $\bar{\theta}$ are updated as

$$\bar{\theta}_k = \beta \bar{\theta}_{k-1} + (1 - \beta)\theta_k, \tag{12}$$

where k indexes optimization steps, $\beta \in [0, 1)$ is the momentum coefficient, and $\bar{\theta}_0 = \theta_0$. The teacher processes complete inputs using its own encoders and fusion classifier, producing the target distribution

$$q(x) = \text{sg}\left[\text{softmax}\left(f_{\bar{\theta}}(x)\right)\right], \tag{13}$$

where sg denotes the stop-gradient operator. The student distributions are defined as

$$p_{\text{full}} = \text{softmax}(g(z)), \quad p_{\text{mask}} = \text{softmax}(g(\tilde{z})), \quad p_m = \text{softmax}(g_m(z_m)). \tag{14}$$

The teacher supervises complete-input, masked-input, and unimodal predictions through consistency distillation:

$$\begin{aligned} \mathcal{L}_{\text{cd}} = \lambda \sum_{m=1}^{M} & D_{\text{KL}}(q(x) \| p_m) \\ & + \Big[\alpha_t D_{\text{KL}}(q(x) \| p_{\text{full}}) + (1 - \alpha_t) D_{\text{KL}}(q(x) \| p_{\text{mask}})\Big], \end{aligned} \tag{15}$$

where D_{KL} is the Kullback–Leibler divergence (Kullback & Leibler, 1951). This supervision encourages the student to approximate predictions informed by all modalities using the remaining evidence, while parameter averaging stabilizes the distillation targets throughout training. The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{cd}}. \tag{16}$$

We optimize its mini-batch average over source-domain samples and their sampled masked views. At inference, the EMA model predicts from complete inputs without removal interventions, with the checkpoint selected using source-domain validation data.

Method	Modality			EPIC-Kitchens				HAC			
	Video	Audio	Flow	D2,D3→D1	D1,D3→D2	D1,D2→D3	Avg.	A,C→H	H,C→A	H,A→C	Avg.
ERM	✓	✓		52.04±0.60	60.73±1.69	57.92±1.40	56.90	72.96±0.69	74.10±3.65	51.96±1.73	66.34
RNA-Net	✓	✓		52.21±1.20	60.06±2.06	56.17±1.87	56.15	73.42±0.99	73.33±2.69	50.06±1.16	65.60
MDJA	✓	✓		54.55±0.50	60.55±0.94	56.82±0.84	57.31	74.21±2.16	77.37±1.42	52.05±0.94	67.88
SimMMDG	✓	✓		56.35±2.19	64.52±0.72	58.21±1.94	59.69	76.54±2.40	75.79±2.14	48.71±1.61	67.01
MOOSA	✓	✓		54.64±3.73	64.93±1.36	61.60±0.51	60.39	74.29±1.40	75.90±0.55	52.79±2.49	67.66
CMRF	✓	✓		56.77±0.30	65.21±0.31	61.79±0.33	61.26	79.21±0.69	79.07±0.82	**55.48±1.41**	**71.25**
MBCD	✓	✓		58.06±0.50	68.10±0.35	63.31±0.21	63.16	80.08±0.44	78.81±0.72	53.09±1.84	70.66
CSI	✓	✓		**60.24±0.45**	**69.05±0.09**	**64.61±0.79**	**64.63**	**81.58±0.90**	**79.25±1.10**	52.48±2.21	71.10
ERM	✓		✓	56.88±0.26	65.74±1.42	58.02±1.32	60.21	74.89±1.12	72.33±1.51	43.35±6.35	63.52
RNA-Net	✓		✓	59.29±0.54	66.37±1.30	57.85±2.02	61.17	77.19±2.40	74.58±2.84	43.11±4.33	64.96
MDJA	✓		✓	58.58±1.06	66.76±0.28	57.28±1.35	60.87	77.72±1.18	72.41±1.88	44.00±3.17	64.71
SimMMDG	✓		✓	57.99±2.29	67.59±0.42	55.85±1.88	60.48	77.48±1.47	73.99±1.76	51.78±6.05	67.75
MOOSA	✓		✓	59.16±1.36	66.44±0.98	61.43±1.14	62.34	75.25±2.25	75.68±0.51	51.75±2.66	67.56
CMRF	✓		✓	63.41±0.35	69.48±0.80	61.52±0.53	64.80	80.32±0.68	77.08±1.52	51.87±3.34	69.76
MBCD	✓		✓	**64.16±0.33**	72.19±0.59	63.05±0.53	66.47	80.37±0.74	78.07±0.75	51.93±1.31	70.12
CSI	✓		✓	63.46±0.31	**72.68±0.39**	**63.75±0.55**	**66.63**	**82.59±0.54**	**79.42±1.60**	**52.39±2.76**	**71.47**
ERM		✓	✓	53.56±1.53	59.14±1.54	56.71±0.94	56.47	54.17±2.28	60.56±3.26	42.19±2.70	52.31
RNA-Net		✓	✓	51.50±1.26	60.35±1.49	56.86±1.38	56.24	54.91±1.07	59.82±2.72	43.23±2.13	52.65
MDJA		✓	✓	53.44±3.74	59.93±3.14	54.79±2.74	56.05	53.57±1.85	62.69±2.17	40.78±0.73	52.35
SimMMDG		✓	✓	57.30±0.89	65.39±1.01	56.19±1.31	59.63	55.61±1.70	63.43±1.57	45.31±0.20	54.78
MOOSA		✓	✓	53.72±1.41	67.20±1.54	60.78±2.03	60.57	56.79±1.28	63.58±0.48	42.59±1.76	54.32
CMRF		✓	✓	59.13±0.87	65.08±0.99	61.30±0.28	61.84	60.49±0.46	64.75±0.32	47.98±1.28	57.74
MBCD		✓	✓	59.25±0.45	68.05±0.59	61.95±0.27	63.08	**63.76±0.91**	65.64±0.69	**49.05±0.19**	**59.48**
CSI		✓	✓	**61.19±1.07**	**69.82±0.31**	**62.55±0.50**	**64.52**	62.96±1.59	**65.78±0.30**	47.65±0.87	58.80
ERM	✓	✓	✓	56.90±1.44	65.61±1.34	60.10±2.34	60.87	74.67±3.33	68.54±0.41	45.40±3.58	62.87
RNA-Net	✓	✓	✓	55.76±2.07	64.05±1.57	60.35±1.22	60.05	76.02±3.14	73.11±1.50	46.42±3.19	65.18
MDJA	✓	✓	✓	59.17±0.92	67.03±1.13	59.76±1.39	61.99	73.64±3.09	73.11±2.87	49.30±2.65	65.35
SimMMDG	✓	✓	✓	60.31±1.96	68.75±1.09	60.97±1.01	63.34	76.38±0.65	74.36±1.70	51.59±2.50	67.44
MOOSA	✓	✓	✓	60.00±0.19	67.38±1.07	64.24±5.15	63.87	75.34±3.11	74.54±1.85	48.44±4.82	66.11
CMRF	✓	✓	✓	61.86±0.74	69.58±0.39	64.97±0.14	65.47	78.25±0.43	78.07±0.50	**55.61±1.23**	70.64
MBCD	✓	✓	✓	63.63±0.39	73.18±0.59	65.92±0.44	67.58	81.28±1.00	78.11±0.32	55.48±1.60	71.62
CSI	✓	✓	✓	**63.87±0.27**	**73.98±0.45**	**66.32±0.62**	**68.06**	**82.30±0.18**	**80.35±0.33**	53.58±1.10	**72.08**

Table 1: Multi-source MMDG accuracy on EPIC-Kitchens and HAC. Results (Top-1 Acc %) are averaged over three seeds; standard deviations are shown in ±. Best results are in bold.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Dataset. We evaluate our method on two benchmark datasets, EPIC-Kitchens (Damen et al., 2020) and Human-Animal-Cartoon (HAC) (Dong et al., 2023), both comprising video, optical flow, and audio modalities. For EPIC-Kitchens, we adopt the domain splits D1, D2, and D3, while HAC consists of three domains: human (H), animal (A), and cartoon (C).

Baselines. We compare our method, CSI, with seven baselines, including a naive multi-modal baseline ERM (based on feature concatenation) (Baltrušaitis et al., 2018) and several state-of-the-art MMDG methods, namely RNA-Net (Planamente et al., 2022), SimMMDG (Dong et al., 2023), MOOSA (Dong et al., 2024), CMRF (Fan et al., 2024), MDJA (Li et al., 2025b), and MBCD (Wang et al., 2026a).

Detailed descriptions of datasets, baselines, and other settings are provided in Appendix A.

4.2 MAIN RESULTS

Multi-modal multi-source DG. Table 1 compares CSI with competitive baselines on EPIC-Kitchens and HAC across four modality configurations. CSI achieves the highest average accuracy in all four configurations on EPIC-Kitchens and in two configurations on HAC. On EPIC-Kitchens, CSI outperforms MBCD by 1.47 percentage points with video–audio inputs and by 1.44 percentage points with audio–flow inputs. On HAC, it improves average accuracy by 1.35 percentage points over MBCD in the video–flow setting. These results demonstrate improved target-domain generalization across multiple modality combinations, although the gains vary with the dataset and configuration. However, the improvements are less consistent when generalizing to Cartoon, where CSI trails MBCD in three of the four modality configurations. Training dynamics show that MBCD increasingly concentrates removal on the dominant visual modality, with the audio removal probability approaching zero, whereas CSI maintains an audio removal probability above approximately

Method	S:	EPIC-Kitchens						Avg.	HAC						Avg.
		D1		D2		D3			H		A		C		
	T:	D2	D3	D1	D3	D1	D2		A	C	H	C	H	A	
ERM		43.42	41.99	42.18	42.24	42.28	51.98	44.02	57.36	41.21	69.38	42.89	56.77	57.28	54.15
RNA-Net		43.72	41.51	41.41	44.56	42.40	53.46	44.51	57.91	34.96	71.14	44.67	63.25	63.13	55.84
SimMMDG		54.09	49.21	53.56	57.27	54.42	64.74	55.55	67.03	42.74	69.72	46.26	66.71	65.82	59.71
MOOSA		54.99	49.76	52.83	55.15	55.49	65.27	55.58	62.21	43.41	73.78	**51.23**	63.90	68.18	60.45
CMRF		59.35	54.38	57.62	60.96	57.01	67.72	59.51	67.84	46.26	74.04	49.36	65.66	65.89	61.51
MBCD		**59.50**	**54.89**	58.11	61.73	57.18	70.77	60.36	**69.24**	**46.81**	76.16	49.36	73.37	71.93	64.48
CSI		59.02 ±0.67	54.57 ±1.19	**60.43** ±0.32	**63.07** ±0.21	**58.31** ±0.47	**71.72** ±0.51	**61.19**	68.54 ±0.62	41.36 ±0.89	**79.38** ±1.80	51.19 ±2.67	**77.79** ±1.01	**72.30** ±0.44	**65.09**

Table 2: Single-source MMDG accuracy on EPIC-Kitchens and HAC. Results (Top-1 Acc %) are averaged over three seeds; standard deviations are shown in ±. S/T denote source and target domains, respectively. Best results are in bold.

EMA	CL	SI	AO	D2,D3→D1	D1,D3→D2	D1,D2→D3	Avg.
				52.04±0.60	60.73±1.69	57.92±1.40	56.90
		✓		57.05±1.29	64.41±0.99	59.92±0.67	60.46
		✓	✓	58.37±0.78	65.30±0.68	60.57±0.83	61.41
✓		✓		58.88±1.14	66.85±0.45	63.52±0.46	63.08
✓		✓	✓	59.18±0.03	67.97±0.12	63.80±0.64	63.65
✓	✓	✓		59.48±0.53	69.00±0.68	64.07±0.08	64.18
✓	✓	✓	✓	**60.24±0.45**	**69.05±0.09**	**64.61±0.79**	**64.63**

Table 3: Ablation study of each component on EPIC-Kitchens with video and audio modalities

Method	D2,D3→D1	D1,D3→D2	D1,D2→D3	Avg.
w/o perturbation	57.97±1.01	67.30±1.34	61.65±0.36	62.31
random	58.63±0.16	65.57±0.70	60.58±0.67	61.59
confidence-based	59.20±0.86	68.07±1.51	62.04±0.53	63.10
sensitivity-based (ours)	**61.19±1.07**	**69.82±0.31**	**62.55±0.50**	**64.52**

Table 4: Comparison of different modality perturbation strategies on EPIC-Kitchens with flow and audio modalities under the MMDG setting for robustness evaluation.

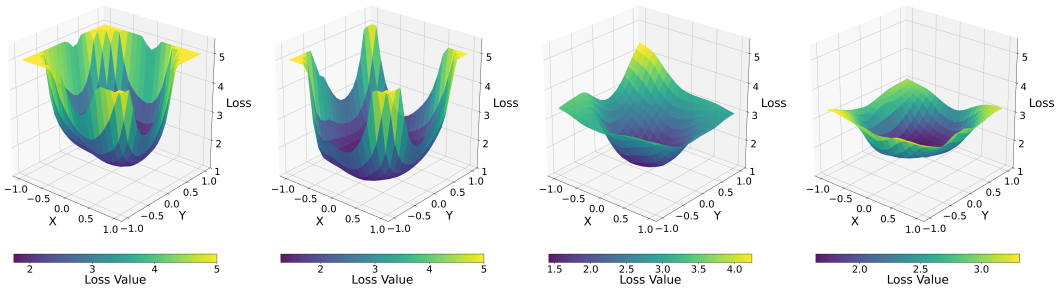
25%. We hypothesize that video undergoes a larger domain shift when transferring from Human and Animal to Cartoon. Concentrating removal on video may therefore better prepare the model to rely on other modalities when visual evidence becomes less transferable.

Multi-modal single-source DG. To further evaluate generalization with limited source-domain diversity, we consider the single-source DG setting, where models are trained on one source domain and evaluated on unseen target domains. As shown in Table 2, CSI achieves the highest average Top-1 accuracy on both benchmarks, reaching 61.19% on EPIC-Kitchens and 65.09% on HAC. These results exceed MBCD by 0.83 and 0.61 percentage points, respectively. On EPIC-Kitchens, CSI outperforms MBCD in four of the six transfer settings, including an improvement of 2.32 percentage points on D2 → D1. On HAC, CSI improves accuracy from 76.16% to 79.38% on A → H, yielding a gain of 3.22 percentage points. These results show that CSI remains effective when training is restricted to a single source domain. Results for additional modality combinations are provided in Appendix B.1.

4.3 Ablation Studies

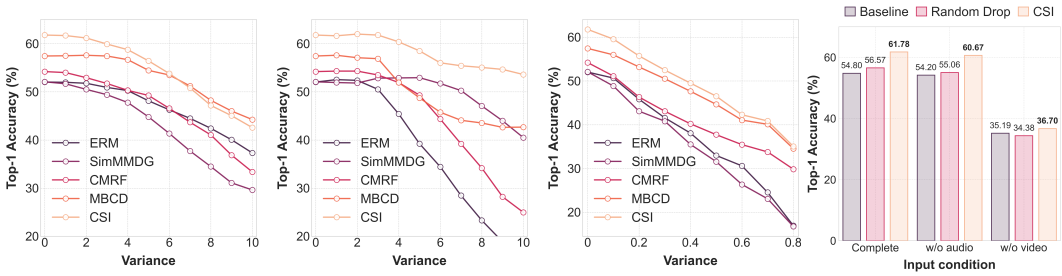
Ablation on each design. Table 3 evaluates the contribution of each component in CSI, where EMA, CL, SI, and AO denote exponential moving average, consistency loss, sensitivity-guided intervention, and annealed objective, respectively. Starting from the ERM baseline, adding SI improves target-domain accuracy, supporting the effectiveness of training with sensitivity-guided modality removal. AO provides further gains, supporting a gradual increase in the emphasis on prediction from remaining evidence. EMA and CL additionally improve performance, indicating the benefits of parameter averaging and complete-input teacher guidance. The full model achieves the best results across all target domains, supporting the combined use of these components for unseen-domain generalization.

Ablation on modality selection. Table 4 compares random, confidence-based, and sensitivity-guided modality selection. Random selection yields lower accuracy than confidence-based selection, while the proposed sensitivity-guided strategy achieves the best performance across all target domains. These results support using sample-specific responses to modality removal to guide intervention selection. By preferentially withholding modalities that strongly affect the current fused prediction, our strategy directs training toward prediction from the remaining evidence, improving unseen-domain generalization.



(a) ERM (b) CMRF (c) MBCD (d) CSI

Figure 3: 3D visualizations of the loss landscape on EPIC-Kitchens (video and audio modalities) with respect to all encoder parameters. CSI achieves a flatter solution compared to baselines.



(a) Video perturbation (b) Audio perturbation (c) Missing robustness (d) Remaining evidence

Figure 4: Robustness Analysis on EPIC-Kitchens with video and audio modalities. Results on HAC are provided in Appendix B.3.

4.4 Further Analysis

Flatness analysis. Prior work has linked flat minima to improved domain generalization (Cha et al., 2021). Motivated by this connection, we examine whether the generalization gains of CSI are accompanied by flatter local loss landscapes. Following Li et al. (2018a), we visualize the loss landscapes on EPIC-Kitchens in Figure 3, with additional HAC results provided in Appendix B.2. ERM exhibits a steep increase in loss under parameter perturbations, whereas CMRF shows a more gradual increase. Both MBCD and CSI exhibit flatter landscapes, with CSI displaying the broadest low-loss region in the visualized parameter directions. These observations suggest that CSI learns solutions with lower local sensitivity to parameter perturbations, complementing its improved target-domain performance.

Robustness to feature perturbations. To evaluate robustness against feature perturbations, we inject Gaussian noise with increasing variance into audio or video features during inference and report the corresponding accuracy degradation in Figure 4a and Figure 4b. The x-axis denotes the variance of the injected noise. Since feature distributions differ across modalities, noise strengths are compared only within the same modality. Overall, all methods exhibit decreasing accuracy as noise variance increases. ERM shows severe degradation, indicating high sensitivity to perturbations. SimMMDG and CMRF improve robustness but still degrade noticeably under large noise variance. In contrast, MBCD and CSI maintain slower performance decay, demonstrating stronger robustness. Notably, CSI consistently achieves the best performance under both audio and video perturbations, especially at high variance levels. These results suggest that the proposed sensitivity-guided regularization promotes stable multimodal representations and reduces sensitivity to noisy features.

Robustness against missing modalities. To evaluate robustness under incomplete multimodal observations, we randomly mask input modalities with different probabilities during inference. Figure 4c shows the Top-1 accuracy as the masking probability increases from 0 to 0.8. As expected, all methods exhibit performance degradation under severe modality missing. ERM and SimMMDG de-

Method	Time/Epoch	Peak Mem.	Params	Acc.
ERM	226.34s	2562.3MB	31.28M	52.04
SimMMDG	228.85s	2949.6MB	64.24M	56.35
CMRF	261.22s	4601.0MB	47.19M	56.77
MBCD	296.61s	6003.7MB	32.73M	58.06
CSI	247.00s	4412.7MB	32.73M	60.24

Table 5: Efficiency comparison on EPIC-Kitchens with video and audio modalities.

teriorate rapidly under missing modalities, indicating limited tolerance to missing inputs. Although CMRF improves robustness to some extent, its performance still drops noticeably at high masking ratios. MBCD shows substantially stronger robustness, validating the benefit of mitigating modality dominance during training. Nevertheless, CSI still achieves the best performance across different masking ratios. These results suggest that CSI encourage flatter and more redundancy-aware multimodal representations, leading to more stable predictions under incomplete modality inputs.

Generalization and Prediction from Remaining Evidence. We compare Baseline, Random Drop, and CSI in the EPIC-Kitchens $\mathrm{D2} + \mathrm{D3} \rightarrow \mathrm{D1}$ setting to examine target-domain generalization and prediction from remaining evidence. For each trained checkpoint, we evaluate complete target inputs and separately zero the audio or video feature block before fusion, reporting accuracy over all target samples. As shown in Fig. 4d, CSI achieves 61.78% accuracy on complete inputs, compared with 54.80% for Baseline and 56.57% for Random Drop. After audio or video removal, CSI achieves 60.67% and 36.70%, respectively, exceeding Random Drop by 5.61 and 2.33 percentage points. These results show that improved complete-input generalization is accompanied by higher accuracy when either modality is withheld, consistent with our motivation of learning to predict from remaining evidence. This comparison assesses the complete training configurations, which differ in multiple components.

Efficiency analysis. Since recent MMDG methods often employ auxiliary teacher networks, we evaluate the efficiency of CSI. Table 5 reports training time per epoch, peak GPU memory, trainable parameters, and out-of-domain accuracy under the same setting. The parameter count includes only trainable parameters involved in backpropagation. Although CSI maintains an EMA teacher during training, its parameters are updated via EMA rather than gradient optimization. During inference, the EMA model shares the same architecture and parameter size as the base ERM model, incurring no additional overhead. Compared with teacher-based approaches such as CMRF and MBCD, CSI achieves better generalization with lower training overhead, reducing both training time and GPU memory consumption relative to MBCD while maintaining the same number of trainable parameters. These results demonstrate an effective trade-off between robustness and efficiency.

5 Conclusion

In this paper, we examine MMDG through cross-domain changes in modality contributions, showing that evidence beneficial in source domains can become less useful or even detrimental in an unseen target domain. Motivated by this observation, we propose CSI, which uses online counterfactual sensitivity to guide sample-specific modality removal and encourage prediction from the remaining evidence. An annealed objective progressively increases the emphasis on masked-input learning, while complete-input guidance from an EMA teacher supports this process through stable distillation targets. Experiments on EPIC-Kitchens and HAC demonstrate improved generalization under both multi-source and single-source DG settings, alongside robustness to modality-specific perturbations and missing inputs. Loss-landscape visualizations further reveal flatter local regions around the learned solutions. Our findings highlight the value of preparing multimodal models to use alternative evidence when modality contributions change across domains.

AI use statement

In this work, we used generative AI tools to improve the language and clarity of the manuscript, assist with literature searches, support experimental coding and figure preparation, and check mathematical derivations for potential errors or inconsistencies. The authors reviewed all AI-assisted outputs, verified references and associated claims against the original publications, checked experi-

mental and plotting code against the underlying data, and independently examined the mathematical derivations. We take full responsibility for the final content of this work, including all text, code, figures, results, and claims produced with the assistance of generative AI.

REFERENCES

Niccolò Ajroldi, Antonio Orvieto, and Jonas Geiping. When, where and why to average weights? *arXiv preprint arXiv:2502.06761*, 2025.

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022.

Aristotelis Ballas and Christos Diou. Gradient-guided annealing for domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20558–20568, 2025.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

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.

Taiqin Chen, Yifeng Wang, Xiaochen Feng, Zhilin Zhu, Hao Sha, Yingjian Li, and Yongbing Zhang. Spectral property-driven data augmentation for hyperspectral single-source domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3038–3046, 2026.

Dong Kyu Cho, Inwoo Hwang, and Sanghack Lee. Peer pressure: Model-to-model regularization for single source domain generalization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15360–15370, 2025.

MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`, 2020.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 36:78674–78695, 2023.

Hao Dong, Eleni Chatzi, and Olga Fink. Towards multimodal open-set domain generalization and adaptation through self-supervision. In *European Conference on Computer Vision*, pp. 270–287. Springer, 2024.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, and Song Guo. Cross-modal representation flattening for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 37: 66773–66795, 2024.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Xudong Han, Senkang Hu, Yihang Tao, Yu Guo, Philip Birch, Sam Tak Wu Kwong, and Yuguang Fang. Dsp-reg: Domain-sensitive parameter regularization for robust domain generalization. *arXiv preprint arXiv:2601.19394*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hai Huang, Yan Xia, Sashuai Zhou, Hanting Wang, Shulei Wang, and Zhou Zhao. Bridging domain generalization to multimodal domain generalization via unified representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22488–22498, 2025.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Aodi Li, Liansheng Zhuang, Xiao Long, Minghong Yao, and Shafei Wang. Seeking consistent flat minima for better domain generalization via refining loss landscapes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15349–15359, 2025a.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018b.

Hongzhao Li, Hualei Wan, Liangzhi Zhang, Mingyuan Jiu, Shupan Li, Mingliang Xu, and Muhammad Haris Khan. Towards robust multimodal domain generalization via modality-domain joint adversarial training. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 180–188, 2025b.

Hongzhao Li, Hao Dong, Hualei Wan, Shupan Li, Mingliang Xu, and Muhammad Haris Khan. Towards multimodal domain generalization with few labels. *arXiv preprint arXiv:2602.22917*, 2026.

Siyuan Li, Zicheng Liu, Juanxi Tian, Ge Wang, Zedong Wang, Weiyang Jin, Di Wu, Cheng Tan, Tao Lin, Yang Liu, et al. Switch ema: A free lunch for better flatness and sharpness. *arXiv preprint arXiv:2402.09240*, 2024.

Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*, 2022.

Chuang Lin, Zehuan Yuan, Sicheng Zhao, Peize Sun, Changhu Wang, and Jianfei Cai. Domain-invariant disentangled network for generalizable object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8771–8780, 2021.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *arXiv preprint arXiv:2411.18704*, 2024.

Ze Peng, Jian Zhang, Yisen Wang, Lei Qi, Yinghuan Shi, and Yang Gao. Leveraging flatness to improve information-theoretic generalization bounds for sgd. *arXiv preprint arXiv:2601.01465*, 2026.

Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1807–1818, 2022.

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12556–12565, 2020.

Sanqing Qu, Yingwei Pan, Guang Chen, Ting Yao, Changjun Jiang, and Tao Mei. Modality-agnostic debiasing for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24142–24151, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3769–3778, 2023.

Xiaohan Wang, Zhangtao Cheng, Ting Zhong, Leiting Chen, and Fan Zhou. Modality-balanced collaborative distillation for multi-modal domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26535–26543, 2026a.

Xiaohan Wang, Xueting Liu, Wenxin Tai, Joojo Walker, Yong Wang, Kai Chen, and Fan Zhou. Invariant learning improves out-of-distribution generalization for ip geolocation. *Information Processing & Management*, 63(5):104641, 2026b.

Changsong Wen, Zelin Peng, Yu Huang, Xiaokang Yang, and Wei Shen. Domain generalization in clip via learning with diverse text prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9559–9569, 2025.

Zhipeng Xu, De Cheng, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Adversarial domain prompt tuning and generation for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18584–18595, 2025.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026.

Yavuz Yarici and Ghassan AlRegib. MER-DG: Modality-entropy regularization for multimodal domain generalization. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=IOUmfPYShr`.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Di Zhao, Jingfeng Zhang, Hongsheng Hu, Philippe Fournier-Viger, Gillian Dobbie, and Yun Sing Koh. Unlearning during training: Domain-specific gradient ascent for domain generalization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=9ufS5Jl0O0`.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. *arXiv preprint arXiv:2410.10373*, 2024.

A EXPERIMENTAL SETTING

Dataset. We evaluate our method on two widely used multimodal domain generalization benchmarks: EPIC-Kitchens (Damen et al., 2020) and Human-Animal-Cartoon (HAC) (Dong et al., 2023), following the experimental protocols of their original works. EPIC-Kitchens is an egocentric action recognition dataset containing eight action categories—'put', 'take', 'open', 'close', 'wash', 'cut', 'mix', and 'pour'—captured in three distinct kitchen environments, denoted as domains D1, D2, and D3. HAC is a cross-domain multimodal action recognition dataset consisting of seven action classes—'sleeping', 'watching TV', 'eating', 'drinking', 'swimming', 'running', and 'opening door'—performed by humans (H), animals (A), and cartoon characters (C), corresponding to domains H, A, and C. C

Both datasets provide three temporally aligned modalities, including RGB video, audio, and optical flow, enabling comprehensive evaluation under multimodal domain shifts. The number of action segments in each domain is summarized in Table 6. For every domain, the data are divided into training and validation subsets following the standard protocol. When a domain is used as a source domain, the training subset is used for model optimization, while the validation subset is employed for model selection and early stopping. In contrast, when a domain is designated as the target domain, the entire dataset, including both training and validation samples, is reserved exclusively for out-of-domain evaluation. This setting ensures that no target-domain information is accessible during training and provides a strict evaluation protocol for measuring generalization performance under domain shifts.

Mode	**EPIC-Kitchens**			**HAC**		
Domain	D1	D2	D3	H	A	C
Training	1543	2495	3897	1111	730	870
Validate	435	750	974	276	176	218
Testing	1978	3245	4871	1387	906	1088

Table 6: Number of action segments per domain.

Baseline. We provide a detailed description of the baselines used in our evaluation.

- **ERM** is adopted as a basic baseline, where multimodal features are directly concatenated and optimized without any dedicated DG strategy.
- **RNA-Net** tackles modality imbalance by introducing a cross-modal objective that aligns the feature magnitudes across different modalities.
- **SimMMDG** separates multimodal representations into shared and modality-specific components, leveraging contrastive learning for shared features and additional constraints to regularize modality-specific representations.
- **MOOSA** enhances generalization through carefully designed multimodal self-supervised objectives, such as cross-modal reconstruction and jigsaw-based pretext tasks.
- **CMRF** aims to improve consistency in the shared representation space by combining cross-modal interpolation with teacher-student distillation, encouraging smoother optimization behavior.
- **MDJA** introduces a tri-discriminator adversarial framework with a stability-aware dynamic weighting strategy, extending unimodal domain-invariant learning to the multimodal setting.
- **MBCD** identifies that differences in optimization dynamics across modalities lead to modality-biased sharp minima, and mitigates this issue via a collaborative distillation framework to alleviate modality imbalance.

Backbone. To ensure fair comparison with prior work, we adopt the same modality-specific backbones as SimMMDG (Dong et al., 2023) and CMRF (Fan et al., 2024), following the implementation protocols in (Damen et al., 2020). Experiments are conducted on three modalities: RGB video, audio, and optical flow. Our implementation is built upon the MMAction2 framework (Contributors, 2020). For visual representation learning, we employ the SlowFast architecture (Feichtenhofer

et al., 2019) initialized with Kinetics-400 pre-trained weights. The audio modality is encoded using a ResNet-18 (He et al., 2016) backbone initialized from the VGGSound checkpoint (Chen et al., 2020). For optical flow, we adopt the slow pathway of SlowFast, also initialized with Kinetics-400 pre-trained weights.

Data augmentation. We employ both input-level augmentation and feature-level augmentation during training to improve robustness and representation stability across modalities. For input-level augmentation, we follow the standard MMAction2 (Contributors, 2020) augmentation pipeline for the RGB video and optical flow modalities, including temporal frame sampling, random resized cropping, horizontal flipping, resizing, and normalization. For the audio modality, raw waveforms are converted into log-spectrograms, followed by random noise perturbation, temporal masking, and normalization during training.

For feature-level augmentation, we follow MBCD (Wang et al., 2026a) and adopt the uncertainty-based feature augmentation strategy proposed in (Li et al., 2022). Since the original method is designed for 2D convolutional networks, we further extend it to the 3D convolutional backbones used for the video and optical flow modalities. Specifically, stochastic perturbations are applied to intermediate feature statistics during training to improve robustness against domain shifts. The same feature augmentation is also applied to the audio modality.

Implementation details. The dimensionality of unimodal feature representations is set to 2304 for video, 512 for audio, and 2048 for optical flow. All models are optimized using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0001 and a batch size of 16. The unimodal loss coefficient λ is fixed at 1.0 throughout all experiments. The annealing coefficient α_t gradually decreases from 0.5 to 0 during training. The sensitivity threshold δ is set to 2 for HAC and 1 for EPIC-Kitchens. Following common practice, the EMA smoothing coefficient β in Eq. 12 is set to 0.999 to stabilize parameter updates and improve training stability.

Our experimental protocol follows (Dong et al., 2023). For each method, we select the model with the best in-domain validation performance and evaluate it on out-of-domain test domains. Results are reported as Top-1 accuracy averaged over three random seeds. All experiments are conducted on a machine equipped with NVIDIA GeForce RTX 4090 D and 4090 GPUs. Each model is trained for 20 epochs, with a total training time of approximately three hours. To ensure statistical robustness, we report both the mean and standard deviation across three runs with different random seeds.

B More Results

B.1 Multi-Modal Single-Source DG

To further investigate the robustness of our method under varying modality availability, we conduct additional experiments using all bimodal combinations (Video+Audio, Video+Flow, Audio+Flow) in the single-source domain generalization setting, as shown in Table 7 and Table 8. Results on both EPIC-Kitchens and HAC datasets demonstrate that our method consistently outperforms existing approaches across all bimodal settings. Notably, our approach achieves the highest average Top-1 accuracy on EPIC-Kitchens and matches or exceeds the strongest baselines on HAC, underscoring its strong generalization capability across diverse modality configurations.

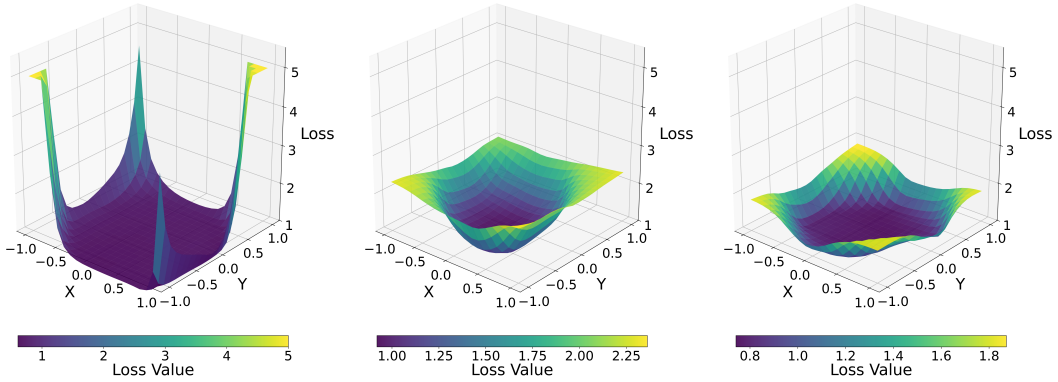
The performance gains are particularly pronounced in combinations involving optical flow, such as Video+Flow and Audio+Flow, suggesting that CSI effectively address the issue of modality imbalance. Moreover, our method exhibits reduced performance variance compared to competing methods, indicating more stable training dynamics. Despite differences in action categories and domain discrepancies across datasets, our method maintains competitive performance, highlighting its practical utility in real-world scenarios where certain modalities may be unreliable.

B.2 3D Visualizations on HAC

We further visualize the loss landscape on the HAC dataset following the same protocol as EPIC-Kitchens, with results shown in Figure 5. Similar trends are observed across different methods. both MBCD and CSI converge to flatter regions, with CSI showing the broadest and smoothest basin

Method	Modality			Source:	EPIC-Kitchens						Avg.
					D1		D2		D3		
	Video	Audio	Flow	Target:	D2	D3	D1	D3	D1	D2	
ERM	✓	✓			50.24±1.79	50.50±1.10	45.25±0.57	52.67±3.12	48.11±0.25	57.67±2.01	50.74
RNA-Net	✓	✓			50.05±2.64	50.77±1.04	45.92±0.91	53.95±0.92	48.08±0.34	56.56±3.04	50.89
SimMMDG	✓	✓			49.23±1.56	44.54±3.77	48.55±0.66	55.62±0.97	48.28±1.50	59.50±2.00	50.95
MOOSA	✓	✓			48.58±1.50	48.66±1.23	49.22±1.42	54.45±1.94	49.48±1.18	59.32±1.45	51.62
CMRF	✓	✓			54.62±0.64	51.63±0.69	53.57±0.52	59.15±0.61	50.57±0.37	62.06±0.35	55.27
MBCD	✓	✓			52.46±2.33	50.21±0.98	54.82±1.20	59.66±1.10	51.85±1.08	65.21±1.12	55.70
CSI	✓	✓			**55.20±0.08**	**52.13±0.47**	**56.64±1.21**	**61.48±0.43**	**53.08±0.51**	**67.22±0.32**	**57.62**
ERM	✓		✓		55.10±0.85	45.97±3.31	51.40±1.67	54.69±1.02	54.28±2.20	61.41±0.77	53.81
RNA-Net	✓		✓		54.20±1.04	46.27±3.09	51.26±2.42	54.55±1.05	53.40±0.82	63.43±1.35	53.85
SimMMDG	✓		✓		51.84±0.94	44.69±2.75	52.21±0.54	55.85±1.50	52.86±1.12	62.39±1.28	53.31
MOOSA	✓		✓		52.32±0.60	47.65±0.18	52.06±1.36	54.60±1.95	56.42±0.39	65.17±0.49	54.70
CMRF	✓		✓		57.02±0.44	51.13±0.84	56.08±0.87	57.37±0.01	58.61±1.52	66.69±0.25	57.82
MBCD	✓		✓		**58.23±0.24**	51.66±1.47	**59.05±0.11**	59.01±1.22	59.34±0.62	**69.18±0.90**	**59.41**
CSI	✓		✓		56.70±1.20	**51.82±1.38**	57.50±0.66	**59.49±0.55**	**59.49±0.10**	68.97±0.08	59.00
ERM		✓	✓		47.93±0.34	46.53±0.45	45.64±0.72	51.87±0.43	48.33±2.23	55.88±0.94	49.36
RNA-Net		✓	✓		48.85±1.15	48.15±2.12	45.79±0.76	51.74±0.61	48.82±1.01	57.10±1.25	50.07
SimMMDG		✓	✓		51.63±0.52	50.13±1.02	48.52±0.93	52.40±0.84	54.03±1.17	62.44±1.09	53.19
MOOSA		✓	✓		47.04±1.50	47.21±0.54	47.99±1.51	53.39±1.08	54.11±1.21	61.51±1.33	51.88
CMRF		✓	✓		54.06±0.14	**52.21±0.54**	52.06±0.27	57.46±0.60	52.92±0.86	61.68±0.78	55.07
MBCD		✓	✓		52.16±1.11	49.84±0.53	52.16±0.27	57.50±0.46	55.49±0.42	65.15±0.61	55.38
CSI		✓	✓		**55.79±0.15**	50.34±0.61	**54.80±0.72**	**58.97±0.53**	**56.89±0.33**	**67.13±0.44**	**57.32**
ERM	✓	✓	✓		43.42±3.08	41.99±3.76	42.18±1.74	42.24±2.00	42.28±1.80	51.98±1.35	44.02
RNA-Net	✓	✓	✓		43.72±3.54	41.51±2.79	41.41±0.82	44.56±3.28	42.40±3.67	53.46±0.55	44.51
SimMMDG	✓	✓	✓		54.09±1.07	49.21±1.20	53.56±3.07	57.27±2.02	54.42±1.69	64.74±1.77	55.55
MOOSA	✓	✓	✓		54.99±0.76	49.76±1.57	52.83±1.35	55.15±2.59	55.49±2.23	65.27±1.06	55.58
CMRF	✓	✓	✓		59.35±0.24	54.38±0.32	57.62±0.37	60.96±0.30	57.01±0.32	67.72±0.61	59.51
MBCD	✓	✓	✓		59.50±0.60	54.89±0.93	58.11±0.84	61.73±0.45	57.18±0.65	70.77±0.44	60.36
CSI	✓	✓	✓		59.02±0.67	54.57±1.19	**60.43±0.32**	**63.07±0.21**	**58.31±0.47**	**71.72±0.51**	**61.19**

Table 7: Multi-modal **single-source** DG accuracy with different modalities on EPIC-Kitchens dataset. The results are averaged over 3 random seeds, with standard deviation displayed as well. The best is in **bold**.



(a) CMRF (b) MBCD (c) CSI

Figure 5: 3D visualizations of the loss landscape on HAC (video and audio modalities) with respect to all encoder parameters. CSI achieves a flatter solution compared to baselines.

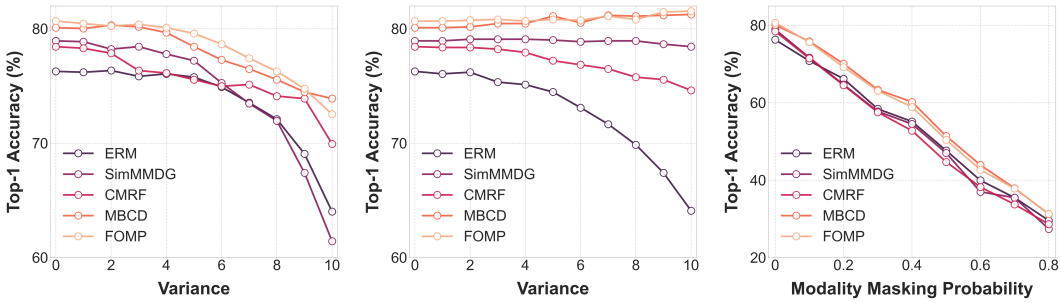
around the minimum. These results further support that the proposed direction-aware regularization promotes flatter and more stable optimization landscapes under domain shifts.

B.3 Robustness Analysis on HAC

We further conduct robustness analysis on the HAC dataset following the same protocol as EPIC-Kitchens, with results shown in Figure 6. Similar trends are observed across both feature perturbation and missing modality experiments. As perturbation variance or masking probability increases, all methods exhibit performance degradation, while MBCD and CSI maintain substantially stronger robustness than other baselines. Notably, CSI consistently achieves the best performance under

Method	Modality			Source:	HAC						Avg.
					H		A		C		
	Video	Audio	Flow	Target:	A	C	H	C	H	A	
ERM	✓	✓			63.32±3.88	40.78±2.12	69.31±2.60	45.25±0.80	62.58±1.68	70.49±1.00	58.62
RNA-Net	✓	✓			63.50±3.73	41.45±3.42	69.02±2.30	44.36±2.54	60.71±1.50	70.49±1.81	58.25
SimMMDG	✓	✓			63.54±3.31	42.37±3.60	71.81±2.97	47.43±4.46	66.96±3.66	68.32±3.91	60.07
MOOSA	✓	✓			65.71±2.21	41.08±3.93	74.38±2.08	47.21±2.20	63.13±3.84	69.87±1.50	60.23
CMRF	✓	✓			67.22±1.10	**44.70±1.78**	74.93±0.39	49.82±1.31	71.09±0.77	70.82±0.21	63.10
MBCD	✓	✓			**68.21±1.26**	38.60±5.23	75.82±2.86	45.62±1.73	65.51±3.84	69.72±1.54	60.58
CSI	✓	✓			67.05±2.26	42.60±0.32	**80.17±0.50**	**50.51±1.61**	**74.62±2.52**	**74.28±0.11**	**64.87**
ERM	✓		✓		64.31±0.68	44.09±3.28	70.87±4.19	43.87±3.16	64.98±2.31	61.63±2.55	58.29
RNA-Net	✓		✓		64.97±3.34	39.09±3.35	71.26±4.25	41.70±3.12	64.36±0.62	63.47±2.77	57.48
SimMMDG	✓		✓		65.78±2.36	41.76±1.15	75.29±0.89	48.10±1.35	65.61±2.95	58.90±2.86	59.24
MOOSA	✓		✓		65.05±1.36	43.44±2.47	73.06±1.49	47.09±4.78	64.43±3.71	61.37±2.71	59.07
CMRF	✓		✓		**67.07±1.71**	**44.27±2.04**	74.21±0.86	**50.55±1.70**	69.31±1.11	64.64±1.52	61.68
MBCD	✓		✓		66.15±2.21	39.74±1.47	77.82±2.65	47.06±2.28	72.94±2.59	63.39±1.36	61.18
CSI	✓		✓		66.72±0.17	38.14±1.01	**81.33±1.15**	48.35±1.29	**78.12±0.47**	**67.11±2.43**	**63.29**
ERM		✓	✓		58.35±1.32	34.04±0.17	53.88±1.62	41.24±1.41	42.73±1.36	46.95±1.30	46.20
RNA-Net		✓	✓		56.48±2.25	34.22±0.43	54.15±0.85	41.33±2.25	40.42±2.12	45.62±1.94	45.37
SimMMDG		✓	✓		59.16±3.80	37.68±1.61	56.04±3.92	40.62±1.72	38.81±1.15	45.84±2.72	46.36
MOOSA		✓	✓		56.44±1.25	34.07±1.72	56.81±0.33	40.59±0.61	38.55±2.69	45.66±4.39	45.35
CMRF		✓	✓		59.60±1.96	**37.90±0.62**	59.89±1.09	**44.06±1.28**	42.85±1.68	48.09±1.99	48.73
MBCD		✓	✓		61.30±1.71	37.71±0.57	**61.84±2.87**	40.72±0.59	48.45±1.71	51.51±2.35	50.25
CSI		✓	✓		**61.53±0.17**	37.50±1.01	60.85±1.23	39.57±1.24	**51.55±0.58**	**52.15±0.39**	**50.52**
ERM	✓	✓	✓		57.36±2.36	41.21±1.95	69.38±2.44	42.89±4.93	56.77±7.48	57.28±2.97	54.15
RNA-Net	✓	✓	✓		57.91±1.58	34.96±5.02	71.14±0.92	44.67±1.46	63.25±3.72	63.13±3.72	55.84
SimMMDG	✓	✓	✓		67.03±2.23	42.74±1.99	69.72±2.97	46.26±5.83	66.71±3.29	65.82±4.26	59.71
MOOSA	✓	✓	✓		62.21±2.40	43.41±2.48	73.78±0.79	**51.23±3.56**	63.90±2.75	68.18±2.37	60.45
CMRF	✓	✓	✓		67.84±0.96	46.26±1.82	74.04±0.44	49.36±0.40	65.66±1.58	65.89±1.00	61.51
MBCD	✓	✓	✓		**69.24±0.75**	**46.81±0.68**	76.16±0.07	49.36±0.26	73.37±1.12	71.93±3.29	64.48
CSI	✓	✓	✓		68.54±0.62	41.36±0.89	**79.38**±1.80	51.19±2.67	**77.79**±1.01	**72.30**±0.44	**65.09**

Table 8: Multi-modal **single-source** DG accuracy with different modalities on HAC dataset. The results are averaged over 3 random seeds, with standard deviation displayed as well. The best is in **bold**.



(a) Video feature perturbation. (b) Audio feature perturbation. (c) Missing modality robustness.

Figure 6: Robustness Analysis on HAC.

both noisy and incomplete modality inputs, further demonstrating the effectiveness of the proposed sensitivity-guided regularization in learning stable multimodal representations.