FEATURE-PULLBACK: RETHINKING MULTIMODAL IMBALANCE THROUGH THE GEOMETRY OF FEATURE SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Modality imbalance prevents multimodal models from fully exploiting heterogeneous information and has become a major issue that limits joint learning performance. Existing studies reduce disparities between modalities by regulating gradients, learning speeds, or contributions of the modality. These methods expect modalities to receive the same optimization signal but rarely examine the feature-space updates induced by this signal in modality representation spaces. To this end, we revisit modality imbalance from the perspective of featurespace geometry. We empirically show that optimization signals undergo modality-specific matrix-valued responses as they propagate through the encoders, inducing feature updates with different directions, scales, and reachability. First, a controlled modality-manifold experiment shows that changing only the geometry of modality representations yields an 8.86% performance improvement. Further local optimization analysis reveals that task signals must pass through modality-specific feature response operators before being converted into feature updates. It also establishes a correspondence between response distortion in feature space and pullback geometry in parameter space. Based on this insight, we propose Feature-Pullback, which uses feature probes and layer-wise Kronecker structures to obtain a scalable approximation. It corrects the backward updates of each modality without altering the forward network. Experiments on four audio-visual and image-text benchmarks show that the method consistently improves modality learning and fusion performance and achieves new state-of-the-art results on all benchmarks.

1 INTRODUCTION

Multimodal learning aims to integrate heterogeneous information from different modalities to obtain representations that are more comprehensive and robust than those learned by unimodal models. In recent years, multimodal learning has been widely applied to tasks such as audio–visual recognition, emotion recognition, action understanding, and cross-modal retrieval (Baltrušaitis et al., 2019). In principle, different modalities provide complementary information, and joint training should therefore yield better predictive performance. However, a large body of research has shown that multimodal models are not always superior to unimodal models (Wang et al., 2020; Huang et al., 2022). During practical training, the model often becomes overly dependent on one dominant modality, while another modality is insufficiently learned, ultimately limiting the overall performance.

This phenomenon is commonly referred to as modality imbalance (Peng et al., 2022). To address this problem, existing methods coordinate the training dynamics of different modalities by adjusting their gradients, learning rates, loss weights, modality contributions, or subnetwork updates (Yao & Mihalcea, 2022; Yang et al., 2025). These methods demonstrate the importance of the optimization process for modality learning. However, their diagnosis and regulation are typically based on losses, predictions, gradients, or parameter updates, while paying relatively little attention to the feature space updates that these optimization signals ultimately induce in the modality representation space. This naturally raises a more fundamental question: **why does the same joint objective produce different effective learning behaviors across different modalities?**

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.

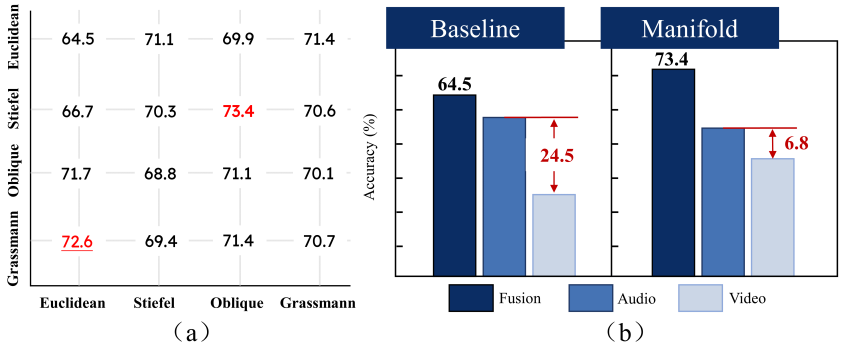


Figure 1: (a) Joint accuracy of different modality manifold combinations on Kinetics-Sounds, where the horizontal axis represents the video modality and the vertical axis represents the audio modality; (b) Fusion and branch performance of the baseline and the optimal manifold configuration. The best existing joint accuracy is 72.3%, and results that outperform existing methods are highlighted in **bold** and underline.

In this paper, we revisit this problem from the perspective of feature space geometry. We first design a controlled modality manifold experiment shown in Figure 1, where distinct manifolds are utilized to characterize diverse feature spaces. The experiment applies a low-dimensional manifold matrix M_m to the representation of modality m through $\widetilde{f}_m = M_m f_m$, thereby influencing the propagation of backward signals in the representation space. Specifically, we impose four predefined manifold structures, namely Euclidean, Stiefel, Oblique, and Grassmann (Absil et al., 2008), on the audio and video representations respectively while keeping all other training settings unchanged. When Euclidean manifolds are adopted for both audio and video branches, this setup degenerates into the conventional baseline method. A particularly intriguing finding can be drawn from our experimental results. Two different combinations of manifolds achieve accuracies of 73.4% and 72.6%, respectively, which even outperform the state-of-the-art rebalancing approaches.

This observation suggests that multimodal learning is highly sensitive to the local structure of representations, and that different modalities may reside in feature spaces with different local geometries. However, the representation structures in real networks are far more complex than these limited predefined manifolds and continuously evolve with the data, parameters, and training process. Therefore, we extend this low-dimensional phenomenon to the optimization process of neural networks. Our local analysis of one SGD update reveals an often-overlooked fact: the task signal returned by the joint objective is not equivalent to the representation update, but instead first undergoes the matrix-valued response induced by the encoder parameterization, $J_m J_m^\top$, where J_m is the local Jacobian of the m-th modality encoder. We define it as the modality representation response operator $\mathcal{R}_m = J_m J_m^\top$.

This indicates that the same joint objective does not induce identical descent geometries across different modalities; each modality encoder imposes different directional transformations, response magnitudes, and reachability constraints on task-related components. Therefore, modality imbalance is not only reflected by the magnitude differences of optimization signals, but also originates from the different ways in which joint supervision is transformed into effective representation learning.

Based on this understanding, we propose **F**eature-**P**ullback (FP), a multimodal training framework based on feature space geometry. FP does not directly match the gradients, performance, or representation trajectories of different modalities. Instead, it takes the representation descent direction provided by the joint objective as the reference and inversely solves for the parameter update that realizes this direction, thereby pulling back the update criterion in the representation space into the parameter space. To avoid explicitly constructing the complete pullback geometry, we further leverage feature probes and a layerwise Kronecker structure to obtain a scalable implementation, without modifying the forward network or fusion architecture. Experiments on four audio–visual and vision–language benchmarks demonstrate that FP consistently improves modality learning and fusion performance across different networks, fusion strategies, and base optimizers, and achieves

new state-of-the-art results on all benchmarks. The main contributions of this paper are summarized as follows:

- We show through controlled manifold experiments that changing the geometry of modality representations alone can substantially affect joint learning, and accordingly revisit modality imbalance from the perspective of feature space geometry.
- We reveal the modality representation response operator $\mathcal{R}_m = J_m B_m J_m^\top$, which characterizes the direction, magnitude, and reachability of the representation updates induced by the task signal after encoder parameterization.
- We propose the Feature-Pullback framework, which inversely solves for parameter updates from target feature space updates and achieves scalable pullback correction through feature probes and a layerwise Kronecker structure.

2 Related Work

Modality Imbalance. To address modality imbalance in joint training, existing methods primarily regulate branch optimization based on the modality states observed during training. G-Blend (Wang et al., 2020) learns loss weights according to the generalization behavior of different prediction branches; OGM-GE (Peng et al., 2022) estimates the degree of modality dominance using unimodal predictions and adjusts the corresponding gradients; Wu et al. (2022) constrain modality learning from the perspective of the multimodal model's preferential reliance on easy-to-learn modalities; PMR (Fan et al., 2023) and AGM (Li et al., 2023) coordinate the updates of different modalities based on class prototypes and relative optimization progress, respectively. Furthermore, MMPareto (Wei & Hu, 2024) derives a Pareto update direction by addressing gradient conflicts between unimodal auxiliary objectives and the joint objective; Yang et al. (2024) dynamically combine supervised learning and contrastive learning according to class fitting difficulty; AMSS (Yang et al., 2025) selects parameter subnetworks to be updated based on modality saliency, and further applies AMSS+ to perform unbiased correction for sampled updates.

Despite differences in their criteria and implementations, these methods follow a similar research paradigm: they identify modality differences from predictions, losses, gradients, or network states, and subsequently regulate the supervisory signals, gradient magnitudes, or parameter update scopes assigned to different modalities. Collectively, these methods demonstrate that coordinating the joint optimization process can effectively alleviate modality imbalance.

3 Method

We propose **F**eature-**P**ullback (FP), a training framework for alleviating multimodal imbalance. The proposed FP advances the analysis endpoint of multimodal optimization from parameter updates to representation updates, and characterizes the local mechanism by which joint task signals are realized as feature updates in different modalities. Based on this analysis, we further derive the ideal parameter updates and provide a scalable implementation of Feature-Pullback. The entire framework does not predefine that modalities should have identical gradients, performance, or representation trajectories, but instead corrects the local response of each modality encoder when transforming task signals into induced feature displacements.

3.1 Theoretical Motivations

In this section, we introduce the theoretical motivation of the proposed FP method. We first give the setup of the problem and the notation. Consider a joint learning model containing M modalities. Specifically, we assume that the data pair is $(\mathbf{x}, y)$, where $\mathbf{x} = (x_1, \ldots, x_M)$ and x_m denotes the m-th modality, and y is the label. For the m-th modality, its encoder E_m maps the input x_m into a representation:

$$f_m = E_m(x_m; \theta_m),$$

where $f_m \in \mathbb{R}^{d_m}$, and $\theta_m \in \mathbb{R}^{p_m}$ denotes the corresponding encoder parameters. The representations from different modalities are passed through a fusion function F and a prediction head C to

obtain the final output. The joint training objective is written as

$$\mathcal{L}(\Theta) = \ell(C(F(f_1, \ldots, f_M)), y), \tag{1}$$

where Θ denotes all trainable parameters θ_1, ..., θ_M of the model and ℓ is the loss function of interest. We denote the task signal returned by the joint objective with respect to the representation of the m-th modality as $h_m \triangleq \nabla_{f_m}\mathcal{L} \in \mathbb{R}^{d_m}$, and define the local Jacobian of the encoder with respect to its parameters as:

$$J_m \triangleq \frac{\partial f_m}{\partial \theta_m} \in \mathbb{R}^{d_m \times p_m}.$$

According to the chain rule, the gradient of objective with respect to the encoder parameter of m-th modality is

$$g_m \triangleq \nabla_{\theta_m}\mathcal{L} = J_m^\top h_m. \tag{2}$$

The benefit of multimodal learning ultimately depends on whether joint training can learn useful representations for each modality. However, parameter gradients and parameter updates are only intermediate quantities in the optimization process, whereas multimodal fusion operates directly on the representations produced by the encoders. Therefore, instead of treating parameter updates as the endpoint of analysis, this section examines the feature updates induced by a single joint optimization step in the modality representation space.

We first consider a local reference case. If f_m is a free variable that can be directly optimized in the Euclidean representation space, then the steepest descent direction given by the joint objective at the current point is $-h_m$. With a step size η, the reference update is

$$\Delta f_m^{\text{direct}} = -\eta h_m. \tag{3}$$

Equation (3) only characterizes the local update that the joint objective intends to produce in the representation space. In a parameterized network, f_m is determined by the encoder parameters θ_m, and the task signal must first be propagated to the parameter space through backpropagation. Taking SGD (Bottou et al., 2018) as an example, then the encoder parameter is updated by $\theta_m \leftarrow \theta_m - \eta\nabla_{\theta_m}\mathcal{L}$, and we define the update of parameter θ_m for one step as

$$\Delta\theta_m \triangleq -\eta\nabla_{\theta_m}\mathcal{L} = -\eta J_m^\top h_m. \tag{4}$$

To determine the effect of this parameter update on the representation, we perform a first-order Taylor expansion of the encoder around the current parameters:

$$f_m(\theta_m + \Delta\theta_m) = f_m(\theta_m) + J_m\Delta\theta_m + \mathcal{O}(\|\Delta\theta_m\|_2^2). \tag{5}$$

Therefore, the update of representation f_m for one step is

$$\begin{aligned}\Delta f_m &= f_m(\theta_m + \Delta\theta_m) - f_m(\theta_m) \approx J_m\Delta\theta_m \\ &= -\eta J_m J_m^\top h_m.\end{aligned} \tag{6}$$

Equation (6) reveals an overlooked mechanism in multimodal joint optimization: the task signal h_m returned by the joint objective does not directly become a representation update, but is transformed by the modality-specific encoder response $J_m J_m^\top$. Because modalities have different encoder parameterizations, the same joint objective induces modality-specific effective learning.

To characterize how this matrix response shapes representation learning, we eigendecompose $J_m J_m^\top$ as

$$J_m J_m^\top = U_m \Lambda_m U_m^\top,$$

where $U_m = [u_{m,1}, \ldots, u_{m,d_m}]$ consists of orthonormal directions in the representation space, and $\Lambda_m = \text{diag}(\lambda_{m,1}, \ldots, \lambda_{m,d_m})$, with $\lambda_{m,i} \geq 0$ denoting the response gain along $u_{m,i}$. We further decompose the task signal as

$$h_m = \sum_{i=1}^{d_m} \alpha_{m,i} u_{m,i}, \qquad \alpha_{m,i} = u_{m,i}^\top h_m,$$

where $\alpha_{m,i}$ is the projection of the task signal onto the i-th response direction. Equation (6) can therefore be written as

$$\Delta f_m \approx -\eta \sum_{i=1}^{d_m} \lambda_{m,i}\alpha_{m,i} u_{m,i}. \tag{7}$$

Equation (7) shows that the representation learning is jointly determined by the task signal and its alignment with the encoder response spectrum. The extent to which the component $\alpha_{m,i}$ along $u_{m,i}$ is realized depends on the response gain $\lambda_{m,i}$. Consequently, task signals with similar norms may still produce different representation displacements. Since $h_m^\top \Delta f_m \approx -\eta \sum_i \lambda_{m,i} \alpha_{m,i}^2 \leq 0$, the positive semidefinite response does not locally reverse the overall descent direction, but instead shapes the direction, magnitude, and reachability of the task signal. This is what we refer to as the *representation update geometry*.

This analysis can be further generalized to other first-order optimization method such as adaptive SGD. For example, the encoder parameter can be updated by $\theta_m \leftarrow \theta_m - \eta B_m \nabla_{\theta_m} \mathcal{L}$, where B_m is the adaptive term, and then the corresponding update of parameter θ_m for one step is $\Delta\theta_m = -\eta B_m J_m^\top h_m$. Following the similar analysis, the representation update can be approximated as

$$\Delta f_m \approx -\eta \underbrace{J_m B_m J_m^\top}_{\mathcal{R}_m} h_m. \tag{8}$$

Here, B_m describes the local update metric imposed by the underlying optimizer in the parameter space. SGD corresponds to $B_m = I$, whereas Adam(Kingma & Ba, 2015) corresponds to a diagonal preconditioner induced by the historical second moments. Detailed derivations and settings are provided in Appendix D.

This result also connects the controlled manifold experiment in Figure 1 to optimization in real networks. The experiment applies the explicit transformation $\widetilde{f}_m = M_m f_m$ to each modality representation. By the chain rule:

$$\frac{\partial \mathcal{L}}{\partial f_m} = \frac{\partial \mathcal{L}}{\partial \widetilde{f}_m} M_m^\top.$$

In addition to modifying the forward representation, $M_m^\top$ explicitly reshapes the direction, scale, and dimensional coupling of the backward task signal. Taking the task signal on the transformed representation as the reference, $M_m^\top$ constitutes an explicitly prescribed matrix response acting on that signal, whereas $\mathcal{R}_m$ in Equation (8) generalizes this role to an endogenous response jointly induced by the encoder parameterization and optimizer. In this sense, $M_m^\top$ can be viewed as a controlled and explicit form of the matrix-valued response characterized by $\mathcal{R}_m$, placing the low-dimensional manifold intervention and the representation-update geometry of real networks under a common operator view.

Accordingly, the analysis of modality imbalance should move beyond asking "what optimization signal does each modality receive?" to a more direct question: after passing through their respective representation responses, what feature updates do the task signals produced by the joint objective ultimately induce in the representation space of each modality?

3.2 Pullback Formulation

To transform previous analysis into a training method, instead of accepting the representation response induced by a given parameter gradient, we start from the reference update $-\eta h_m$ in the representation space and solve inversely for the parameter change that realizes this target. The objective here is not to match the gradients, update magnitudes, or performance across different modalities, but to pull back the representation-space update requirement of each modality into its parameter space. Under SGD, this objective can be written as

$$\Delta\theta_m^\star = \arg\min_{\Delta\theta_m} \|J_m \Delta\theta_m + \eta h_m\|_2^2, \tag{9}$$

If $v \in \mathrm{Null}(J_m)$, then $\Delta\theta_m$ and $\Delta\theta_m + v$ produce the same representation update under the first-order approximation. To eliminate these redundant parameter components that do not change the representation, we select the solution of Eq. (9) with the minimum parameter norm. According to the Moore–Penrose pseudoinverse (Penrose, 1955), the solution is $\Delta\theta_m^\star = -\eta J_m^\dagger h_m$. Using $J_m^\dagger = (J_m^\top J_m)^\dagger J_m^\top$ together with the standard task gradient $g_m = J_m^\top h_m$, it can be written as

$$\Delta\theta_m^\star = -\eta \left(G_m^{\mathrm{FP}}\right)^\dagger g_m, \qquad G_m^{\mathrm{FP}} \triangleq J_m^\top J_m. \tag{10}$$

Equation (10) does not artificially specify a parameter preconditioner, but is directly derived from the representation update objective in Eq. (9). To understand its geometric meaning, consider an arbitrary local parameter perturbation $d\theta_m$, whose induced representation change satisfies

$$\|df_m\|_2^2 = \|J_m d\theta_m\|_2^2 = d\theta_m^\top G_m^{\mathrm{FP}} d\theta_m.$$

Therefore, G_m^{FP} is the local *pullback geometry* obtained by pulling back the Euclidean metric of the representation space to the parameter space through the encoder mapping. It measures, in parameter coordinates, the feature displacement induced by a parameter perturbation. Consequently, the "pull-back" in Feature-Pullback does not refer to the mapping of the task signal by $J_m^\top h_m$ in standard backpropagation, but to pulling back the update criterion and its metric in the representation space into the parameter space, and conditioning the joint-loss task gradient accordingly.

This parameter-space correction forms a closed loop with the representation response in Section 3.1. Let $J_m = U_m \Sigma_m V_m^\top$. The response induced by standard SGD is $J_m J_m^\top = U_m \Sigma_m^2 U_m^\top$, where each reachable direction is non-uniformly modulated by $\sigma_{m,i}^2$. In contrast, the ideal Feature-Pullback satisfies

$$J_m \left(G_m^{\mathrm{FP}}\right)^\dagger J_m^\top = U_m U_m^\top = P_{\mathrm{Range}(J_m)}.$$

Therefore, under the ideal setting of SGD with the exact pseudoinverse, Feature-Pullback produces $\Delta f_m^\star \approx -\eta P_{\mathrm{Range}(J_m)} h_m$, which, relative to the reference update in Eq. (3), corrects the response gain along the reachable directions from $\sigma_{m,i}^2$ to a unit response while leaving the unreachable directions unchanged. The proposed method does not require different modalities to obtain identical updates. Instead, it independently reduces the dependence of the actual update of each modality on the non-uniform response gain. The composite form with Adam is provided in the Appendix D.

3.3 Scalable Approximation

The storage complexity of the target matrix $G_m^{\mathrm{FP}} \in \mathbb{R}^{p_m \times p_m}$ is $O(p_m^2)$. When $p_m = 10^7$, a dense float32 matrix alone requires approximately 400 TB of memory. Our key observation is that the gradient outer product of each network layer naturally exhibits a Kronecker structure (Martens & Grosse, 2015), allowing the complete pullback to be decomposed into a set of computable layerwise factors.

Consider the l-th linear layer, $s_{m,l} = W_{m,l} a_{m,l}$, where $a_{m,l}$ denotes the input activation. For the representation coordinate f_m^c, let its sensitivity with respect to the output of this layer be $\delta_{m,l}^c = \frac{\partial f_m^c}{\partial s_{m,l}}$. According to the chain rule,

$$\frac{\partial f_m^c}{\partial W_{m,l}} = \delta_{m,l}^c a_{m,l}^\top, \qquad \mathrm{vec}\left(\frac{\partial f_m^c}{\partial W_{m,l}}\right) = a_{m,l} \otimes \delta_{m,l}^c.$$

Therefore, the pullback block of the l-th layer can be written as

$$\begin{aligned}\left[G_m^{\mathrm{FP}}\right]_l &= \mathbb{E}\left[(a_{m,l} a_{m,l}^\top) \otimes \left(\sum_c \delta_{m,l}^c \delta_{m,l}^{c\top}\right)\right] \\ &\approx \underbrace{\mathbb{E}\left[a_{m,l} a_{m,l}^\top\right]}_{A_{m,l}} \otimes \underbrace{\mathbb{E}\left[\sum_c \delta_{m,l}^c \delta_{m,l}^{c\top}\right]}_{S_{m,l}^f}.\end{aligned} \tag{11}$$

Equation (11) approximates the pullback block of the l-th layer by two factors with clear interpretations. $A_{m,l}$ describes the second-order structure of the activations entering this layer and can be directly estimated during the forward pass. $S_{m,l}^f$ describes the *feature sensitivity* of how parameter changes in this layer are propagated to the modality representation, which constitutes the primary computational bottleneck. Directly computing the latter requires backpropagation with respect to every representation coordinate. To aggregate the sensitivities of all coordinates within a single backward pass, we construct isotropic *feature probes* in the representation space:

$$\xi_m^{(r)} \sim \mathcal{N}(0, I), \qquad q_m^{(r)} = \langle \xi_m^{(r)}, f_m \rangle.$$

Table 1: Comparison of Feature-Pullback with imbalanced multi-modal learning methods, where **bold** and underline indicate the best and second-best performance respectively.

Method	Kinetics-Sound (Audio+Video)		CREMA-D (Audio+Video)		Sarcasm-Detection (Text+Image)		Twitter15 (Text+Image)	
	Acc	mAP	Acc	mAP	Acc	Mac-F1	Acc	Mac-F1
Traditional Multi-modal Fusion Methods								
Concat	64.55	71.30	63.31	68.41	82.86	82.40	70.11	63.86
Affine	64.24	69.31	66.26	71.93	82.40	81.88	72.03	59.92
Sum	64.90	71.03	63.44	69.08	82.94	82.47	73.00	66.61
Channel	63.51	68.66	66.13	71.70	-	-	-	-
Rebalanced Multi-modal Learning Methods								
MSES	64.71	70.63	61.56	66.83	84.18	83.60	71.84	66.55
OGR-GB	67.10	71.39	64.65	68.54	83.35	82.71	74.35	68.69
OGM-GE	66.06	71.44	66.94	71.73	83.23	82.66	74.92	68.74
Greedy	66.52	72.81	66.64	72.64	-	-	-	-
DOMFN	66.25	72.44	67.34	73.72	83.56	82.62	74.45	68.57
MSLR	65.91	71.96	65.46	71.38	84.23	83.69	72.52	64.39
PMR	66.56	71.93	66.59	70.30	83.60	82.49	74.25	68.60
AGM	66.02	72.52	67.07	73.58	84.02	83.44	74.83	69.11
MMPareto	70.13	-	<u>75.13</u>	-	-	-	-	-
AMSS	68.96	74.89	67.61	73.97	84.14	83.69	<u>75.89</u>	<u>69.81</u>
AMSS+	<u>72.25</u>	<u>79.13</u>	70.30	<u>76.14</u>	<u>84.35</u>	<u>83.77</u>	75.12	69.23
FP (Ours)	**76.78**	**82.67**	**76.93**	**82.35**	**86.12**	**84.92**	**76.75**	**73.33**

Backpropagating $q_m^{(r)}$ to the l-th layer yields

$$\delta_{m,l}^{(r)} = \frac{\partial q_m^{(r)}}{\partial s_{m,l}} = \sum_c \xi_{m,c}^{(r)} \delta_{m,l}^c, \qquad \mathbb{E}_\xi \left[\delta_{m,l}^{(r)} \delta_{m,l}^{(r)\top} \right] = \sum_c \delta_{m,l}^c \delta_{m,l}^{c\top}. \tag{12}$$

Equation (12) shows that the feature probe aggregates the sensitivities of all representation coordinates into a single vector–Jacobian product, whose second-order moment exactly recovers the target factor in Eq. (11). A finite number of probes provides a Monte Carlo estimate of this expectation. The probe gradients are used only to construct $S_{m,l}^f$ and are never accumulated into the task gradients of the model. The Kronecker structure further decomposes the inversion of a large matrix into the inversion of two small matrices. Using $(A \otimes S)^{-1} = A^{-1} \otimes S^{-1}$, together with the vectorization identity

$$(A^{-1} \otimes S^{-1}) \operatorname{vec}(G) = \operatorname{vec}\left(S^{-1} G A^{-1}\right),$$

the Feature-Pullback gradient for the l-th layer is finally given by

$$\widehat{\nabla}_{W_{m,l}} \mathcal{L} = \left(S_{m,l}^f + \lambda I\right)^{-1} \nabla_{W_{m,l}} \mathcal{L} \left(A_{m,l} + \lambda I\right)^{-1}. \tag{13}$$

Here, λ is a damping coefficient that ensures stable inversion of the factors. Equation (13) is directly applied to the joint-loss gradient produced by the joint loss. We instantiate the ideal pullback formulation with a layerwise Kronecker approximation, yielding the scalable Feature-Pullback algorithm used throughout our experiments. The detailed updating steps of the proposed FP method are presented in Algorithm 1.

4 Experiments

4.1 Dataset and Experiments Settings

Following (Yang et al., 2025), we adopt four benchmarks across different modalities: Kinetics-Sound (Arandjelovic & Zisserman, 2017), CREMA-D (Cao et al., 2014), Sarcasm Detection (Cai et al., 2019), and Twitter-15 (Yu & Jiang, 2019). Detailed dataset descriptions and experimental settings are provided in Appendix B.

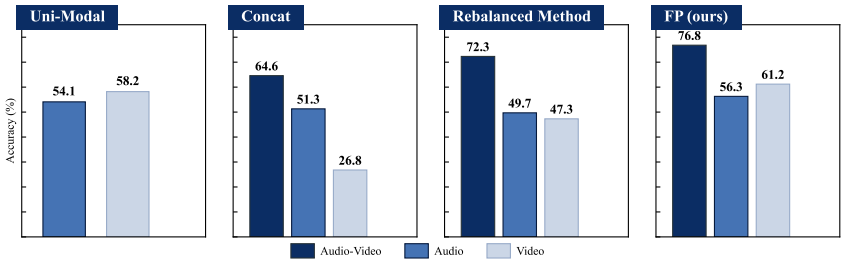


Figure 2: We compare the fusion and branch performance of unimodal references, Concat, representative rebalancing methods, and FP on Kinetics-Sound.

4.2 Comparison with Multi-Modal Learning Methods

Table 1 summarizes the results of Feature-Pullback, unimodal baselines, conventional fusion methods, and representative modality rebalancing methods (Fujimori et al., 2020; Yang et al., 2022) on four benchmarks. FP achieves the best performance on all eight evaluation metrics. Based on these results, we make the following observations:

(1) **Surpassing conventional forward fusion.** Compared with fusion methods including Concat, Affine (Perez et al., 2018), Sum, and Channel (Joze et al., 2020), FP achieves significant improvements on all four datasets. Specifically, it improves over the best conventional fusion method by 11.88 and 10.67 percentage points on Kinetics-Sound and CREMA-D, respectively. Different fusion methods only modify the forward aggregation form of modality representations, while their relative performance varies across datasets. FP further improves performance without changing the forward fusion architecture, indicating that the performance bottleneck of joint learning cannot be explained solely by how representations are aggregated. How task signals are realized as encoder-induced feature updates through different modality encoders also plays an important role.

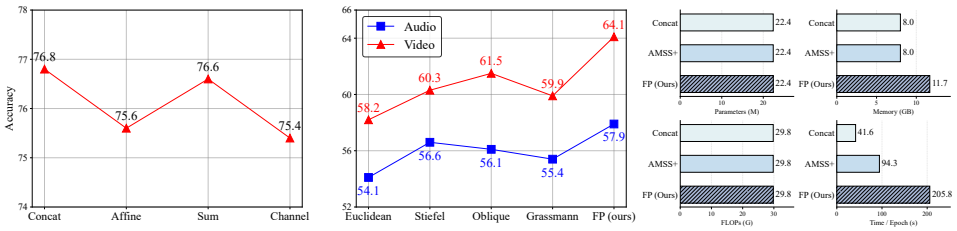
(2) **Superior to existing modality rebalancing strategies.** Compared with methods based on gradient modulation, learning speed coordination, modality contribution estimation, and parameter update selection, FP consistently achieves leading performance across all datasets and evaluation metrics, improving over the strongest existing results by 0.86 to 6.21 percentage points on different metrics. Existing methods mainly coordinate modality training based on optimization signals and their statistical characteristics, while FP further addresses the matrix-valued response induced by these signals through the parameterized encoder structures. The consistent improvements demonstrate that conditioning the response from the perspective of representation updates provides an effective improvement dimension beyond optimization-signal modulation.

(3) **Consistency across modalities and training configurations.** The four benchmarks cover four modalities, including audio, video, text, and image, and involve convolutional encoders represented by ResNet (He et al., 2016), Transformer (Vaswani et al., 2017) encoders represented by BERT (Devlin et al., 2019), as well as two types of optimizers, SGD and Adam. FP adopts the same algorithmic formulation across these configurations, without requiring modality-specific, encoder-specific, or optimizer-specific adjustment rules, and achieves consistent improvements under all settings. This indicates that Feature-Pullback does not address a training phenomenon specific to a particular network architecture or modality, but rather the general process of transforming task signals into representation updates in parameterized encoders.

4.3 Revisiting Multimodal Rebalancing

Existing rebalancing methods identify modality dominance relationships based on relative signals such as gradients, learning speed, or modality contributions, and accordingly adjust the updates of different branches. As shown in Figure 2, methods represented by AMSS+ recover part of the video performance, reduce the modality gap, and improve the fusion results. However, this convergence raises a question: does the reduction of the modality gap result from the recovery of the underperforming modality, or from the suppression of the other modality?

The independently trained results in Figure 2 provide a counterfactual reference: without the influence of joint training, the video encoder can achieve higher performance than the audio encoder.



(a) Fusion Strategy (b) FP On Uni-Modal (c) Computational Cost

Figure 3: Ablation studies and computational cost analysis of FP. (a) Performance with different fusion strategies on the KS dataset. (b) Performance of FP on unimodal audio and video models compared with predefined manifold alternatives. (c) Computational cost in terms of model parameters, GPU memory, FLOPs, and training time per epoch.

Existing rebalancing methods make the two branches closer to each other, but the audio performance becomes lower than its independent reference. This phenomenon indicates that modality performance convergence does not necessarily imply that each modality has been sufficiently learned. Therefore, modality imbalance should not be defined by whether the two branches become equal, but should instead focus on whether joint training causes a modality to deviate from its own achievable learning state.

This observation further connects to the representation update analysis in Section 3: the effective learning of each modality also depends on the representation updates induced by the joint signal through the matrix-valued response of the encoder. FP does not use unimodal performance as the training objective, but directly conditions this response process. On Kinetics-Sound, FP further improves the fusion performance to 76.8%, while enabling both the audio and video branches during joint training to exceed their independent training references of 54.3% and 58.2%, respectively. The same observation is obtained on CREMA-D (see Appendix Figure 4). These results validate the rebalancing perspective of this paper: rebalancing does not aim to make modalities converge by strengthening weaker modalities and weakening stronger ones, but instead corrects the transformation from the joint task signal to theinduced feature displacement of each modality, thereby improving the collaborative realization of the same joint objective across different modality encoders.

4.4 Ablation

Fusion Strategies. We integrate FP into Concat, Affine, Sum, and Channel fusion, respectively. FP achieves clear improvements over all baseline fusion methods and consistently maintains the best performance. Different fusion strategies alter the task signals returned by the joint objective to each modality, while FP consistently conditions the process by which these signals are realized as feature updates through the encoder. This demonstrates that its effectiveness does not depend on a specific forward fusion structure.

FP on Uni-Modal. We further apply FP independently to audio and video models, and simultaneously present the unimodal performance of predefined manifolds, including Euclidean, Stiefel, Oblique, and Grassmann, in Figure 3 (b). FP consistently improves both audio and video unimodal training and achieves the best results. This indicates that FP does not rely on the relative states among modalities and demonstrates the potential to be extended as a general training principle for parameterized representation learning.

5 Conclusion

We revisit multimodal imbalance through feature space geometry, showing that task signals are realized as feature updates through modality-specific response operators. Based on this view, we introduce Feature-Pullback, a scalable pullback conditioning framework for correcting this transformation. Experiments across audio, video, text, and image demonstrate consistent improvements across architectures, fusion strategies, and optimizers. Our results further suggest that rebalancing should not enforce modality equality, but improve how the joint objective is realized in each modality.

6 AI USE STATEMENT

In this work, we used generative AI tools solely for English translation of text written by the authors. For other tasks, we did not use generative AI tools.

All AI-assisted translations were reviewed and revised by the authors to ensure that they accurately reflected the intended meaning of the original text and did not alter the scientific content. We take responsibility for the final content of this work, including all text, claims, and other materials presented in the paper.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3.

Relja Arandjelovic and Andrew Zisserman. Look, Listen and Learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017. doi: 10.1109/ICCV.2017.73.

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, 2019. doi: 10.1109/TPAMI.2018.2798607.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2506–2515. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1239.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical Modal Rebalance for Multimodal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023. doi: 10.1109/CVPR52729.2023.01918.

Naotsuna Fujimori, Rei Endo, Yoshihiko Kawai, and Takahiro Mochizuki. Modality-Specific Learning Rate Control for Multimodal Classification. In *Proceedings of the 5th Asian Conference on Pattern Recognition*, volume 12047 of *Lecture Notes in Computer Science*, pp. 412–422. Springer, 2020. doi: 10.1007/978-3-030-41299-9_32.

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. doi: 10.1109/CVPR.2016.90.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality Competition: What Makes Joint Training of Multi-Modal Network Fail in Deep Learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 2022.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. MMTM: Multimodal Transfer Module for CNN Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13286–13296, 2020. doi: 10.1109/CVPR42600.2020.01330.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting Multi-Modal Model Performance with Adaptive Gradient Modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22157–22167, 2023. doi: 10.1109/ICCV51070.2023.02030.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-Factored Approximate Curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022. doi: 10.1109/CVPR52688.2022.00806.

Roger Penrose. A Generalized Inverse for Matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955. doi: 10.1017/S0305004100030401.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Weiyao Wang, Du Tran, and Matt Feiszli. What Makes Training Multi-Modal Classification Networks Hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12692–12702, 2020. doi: 10.1109/CVPR42600.2020.01271.

Yake Wei and Di Hu. MMPareto: Boosting Multimodal Learning with Innocent Unimodal Assistance. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52559–52572. PMLR, 2024.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and Overcoming the Greedy Nature of Learning in Multi-Modal Deep Neural Networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055. PMLR, 2022.

Yang Yang, Jingshuai Zhang, Fan Gao, Xiaoru Gao, and Hengshu Zhu. DOMFN: A Divergence-Orientated Multi-Modal Fusion Network for Resume Assessment. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1612–1620. ACM, 2022. doi: 10.1145/3503161.3548203.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating Multimodal Classification via Dynamically Learning Modality Gap. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62108–62122, 2024.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to Rebalance Multi-Modal Optimization by Adaptively Masking Subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4553–4566, 2025. doi: 10.1109/TPAMI.2025.3547417.

Yiqun Yao and Rada Mihalcea. Modality-Specific Learning Rates for Effective Multimodal Additive Late-Fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1824–1834. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.143.

Jianfei Yu and Jing Jiang. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 5408–5414, 2019. doi: 10.24963/IJCAI.2019/751.

Table 2: This table summarizes the main symbols used in the main text and appendix. Unless otherwise specified, vectors are treated as column vectors; the modality index is denoted by m, the training step index by t, and the network layer index by l.

Notation	Explanation	Notation	Explanation
M, m	Number of modalities and modality index.	x, y	Multimodal input and supervision label.
$E_m(\cdot;\theta_m)$	Encoder of the m-th modality.	θ_m, p_m	Parameters and parameter dimension of the m-th encoder.
f_m, d_m	Representation of the m-th modality and its dimension.	F, C	Fusion function and prediction head.
$\ell, \mathcal{L}$	Sample loss function and joint training objective.	Θ	All trainable parameters of the model.
h_m	Task signal returned from the joint objective to the m-th modality representation.	J_m	Jacobian of the m-th encoder with respect to its parameters.
g_m	Parameter gradient of the m-th encoder.	η	Learning rate.
$\Delta\theta_m, \Delta f_m$	One-step parameter change and the induced local feature change.	L_J	Local Lipschitz constant of the Jacobian.
$r(\Delta\theta)$	First-order Taylor residual.	$\mathcal{R}_m$	Modality representation response operator, $\mathcal{R}_m = J_m B_m J_m^\top$.
U_m, Λ_m	Orthogonal eigen-directions and response gain matrix of $\mathcal{R}_m$.	$u_{m,i}, \lambda_{m,i}, \alpha_{m,i}$	The i-th response direction, corresponding gain, and task signal projection.
M_m	Modality manifold transformation matrix.	$(\cdot)^\dagger$	Moore–Penrose pseudoinverse.
$P_{\mathrm{Range}(J)}$	Orthogonal projection onto the range of J.	z	Arbitrary parameter-space vector in the least-squares solution.

A Appendix

To improve the readability for the general audience, Table 2 lists all the notation used in this paper.

B Datasets and Settings

B.1 Dataset

To verify the generalization ability and robustness of the Feature-Pullback framework, we conduct extensive experiments on four widely used benchmark datasets covering Audio-Video and Image-Text interactions.

B.1.1 Audio-Video Tasks

Kinetics-Sound is a subset derived from Kinetics-400, containing 31 human action categories. The dataset consists of approximately 19k 10-second video clips. We follow the standard split, with 15k clips used for training and 1.9k clips for each of validation and testing. **CREMA-D** is an audio-visual emotion dataset recorded by 91 actors, containing 7,442 clips. We split it into 6,698 samples for training/validation and 744 samples for testing.

B.1.2 Image-Text Tasks

Sarcasm Detection is a multimodal sarcasm detection dataset collected from Twitter. It contains 24,635 image-text pairs, with 19,816, 2,410, and 2,409 samples for training, validation, and testing, respectively. **Twitter-15** is a multimodal sentiment analysis dataset containing three polarities: positive, neutral, and negative. It consists of 5,338 image-text pairs, split into 3,179 (training), 1,122 (validation), and 1,037 (testing) samples.

B.2 Experiments Settings

For Audio-Video tasks, we adopt ResNet-18 as the feature extractor for both modalities and use SGD with an initial learning rate of $1e^{-2}$. For Vision-Language tasks, we use ResNet-50 and BERT to extract visual and textual features, respectively, and adopt Adam with an initial learning rate of e^{-5}.

For the video modality, we uniformly sample 3 frames from each clip as input. For the audio modality, to adapt ResNet-18 to the input structure, we modify the number of input channels in the

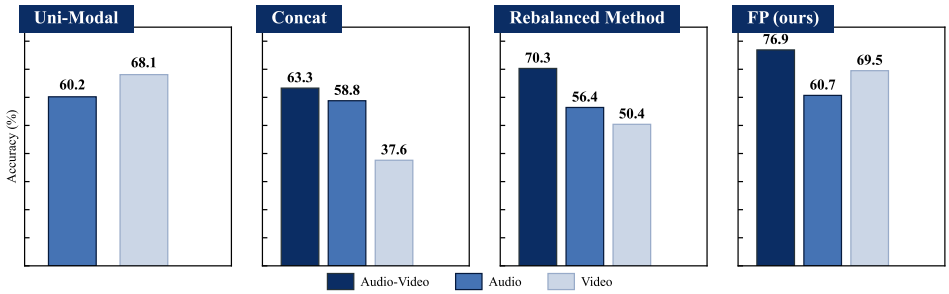


Figure 4: Fusion and modality branch performance of different training strategies on CREMA-D. The figure compares the unimodal references trained independently, Concat, the representative rebalancing method AMSS+, and FP. Fusion denotes the fusion accuracy of the joint model, while Audio and Video denote the accuracy of the corresponding modality branches in the joint model.

first convolutional layer from 3 to 1 to process spectrogram inputs. Kinetics-Sound and CREMA-D are trained from scratch, while Sarcasm Detection and Twitter-15 use the official PyTorch pretrained models. All experiments are conducted on a 4090 GPU.

C Modality Rebalancing Analysis on CREMA-D

To supplement the re-examination of modality imbalance in Section 4.3, this section further presents and analyzes the corresponding experimental results on CREMA-D. The results on CREMA-D exhibit a phenomenon consistent with Kinetics-Sounds: existing rebalancing methods can reduce the performance gap between two modality branches, but modality convergence does not necessarily indicate that both modalities are sufficiently learned. FP does not take unimodal performance as the training objective; instead, it conditions the transformation from joint task signals to feature updates, enabling improvements in both branches and further achieving higher fusion performance. This result further demonstrates that the observations on modality imbalance in the main text are not a phenomenon specific to Kinetics-Sounds.

D Supplementary Optimization Analysis

D.1 Applicability Scope of the Local Response Analysis

The main text is based on the single-step update of standard SGD and performs a first-order local analysis around the current parameters. This analysis aims to isolate the process by which the joint task signal forms feature updates through modality encoders: when $g_t = J_t^\top h_t$ and $\Delta\theta_t = -\eta_t g_t$, we have $\Delta f_t \approx -\eta_t J_t J_t^\top h_t$. Therefore, the modality representation response operator characterizes the local realization of the current joint task signal, rather than a global closed-form description of the complete optimization trajectory.

D.2 Representation Response Induced by Adam

For Adam, we still denote the current task gradient as $g_t = J_t^\top h_t$. Its first- and second-moment updates are

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t, \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)(g_t \odot g_t), \end{aligned} \tag{14}$$

The bias-corrected moment estimates are

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t}, \qquad \widehat{v}_t = \frac{v_t}{1-\beta_2^t}. \tag{15}$$

Algorithm 1 Feature-Pullback

Require: Model Θ, base optimizer $\mathcal{O}$, number of probes R

for each mini-batch **do**
  Compute modality features $\{f_m\}_{m=1}^{M}$ and loss $\mathcal{L}$
  for each modality m and encoder layer l **do**
    Compute $A_{m,l} \leftarrow \mathbb{E}[a_{m,l} a_{m,l}^{\top}]$ {equation 11}
    Draw isotropic feature probes $\{\xi_m^{(r)}\}_{r=1}^{R}$
    $q_m^{(r)} \leftarrow \langle \xi_m^{(r)}, f_m \rangle$
    $\delta_{m,l}^{(r)} \leftarrow \partial q_m^{(r)} / \partial s_{m,l}$ {equation 12}
    $S_{m,l}^{f} \leftarrow \frac{1}{R} \sum_{r=1}^{R} \delta_{m,l}^{(r)} (\delta_{m,l}^{(r)})^{\top}$ {equation 11–equation 12}
  end for
  Back-propagate $\mathcal{L}$ to obtain $\{G_{m,l}\}$, where $G_{m,l} = \nabla_{W_{m,l}} \mathcal{L}$
  for each modality m and encoder layer l **do**
    $\widetilde{G}_{m,l} \leftarrow (S_{m,l}^{f})^{\dagger} G_{m,l} A_{m,l}^{\dagger}$ {Undamped form of equation 13}
    $\nabla_{W_{m,l}} \mathcal{L} \leftarrow \widetilde{G}_{m,l}$
  end for
  Update Θ using $\mathcal{O}$
end for

Define

$$D_t = \operatorname{diag}\left((\sqrt{\widehat{v}_t} + \epsilon)^{-1} \right), \qquad c_t = \frac{1 - \beta_1}{1 - \beta_1^t}. \tag{16}$$

Then, conditioned on the current moment estimates, the component of the current gradient g_t in $\widehat{m}_t$ is $c_t g_t$, and its induced local feature update is

$$\Delta f_t^{\mathrm{cur}} \approx -\eta_t c_t J_t D_t J_t^{\top} h_t. \tag{17}$$

Therefore, the conditional local response induced by Adam for the current task signal is $c_t J_t D_t J_t^{\top}$. Since the moment estimates simultaneously depend on the current gradient and the existing optimizer states, Equation (17) is only used to characterize the effect of the current task signal and does not constitute an equivalent linear expression of the complete cross-step dynamics of Adam.

D.3 Combination of FP and Adam

When the base optimizer is Adam, FP first conditions the parameter gradient generated by the current joint loss:

$$\bar{g}_t = P_t^{\mathrm{FP}} g_t, \tag{18}$$

where P_t^{FP} denotes the Feature-Pullback conditioning operator obtained from the layerwise approximation in the main text. Subsequently, Adam updates its first- and second-moment estimates using $\bar{g}_t$. Let D_t^{FP} denote the diagonal preconditioning matrix formed by the conditioned gradient and its optimizer states. Then, the local feature update induced by the current gradient component is

$$\Delta f_t^{\mathrm{cur,FP+Adam}} \approx -\eta_t c_t J_t D_t^{\mathrm{FP}} P_t^{\mathrm{FP}} J_t^{\top} h_t. \tag{19}$$

Equation (19) shows that FP first pulls back the update criterion in feature space into parameter space to condition the current joint task gradient, and Adam subsequently performs adaptive updates according to its optimizer states. This combination indicates that FP can be integrated with different base optimizers, but it does not imply that its composition with Adam is equivalent to the ideal projection derived from the local SGD analysis in the main text.

Table 3: Comparison of performance and training overhead with different numbers of feature probes R on Kinetics-Sounds. $R = 2$ is the default setting.

Probes R	Acc (%)	mAP (%)	Time/Epoch (s)
1	76.68	82.59	**182.9**
2	**76.78**	82.67	205.8
4	76.68	82.81	247.3
8	76.72	**82.88**	331.9

E Ablation

E.1 Sensitivity Analysis of the Number of Feature Probes

We analyze the sensitivity of FP to the number of feature probes R on Kinetics-Sounds. When $R \geq 2$, the model performance remains largely stable, while the training time increases with R. Therefore, we set $R = 2$ as the default configuration to balance estimation quality and training cost.

E.2 Computational Cost

Figure 3 (c) compares the computational overhead of FP with Concat and AMSS+. FP has the same number of parameters and inference FLOPs as the two baselines, indicating that its performance gains do not stem from increased model capacity or additional forward computation, but rather from explicit regulation of the representation update process. The additional computation introduced by FP occurs only during training, mainly due to the estimation and correction of matrix-valued representation responses, thereby resulting in additional GPU memory consumption and longer training time per epoch. Notably, FP does not alter the forward inference process of the model. After training, the model still follows the same encoding, fusion, and prediction pipeline as the original baseline. Therefore, the effect of FP is confined to the optimization stage and does not rely on additional inference modules or specific forward architectures. Together with the consistent improvements observed under different fusion strategies, these results further indicate that FP exhibits good structural compatibility and methodological generalizability.

F Limitation

The main limitation of FP arises from the fact that explicitly constructing and solving the full pullback geometry requires storing and processing matrices at the scale of the model parameters, resulting in prohibitive computational, memory, and time overheads for modern encoders. Therefore, scalable FP employs a finite number of feature probes together with layer-wise Kronecker-structured approximations. Although experiments show that model performance is not sensitive to the number of probes, this approximation cannot preserve all cross-layer dependencies in the full Jacobian geometry and introduces additional time and memory overhead during training, while leaving the model parameter count and inference computation unchanged. Future work will explore geometric approximations that balance fidelity to the response structure with training efficiency.