BALANCING MULTIMODAL LEARNING VIA FUNCTIONAL PROGRESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning often suffers from modality imbalance, where the joint optimization process is dominated by a single modality. Existing methods typically estimate modality imbalance from score disparities derived from prediction uncertainty or optimization statistics. However, due to distinct prediction uncertainty and learning dynamics across modalities, direct comparison of such scores may misinterpret intrinsic modality differences as progress gaps, leading to biased imbalance estimation. In this paper, we propose **F**unction-Space **G**uided **M**ultimodal **O**ptimization **(FGMO)**, which leverages a function-space progress signal to assess modality-wise optimization progress and coordinate optimization across modalities to alleviate modality imbalance. Specifically, we introduce Functional Progress Estimation (FPE) to measure each modality's update-induced function-space response and calibrate it against a loss-aligned unimodal reference, producing a comparable progress signal. Based on this signal, Functional Response Control (FRC) redistributes modality-level function-space budgets and realizes the target responses through tensor-wise learning-rate adjustment. Theoretical analysis establishes a one-step target-contraction property of FRC under bounded controller-state mismatch, and extensive experiments demonstrate the effectiveness of FGMO across multiple multimodal benchmarks.

1 INTRODUCTION

Multimodal learning (Baltrušaitis et al., 2019) aims to integrate information from multiple modalities, such as vision, language, and audio. Yet, incorporating more modalities does not always lead to better performance (Wang et al., 2020). Previous studies have shown that different modalities often learn and converge at different rates during joint training (Yang et al., 2024). As a result, the easier-to-learn modality may dominate the training process, while the other modalities remain under-optimized (Hua et al., 2024; Huang et al., 2025). This phenomenon is commonly referred to as modality imbalance (Yang et al., 2024; Jiang et al., 2025).

To alleviate modality imbalance, existing methods typically estimate modality-wise imbalance scores from prediction uncertainty or optimization statistics and adjust the optimization accordingly (Peng et al., 2022; Fan et al., 2023; Wei et al., 2025; Gao et al., 2026). Prediction-based methods derive such scores from signals such as confidence or prediction variance (Peng et al., 2022; Fan et al., 2023; Wei et al., 2025), while optimization-based methods use statistics such as gradient or update magnitudes (Wei & Hu, 2024; Gao et al., 2026). Typically, these scores are compared directly across modalities to determine which modality is under-optimized. However, different modalities naturally exhibit distinct prediction uncertainty and learning dynamics, making cross-modal score gaps poorly calibrated to the actual rebalancing demand. Consequently, the magnitude of a score disparity may not accurately indicate which modality is relatively under-optimized or how strongly it should be adjusted, potentially leading to inappropriate rebalancing.

To address this limitation, we assess each modality's optimization during joint training relative to its own unimodal training, rather than directly comparing absolute scores across modalities. Specifically, we assess modality optimization from a function-space perspective (Milsom et al., 2025), which directly captures the model-output change induced by the current update while being independent of heterogeneous parameterizations. We then normalize each modality's current functional response by its loss-aligned unimodal reference to obtain a function-space progress signal termed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

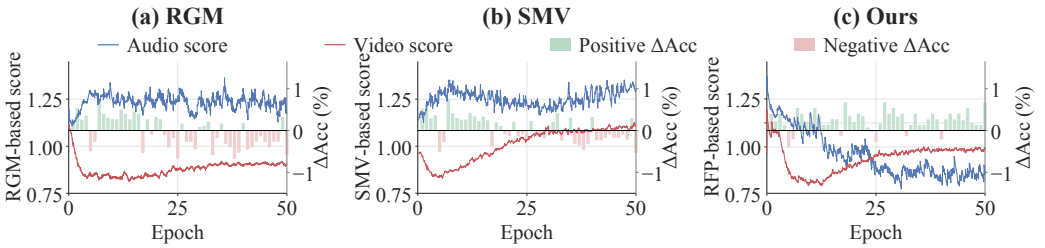


Figure 1: Controlled intervention experiments on CREMA-D. We save 50 checkpoints along a vanilla training trajectory and, from each checkpoint, construct a vanilla continuation and three intervention branches. All branches start from the same model and optimizer states and continue for 10 epochs, with ΔAcc measured relative to the vanilla continuation. Across panels, blue/red curves denote audio/video scores and green/red bars denote positive/negative ΔAcc.

Relative Functional Progress (RFP), which accounts for modality-specific learning dynamics and provides a more comparable signal for modality rebalancing.

To evaluate the rebalancing reliability of RFP, we conduct controlled intervention experiments on CREMA-D and compare RFP with the modality-level rebalancing scores derived from RGM (Gao et al., 2026) and SMV (Wei et al., 2024), representing optimization-based and prediction-based signals, respectively. To isolate the quality of these signals, we retain only their modality-level scores and convert all three signals into intervention coefficients using the same mapping rule. Specifically, we select several checkpoints along the same vanilla training trajectory and, from each checkpoint, construct four continuation branches: one vanilla branch that preserves the original optimization process and three rebalancing branches guided by the RGM-derived score, SMV-derived score, and RFP, respectively. We measure the intervention benefit of each signal at the current training stage by comparing the final accuracy of its guided branch with that of the corresponding vanilla continuation. As illustrated in Figure 1, while all three signals show a consistent audio-over-video ordering with generally beneficial interventions in the early stage of training, the RGM- and SMV-derived scores retain clear audio–video gaps in the later stage even as their interventions become harmful. This suggests that these cross-modal scores can be biased by modality-specific differences and consequently misguide rebalancing. By contrast, the RFP gap gradually narrows and eventually reverses its ordering, while RFP-guided interventions remain beneficial at most checkpoints, suggesting that RFP provides a more comparable measure of optimization progress across modalities.

To this end, we propose **F**unction-Space **G**uided **M**ultimodal **O**ptimization **(FGMO)**, which coordinates modality optimization using RFP, a function-space progress signal. Specifically, we introduce Functional Progress Estimation (FPE) to recover unit-learning-rate updates, measure their tensor-wise responses in function space, and derive RFP. Using the resulting RFP, Functional Response Control (FRC) further allocates modality-level response budgets toward slower-progress modalities under KL regularization. In turn, the assigned budgets are converted into tensor-wise learning rates according to the reference profile and current functional responses, which modulate the optimizer updates. Together, these components form a feedback loop that continuously reassesses and regulates modality progress throughout training. Our contributions are as follows:

- **We rethink modality imbalance in function space and introduce Relative Functional Progress (RFP)**, which provides a more comparable signal for modality rebalancing.
- **We propose Functional Progress Estimation (FPE) and Functional Response Control (FRC) to form a closed-loop optimization framework,** which dynamically coordinates modality optimization to alleviate modality imbalance.
- **We provide a theoretical analysis of the proposed closed-loop controller** and establish its one-step target-contraction property for reducing RFP discrepancies. Extensive experiments on six multimodal benchmarks demonstrate the effectiveness of our method.

2 Related Work

Imbalanced Multimodal Learning. Compared to unimodal learning, multimodal learning (Baltrušaitis et al., 2019) can integrate multi-source information to improve performance. However, traditional fusion methods (Baltrušaitis et al., 2019; Nie et al., 2021) often suffer from modality imbalance, where the strong modality rapidly dominates the training process, leaving the weak modality under-optimized (Wang et al., 2020; Huang et al., 2022; Peng et al., 2022). To alleviate this issue, one line of work estimates modality status from prediction-related signals. OGM (Peng et al., 2022) estimates modality imbalance from prediction contributions, while methods such as PMR (Fan et al., 2023) and AGM (Li et al., 2023) introduce modality-specific classifiers or predictions to characterize the learning state of each modality. More recently, ARL (Wei et al., 2025) and AUG (Jiang et al., 2025) calibrate modality optimization based on prediction variance and bias and modality classification ability. Another line of work characterizes modality status from optimization-related signals. G-Blend (Wang et al., 2020) and Greedy (Wu et al., 2022) adjust modality optimization according to generalization dynamics. Recent methods such as AMSS+ (Yang et al., 2025) and RGM (Gao et al., 2026) exploit fine-grained parameter information or cumulative gradient energy to regulate modality optimization. Despite their different forms, both categories estimate modality imbalance by directly comparing modality-wise scores. However, different modalities naturally exhibit distinct prediction uncertainty and learning dynamics, making these scores not inherently comparable across modalities. Consequently, score disparities may reflect intrinsic modality differences rather than actual differences in optimization progress.

Function-Space Methods. To more faithfully characterize how neural networks change during learning, early studies began to examine optimization from a function-space perspective. HCGD compared parameter- and function-space distances during optimization (Benjamin et al., 2019). Subsequent works such as S-FSVI regularized model outputs in function space to preserve previously learned knowledge (Rudner et al., 2022). More recently, SFR introduced a function-space parameterization for sequential learning (Scannell et al., 2024), while FLeRM proposed an efficient first-order estimator to quantify function-space changes induced by parameter updates (Milsom et al., 2025). Despite these advances, existing function-space methods mainly focus on single-network or unimodal learning scenarios. How to leverage function space to measure and balance modality competition in joint multimodal training remains under-explored.

3 Method

In this section, we present Function-Space Guided Multimodal Optimization (FGMO), whose overall framework is illustrated in Figure 2. Specifically, we introduce FPE that measures the function-space response and calibrates it against a loss-aligned unimodal reference to obtain RFP. Based on RFP, FRC further redistributes the modality-level function-space response budget and realizes the assigned budget. Together, FPE and FRC form a closed-loop optimization framework that dynamically coordinates modality optimization toward balanced RFP across modalities.

3.1 Multimodal Learning

We consider a multimodal classification task involving M distinct modalities. Let $\mathcal{D} = \{(\mathbf{x}_i^1, \ldots, \mathbf{x}_i^M, \mathbf{y}_i)\}_{i=1}^N$ denote a training set of N i.i.d. samples, where $\mathbf{x}_i^m \in \mathcal{X}^m$ represents the input of the m-th modality and $\mathbf{y}_i \in \{0,1\}^C$ is the corresponding one-hot label over C classes. For each modality m, a dedicated encoder $\phi_m(\cdot;\boldsymbol{\theta}_m)$ and classifier $g_m(\cdot;\boldsymbol{\psi}_m)$ produce the modality-specific logits $\mathbf{z}_i^m = g_m(\phi_m(\mathbf{x}_i^m;\boldsymbol{\theta}_m);\boldsymbol{\psi}_m)$, yielding the unimodal prediction $\mathbf{p}_i^m = \mathrm{softmax}(\mathbf{z}_i^m)$. During joint training, the modality-specific logits are averaged to obtain the joint prediction $\mathbf{p}_i = \mathrm{softmax}\left(\frac{1}{M}\sum_{m=1}^M \mathbf{z}_i^m\right)$. Let $\boldsymbol{\Theta} = \{(\boldsymbol{\theta}_m, \boldsymbol{\psi}_m)\}_{m=1}^M$ collect all trainable encoder and classifier parameters. The entire network is optimized with an overall objective combining the multimodal joint loss and the unimodal losses from individual branches:

$$\mathcal{L}(\boldsymbol{\Theta}) = -\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_i^\top \log(\mathbf{p}_i) + \sum_{m=1}^{M}\mathbf{y}_i^\top \log(\mathbf{p}_i^m)\right). \tag{1}$$

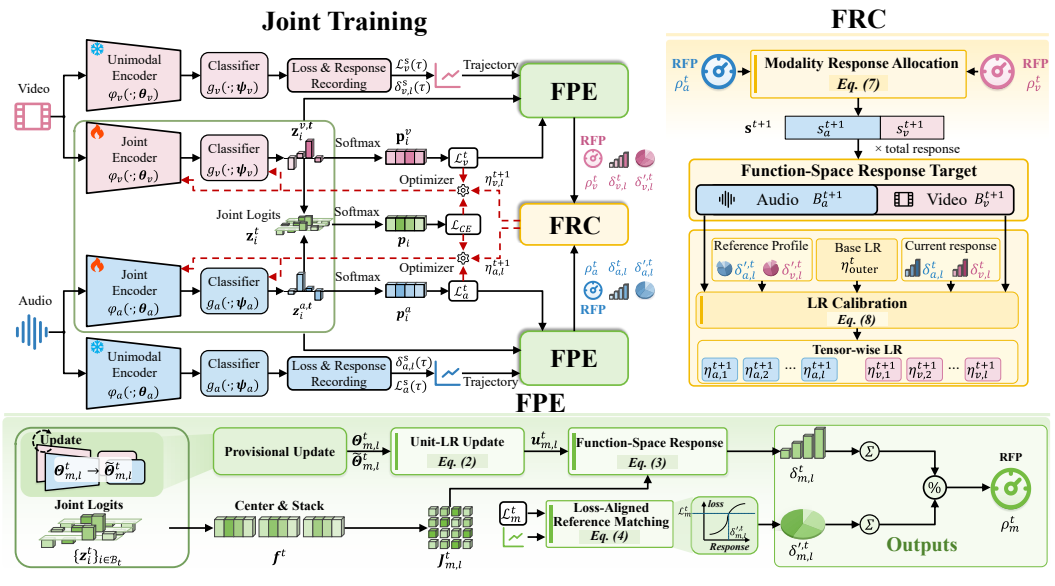


Figure 2: The overall framework of Function-Space Guided Multimodal Optimization (FGMO). FPE estimates Relative Functional Progress (RFP) from update-induced function-space responses and loss-aligned unimodal references, while FRC uses RFP feedback to perform modality-level response allocation and tensor-wise learning-rate adjustment.

3.2 Functional Progress Estimation

Based on Eq. (1), at each optimization update, the optimizer produces parameter updates from the current mini-batch. Due to heterogeneous parameterizations and learning dynamics, these updates cannot directly reflect the optimization progress of each modality. Hence, FGMO characterizes modality-wise optimization progress by estimating the function-space response and calibrating it against a loss-aligned unimodal reference to obtain the function-space progress signal.

Specifically, we perform progress estimation and response adjustment every K optimization updates and refer to each such invocation as an adjustment event and index it by t. At adjustment event t, let $\mathcal{B}_t$ denote the current mini-batch and $\mathcal{T}_m$ the set of parameter tensors belonging to modality m. For each $l \in \mathcal{T}_m$, we first perform a provisional optimizer update on $\mathcal{B}_t$ and then restore the parameters to their pre-update values. Since we dynamically adjust tensor-wise learning rates during training, directly using the parameter change after the optimizer update would make the measured response depend on the current learning-rate scale. We therefore recover the unit-learning-rate update as:

$$\mathbf{u}_{m,l}^{t} = \frac{\widetilde{\mathbf{\Theta}}_{m,l}^{t} - \mathbf{\Theta}_{m,l}^{t}}{\eta_{m,l}^{t}}, \tag{2}$$

where $\mathbf{\Theta}_{m,l}^{t}$ and $\widetilde{\mathbf{\Theta}}_{m,l}^{t}$ denote the parameters of the l-th tensor for the m-th modality before and after the provisional update, $\eta_{m,l}^{t}$ is the current learning rate for tensor l of modality m, and $\mathbf{u}_{m,l}^{t}$ is the corresponding unit-learning-rate update.

Inspired by an efficient function-space estimator (Milsom et al., 2025), we characterize the functional effect of each tensor update through its induced first-order change in the model output. For the current mini-batch of n_b samples, let $\mathbf{z}_i^t = \frac{1}{M}\sum_{m=1}^{M} \mathbf{z}_i^{m,t}$ denote the joint logits of sample i, obtained using the same logit-level fusion as in inference. We subtract the class-wise mean from each logit vector and stack the resulting centered logits as $\mathbf{f}^t \in \mathbb{R}^D$, where $D = n_b C$. Let $\mathbf{J}_{m,l}^{t} = \partial \mathbf{f}^t / \partial \mathbf{\Theta}_{m,l}^{t}$ denote the Jacobian with respect to tensor l of modality m. Under a first-order approximation, we define the tensor-wise function-space response as:

$$\delta_{m,l}^{t} = \frac{1}{\sqrt{D}} \left\| \mathbf{J}_{m,l}^{t} \mathbf{u}_{m,l}^{t} \right\|_2, \tag{3}$$

which measures the normalized functional change induced by the unit-learning-rate update. In practice, we avoid explicitly constructing $\mathbf{J}_{m,l}^t$ and estimate $\delta_{m,l}^t$ using Gaussian random projection with a low-variance Kronecker-factored EMA estimator (Milsom et al., 2025). The derivation and implementation details are provided in Appendices B.1 and B.2.

To construct the unimodal reference, we independently train each modality and evaluate the unimodal loss $\mathcal{L}_m^{\mathrm{s}}(\tau)$ and tensor-wise response $\delta_{m,l}^{\mathrm{s}}(\tau)$ every K optimization updates, where τ indexes the corresponding response estimation event. Meanwhile, during joint training, we align the joint and unimodal optimization states by loss rather than epoch:

$$\tau_m^t = \arg\min_{\tau} \left| \log \mathcal{L}_m^t - \log \mathcal{L}_m^{\mathrm{s}}(\tau) \right|, \tag{4}$$

where $\mathcal{L}_m^t$ denotes the modality-specific loss during joint training, and $\delta_{m,l}^{\prime,t} = \delta_{m,l}^{\mathrm{s}}(\tau_m^t)$ denotes the matched unimodal response. Using the current and reference responses, we obtain the Relative Functional Progress (RFP), denoted by ρ_m^t, of modality m as:

$$\rho_m^t = \frac{\sum_{l \in \mathcal{T}_m} \delta_{m,l}^t + \epsilon}{\sum_{l \in \mathcal{T}_m} \delta_{m,l}^{\prime,t} + \epsilon}, \tag{5}$$

where $\epsilon > 0$ is a small constant for numerical stability. The ratio measures the overall function-space response during joint training relative to the modality's own reference response, reducing the influence of modality-specific response scales and enabling comparison through relative rather than absolute progress. The resulting RFP is then passed to FRC as the feedback signal for subsequent function-space response allocation.

3.3 Functional Response Control

Through Eq. (5) in FPE, we obtain the RFP ρ_m^t of each modality at the current optimization state. To use this feedback to coordinate subsequent optimization, we propose Functional Response Control (FRC), which dynamically redistributes modality-level function-space responses according to RFP and further realizes the assigned response targets through tensor-wise learning-rate adjustment.

Specifically, at adjustment event t, we use $\mathbf{s}^t = (s_1^t, \ldots, s_M^t) \in \Delta_M$ to represent the current modality response shares, where s_m^t denotes the response fraction assigned to modality m. Since a smaller RFP indicates slower functional progress relative to its unimodal reference, directly favoring modalities with smaller RFP may over-concentrate the response and ignore the previous allocation state. We therefore use $\log \rho_m^t$ as the progress cost and constrain the change from the previous allocation using the Kullback–Leibler divergence:

$$\mathbf{s}^{t+1} = \arg\min_{\mathbf{s} \in \Delta_M} \left[\sum_{m=1}^{M} s_m \log \rho_m^t + \frac{1}{\lambda} D_{\mathrm{KL}}(\mathbf{s} \| \mathbf{s}^t) \right]. \tag{6}$$

where $\lambda > 0$ controls the response strength of FRC to the current RFP. The first term adjusts modality responses according to the current relative functional progress, while the KL term preserves the allocation state across consecutive adjustment events. The objective admits the closed-form solution:

$$s_m^{t+1} = \frac{s_m^t (\rho_m^t)^{-\lambda}}{\sum_{j=1}^{M} s_j^t (\rho_j^t)^{-\lambda}}. \tag{7}$$

This update adapts the modality response allocation to the current RFP as relative progress evolves during training. We then convert the updated response share into the modality-level function-space response target $B_m^{t+1} = s_m^{t+1} \sum_{k=1}^{M} \sum_{l \in \mathcal{T}_k} \delta_{k,l}^t$, where B_m^{t+1} denotes the function-space response assigned to modality m at the next adjustment.

Accordingly, to realize the modality-level response target across parameter tensors, we use the loss-aligned unimodal response matched by Eq. (4) to define the loss-aligned reference profile of modality m, whose entry for tensor l is $\delta_{m,l}^{\prime,t} / \sum_{j \in \mathcal{T}_m} \delta_{m,j}^{\prime,t}$. Since $\delta_{m,l}^t$ characterizes the function-space response under a unit learning rate, we scale the learning rate inversely with the current response to realize the assigned tensor-level target:

$$\eta_{m,l}^{t+1} = \eta_{\mathrm{outer}}^t \frac{B_m^{t+1} \delta_{m,l}^{\prime,t}}{(\delta_{m,l}^t + \epsilon) \sum_{j \in \mathcal{T}_m} \delta_{m,j}^{\prime,t}}, \qquad l \in \mathcal{T}_m, \tag{8}$$

where η_{outer}^t denotes the base learning rate. Finally, the tensor-wise learning rates produced by FRC rescale the parameter updates generated by the optimizer, allowing each tensor to be updated along the optimizer direction with its assigned step size. In this way, FRC maps the modality-level response target to tensor-wise learning rates while accounting for heterogeneous functional responses across tensors. The complete optimization procedure of FGMO is summarized in Algorithm 1 in Appendix A.

3.4 Theoretical Analysis

We analyze FRC via one-step target contraction to characterize whether its response target reduces the current modality optimization discrepancy. For clarity, we consider two modalities with input RFP ratio $Q^t = \rho_1^t/\rho_2^t$, where $\log Q^t = 0$ indicates balanced relative functional progress, and let $Q_{\text{tar}}^{t+1|t}$ denote the FRC-assigned target ratio.

Theorem 1 (One-Step Target Contraction, Informal). The target ratio satisfies

$$\log Q_{\text{tar}}^{t+1|t} = (1-\lambda)\log Q^t + d_t, \tag{9}$$

where λ is the controller gain and d_t is the controller-state mismatch. If $0 < \lambda < 2$ and $|d_t| < (1-|1-\lambda|)|\log Q^t|$, then

$$|\log Q_{\text{tar}}^{t+1|t}| < |\log Q^t|. \tag{10}$$

Theorem 1 shows that, under bounded controller-state mismatch and $0 < \lambda < 2$, the FRC target strictly contracts toward balance. This theoretically characterizes FRC as a negative-feedback controller over RFP, supporting the use of RFP feedback to coordinate modality optimization. Detailed proofs are provided in Appendix B.6.

4 Experiments

4.1 Experimental Setup

Dataset: We conduct experiments on six widely used multimodal datasets, i.e., CREMA-D (Cao et al., 2014), KSounds (Arandjelović & Zisserman, 2017), VGGSound (Chen et al., 2020), Twitter (Yu & Jiang, 2019), Sarcasm (Cai et al., 2019), and NVGesture (Molchanov et al., 2016), covering audio-video, image-text, and trimodal settings. CREMA-D is an audio-visual dataset for emotion recognition, containing 7,442 video clips from 91 actors. KSounds is an audio-visual dataset for human action recognition with 31 action classes. VGGSound is a large-scale audio-visual dataset with nearly 200K video clips. Twitter is an image-text dataset for emotion recognition with three classes. Sarcasm is designed for sarcasm detection and contains 24,635 image-text pairs. NVGesture is a trimodal dataset for gesture recognition, containing 1,532 samples from RGB, optical flow (OF), and depth modalities.

Baselines: We compare FGMO with traditional multimodal fusion methods, including Concat (Baltrušaitis et al., 2019), Affine (Perez et al., 2018), and ML-LSTM (Nie et al., 2021), and modality rebalancing methods, including G-Blend (Wang et al., 2020), MSLR (Yao & Mihalcea, 2022), OGM (Peng et al., 2022), PMR (Fan et al., 2023), AGM (Li et al., 2023), MMPareto (Wei & Hu, 2024), SMV (Wei et al., 2024), MLA (Zhang et al., 2024), DI-MML (Fan et al., 2024), ReconBoost (Hua et al., 2024), LFM (Yang et al., 2024), AMSS+ (Yang et al., 2025), InfoReg (Huang et al., 2025), ARL (Wei et al., 2025), AUG (Jiang et al., 2025), DecAlign (Qian et al., 2026) and RGM (Gao et al., 2026).

Evaluation Protocols: Following established evaluation protocols in prior multimodal studies (Yu & Jiang, 2019; Cai et al., 2019; Peng et al., 2022), we adopt accuracy, mean average precision (mAP), and Macro-F1 as evaluation metrics. Accuracy measures the proportion of correctly classified samples. mAP reports the mean average precision over all classes, while Macro-F1 computes the average F1 score across classes.

Implementation Details: Following OGM (Peng et al., 2022), we employ ResNet-18 (He et al., 2016) as the backbone for both audio and visual encoders on CREMA-D, KSounds, and VGGSound. All backbone parameters are randomly initialized. For NVGesture, we employ I3D (Carreira & Zisserman, 2017) as the encoder for RGB, OF, and depth modalities and initialize it with pretrained

Table 1: Comparison with SOTA multimodal learning methods. The best results are highlighted in **bold**. The underline denotes the second-best performance. Gray-shaded results indicate that multimodal performance is lower than that of the best unimodal approach.

Method	CREMA-D		KSounds		VGGSound		Twitter		Sarcasm		NVGesture	
	Acc.	mAP	Acc.	mAP	Acc.	mAP	Acc.	F1	Acc.	F1	Acc.	F1
Unimodal-1	.6317	.5879	.5412	.5669	.4655	.4701	.5863	.4333	.7181	.7073	.7822	.7833
Unimodal-2	.4583	.6861	.5562	.5837	.3494	.3478	.7367	.6849	.8136	.8056	.7863	.7865
Unimodal-3	–	–	–	–	–	–	–	–	–	–	.8154	.8183
Concat	.6361	.6841	.6455	.7130	.5116	.5352	.7011	.6386	.8286	.8240	.8237	.8270
Affine	.6626	.7193	.6424	.6931	.5001	.5155	.7203	.5992	.8240	.8188	.8278	.8281
ML-LSTM	.6290	.6473	.6394	.6902	.4966	.5139	.7068	.6564	.8277	.8205	.8320	.8330
G-Blend	.6465	.7392	.6722	.7274	.5086	.5555	.7309	.6799	.8286	.8215	.8299	.8305
MSLR	.6868	.7412	.6756	.7282	.4987	.5415	.7232	.6382	.8439	.8378	.8237	.8284
OGM	.6612	.7372	.6582	.7159	.4829	.4978	.7058	.6435	.8360	.8293	–	–
PMR	.6659	.7058	.6675	.7274	.4647	.4866	.7357	.6636	.8310	.8256	–	–
AGM	.6733	.7807	.6791	.7388	.4711	.5198	.7261	.6502	.8306	.8293	.8279	.8284
MMPareto	.7487	.8535	.7000	.7850	.5125	.5473	.7358	.6729	.8348	.8284	.8382	.8424
SMV	.7872	.8417	.6900	.7426	.5031	.5362	.7428	.6817	.8418	.8368	.8352	.8341
MLA	.7943	.8572	.7004	.7945	.5165	.5473	.7352	.6713	.8426	.8348	.8340	.8372
DI-MML	.8158	.8592	.7203	.7426	.5173	.5479	.7248	.6686	.8411	.8315	–	–
ReconBoost	.7557	.8140	.6855	.7662	.5097	.5387	.7442	.6832	.8437	.8317	.8386	.8434
LFM	.8362	.9006	.7253	.7897	.5274	.5694	.7501	.7057	.8497	.8457	.8436	.8468
AMSS+	.7030	.7641	.7255	.7913	–	–	.7558	.6981	.8435	.8377	.8464	.8494
InfoReg	.7498	.8603	.7189	.7986	.5289	.5712	.7403	.6847	.8479	.8396	–	–
ARL	.8379	.8927	.7198	.7826	.5234	.5583	.7386	.6786	–	–	.8375	.8432
AUG	.8515	.9103	.7263	.7901	.5301	.5826	.7512	.6962	.8510	.8458	.8501	.8533
DecAlign	.8426	.9027	.7298	.7926	.5268	.5677	.7396	.6786	–	–	.8397	.8429
RGM	.8327	.8993	.7309	.8088	–	–	.7512	.7035	.8449	.8397	.8406	.8470
FGMO	**.8589** ±0.0037	**.9126** ±0.0012	**.7565** ±0.0033	**.8210** ±0.0058	**.5341** ±0.0023	**.5864** ±0.0024	**.7628** ±0.0046	**.7131** ±0.0042	**.8547** ±0.0021	**.8486** ±0.0012	**.8548** ±0.0034	**.8582** ±0.0028

weights. We use SGD with a momentum of 0.9 and a weight decay of 1×10^{-4} for these datasets. The initial learning rates are 7.5×10^{-3} for CREMA-D, 1×10^{-2} for KSounds and VGGSound, and 8×10^{-3} for NVGesture. For Twitter and Sarcasm, we adopt BERT (Devlin et al., 2019) as the text encoder and ResNet-50 (He et al., 2016) as the image encoder. We use Adam (Kingma & Ba, 2015) with an initial learning rate of 2×10^{-5}. The batch sizes are set to 64, 32, 16, 2, 32, and 64 for CREMA-D, KSounds, VGGSound, NVGesture, Twitter, and Sarcasm, respectively. For fair comparison, all baselines use the same backbone and initialization strategy. All models are implemented with PyTorch (Paszke et al., 2019) and experiments are conducted on NVIDIA GeForce RTX 4090 GPUs.

4.2 Main Results

Comparison with Multimodal Baselines: Table 1 reports the results on all six datasets. Unimodal-1/2 denote audio/video for CREMA-D, KSounds, and VGGSound, and image/text for Twitter and Sarcasm, respectively. Unimodal-1/2/3 denote RGB/OF/Depth for NVGesture, respectively. Gray-shaded entries indicate multimodal results below the best unimodal performance on the corresponding metric. Simple fusion methods, including Concat (Baltrušaitis et al., 2019) and Affine (Perez et al., 2018), underperform the best unimodal model on Twitter, showing that combining modalities alone does not guarantee better performance. Existing modality rebalancing methods improve performance in many settings, but their relative performance varies across datasets and metrics. In contrast, FGMO achieves the best results on both reported metrics across all six datasets, outperforming competitive methods such as AUG (Jiang et al., 2025) and RGM (Gao et al., 2026). Notably, on KSounds, FGMO achieves an accuracy of 0.7565 and an mAP of 0.8210, exceeding the strongest baseline, RGM, by 2.56% and 1.22%, respectively. These results support the effectiveness of our reference-calibrated function-space rebalancing across different modality combinations.

4.3 Ablation Study

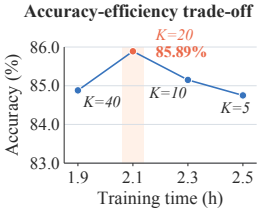
We conduct ablation studies on Sarcasm and KSounds to evaluate four key designs in FGMO: function-space response measurement, the relative progress metric RFP, loss-aligned reference matching, and tensor-wise response allocation. The results are reported in Table 2.

As shown in Table 2, removing any of the four designs consistently reduces multimodal performance. Replacing function-space responses with parameter-space update magnitudes reduces mul-

Table 2: Ablation study on Sarcasm and KSounds. We ablate function-space measurement, relative progress normalization (RFP), loss-aligned reference matching, and tensor-wise allocation.

Variant	Space	Align.	Metric	Granularity	Sarcasm (Acc.)			KSounds (Acc.)		
					Multi	Image	Text	Multi	Audio	Video
Baseline	–	–	–	–	0.8286	0.6907	0.8066	0.6455	0.4998	0.4798
(a) w/o Function-space	Param.	Loss	Relative	Tensor	0.8355	0.7001	0.8120	0.6790	0.5207	0.5490
(b) w/o RFP	Func.	Loss	Absolute	Tensor	0.8348	0.6978	0.8132	0.6952	0.5191	0.5508
(c) w/o Loss-Align	Func.	Epoch	Relative	Tensor	0.8389	0.7024	0.8177	0.7352	0.5276	0.5551
(d) w/o Tensor-wise	Func.	Loss	Relative	Modality	0.8427	0.7082	0.8223	0.7449	0.5222	0.5613
FGMO (Ours)	**Func.**	**Loss**	**Relative**	**Tensor**	**0.8547**	**0.7165**	**0.8281**	**0.7565**	**0.5319**	**0.5991**

Figure 3: Training time and accuracy under different K on CREMA-D dataset.



timodal accuracy by 1.92% and 7.75% on Sarcasm and KSounds, respectively, supporting the benefit of function-space measurement even when reference normalization is retained. Replacing RFP with absolute function-space responses reduces multimodal accuracy by 1.99% and 6.13% on Sarcasm and KSounds, respectively, which highlights the importance of measuring modality progress relative to its own reference rather than directly comparing raw response magnitudes. Replacing loss alignment with epoch alignment also decreases performance, indicating that matching comparable optimization states is important for reference calibration. Finally, removing tensor-wise allocation consistently lowers accuracy, showing that fine-grained response control provides additional gains beyond modality-level allocation. Combining all four designs achieves the best multimodal and modality-specific performance on both datasets.

4.4 Hyperparameter Analysis

We investigate the influence of the adjustment interval K and the controller gain λ on the model performance and its internal optimization dynamics on the CREMA-D and Sarcasm datasets.

Analysis of Adjustment Interval K: The interval K controls how frequently FGMO estimates RFP and invokes FRC to update the modality budgets. A small K provides frequent feedback, but makes the controller more responsive to short-term optimization fluctuations and increases the function-space probing overhead. In contrast, a large K reduces the computational cost but introduces delayed feedback, limiting timely correction of evolving modality imbalance. As shown in Figure 3, $K = 20$ achieves the highest accuracy of 0.8589 with a training time of 2.1796 hours. These results suggest that a moderate update frequency provides a better balance among feedback stability, control responsiveness, and training efficiency. We therefore set $K = 20$ as the default.

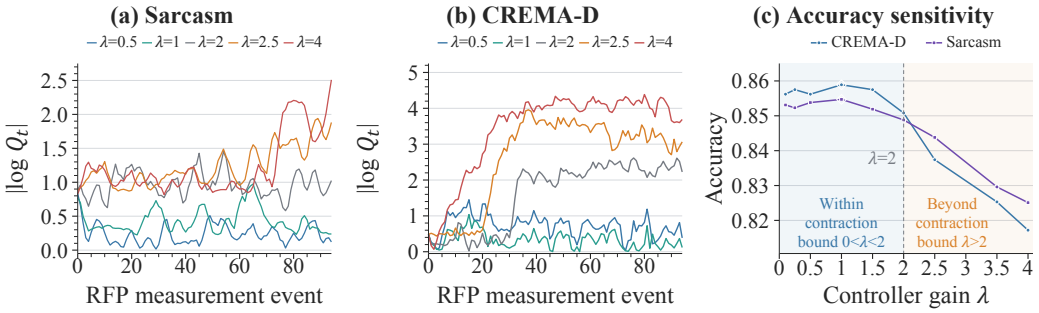


Figure 4: Sensitivity analysis of the controller gain λ on the Sarcasm and CREMA-D datasets. (a) and (b) Evolution of the absolute log RFP ratio $|\log Q^t|$ under different controller gains on Sarcasm and CREMA-D, respectively. (c) Impact of λ on the multimodal classification accuracy. The dashed line ($\lambda = 2$) denotes the boundary of the ideal target-contraction range.

Analysis of Controller Gain λ: To empirically examine the target-contraction behavior characterized in **Theorem 1**, we evaluate the controller gain across a broad range $\lambda \in \{0.1, 0.25, 0.5, 1, 1.5, 2, 2.5, 3.5, 4\}$. As illustrated in Figure 4 (c), when λ falls within the theoretically derived target-contraction range ($0 < \lambda < 2$), the model accuracy remains stably high, peaking at $\lambda = 1$ (0.8589, 0.8547). However, once λ reaches or exceeds this range ($\lambda \geq 2$), the

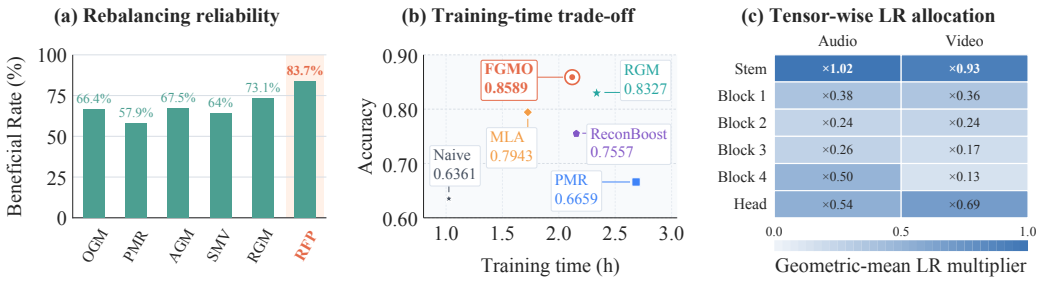


Figure 5: (a) Beneficial Rates of different modality signals under controlled interventions. (b) Training-time comparison with representative multimodal learning methods. (c) Tensor-wise learning-rate allocation across the audio and video branches, where each cell reports the geometric-mean learning-rate multiplier.

performance degrades significantly. Figure 4 (a), (b) further shows training dynamics consistent with the theoretical analysis. Within the target-contraction range (e.g., $\lambda = 0.5, 1$), the closed-loop controller effectively maintains substantially smaller RFP discrepancies. Conversely, an excessively large gain (e.g., $\lambda = 2.5, 4$) lies outside the target-contraction range and is accompanied by pronounced oscillations and an enlarged RFP discrepancy.

4.5 Further Analysis

Reliability of Rebalancing Signals. To evaluate signal reliability across training stages, we extend the controlled interventions (Figure 1) to six datasets and 1,000 checkpoints. We compare RFP with five baselines using the *Beneficial Rate*—the percentage of checkpoints where signal-guided interventions improve accuracy over vanilla continuations. As Figure 5 (a) shows, RFP achieves the highest beneficial rate of 83.7%, outperforming the competitive baselines and demonstrating superior and more reliable guidance for stage-dependent modality rebalancing.

Computational Cost. FGMO introduces additional computation for unimodal reference construction and periodic function-space probing. Figure 5 (b) compares joint-training time, excluding reference construction. The same fixed references are reused across training seeds (Appendix C.4), allowing their construction cost to be amortized over repeated runs under the same setting. Although reference construction increases the end-to-end training cost, FGMO achieves improved multimodal performance through reference-calibrated rebalancing.

Interpretable LR Allocation. As Figure 5 (c) shows, FRC exhibits a clear hierarchical pattern in learning-rate allocation rather than a simple depth-wise decay. Stem layers retain multipliers near $1\times$, suggesting FRC preserves the adaptation capacity of low-level feature extractors. Meanwhile, intermediate and deep blocks are substantially suppressed to stabilize high-level representations, before the multipliers increase again at the prediction heads. Moreover, audio and video show distinct allocation patterns in deeper blocks, confirming that FRC dynamically learns a modality-specific optimization strategy instead of applying a predefined schedule.

5 Conclusion

In this paper, we propose Function-Space Guided Multimodal Optimization (FGMO) to alleviate modality imbalance through comparable progress estimation and coordinated optimization. Functional Progress Estimation (FPE) measures update-induced function-space responses and calibrates them against loss-aligned unimodal references to produce a comparable progress signal termed RFP. Guided by this signal, Functional Response Control (FRC) redistributes modality-level function-space budgets and realizes target responses through tensor-wise learning-rate adjustment. Theoretical analysis establishes a one-step target-contraction property of FRC under bounded controller-state mismatch, and extensive experiments demonstrate the effectiveness of FGMO across multiple multimodal benchmarks.

AI USE STATEMENT

We used GPT solely to assist with language polishing and improving the readability of the manuscript. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *ICCV*, pp. 609–617. IEEE, 2017.

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.

Ari S. Benjamin, David Rolnick, and Konrad P. Körding. Measuring and regularizing networks in function space. In *ICLR*. OpenReview.net, 2019.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *ACL*, pp. 2506–2515. Association for Computational Linguistics, 2019.

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.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pp. 4724–4733. IEEE, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, pp. 721–725. IEEE, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *CVPR*, pp. 20029–20038. IEEE, 2023.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junhong Liu, and Song Guo. Detached and interactive multimodal learning. In *ACM Multimedia*, pp. 5470–5478. ACM, 2024.

Xiyuan Gao, Bing Cao, Baoquan Gong, and Pengfei Zhu. Reconcile gradient modulation for harmony multimodal learning. In *AAAI*, pp. 21225–21233. AAAI Press, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778. IEEE, 2016.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. ReconBoost: Boosting can achieve modality reconcilement. In *ICML*, pp. 19573–19597. PMLR, 2024.

Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition. In *CVPR*, pp. 25854–25863. IEEE, 2025.

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 *ICML*, pp. 9226–9259. PMLR, 2022.

Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *NeurIPS*, pp. 129580–129607. Curran Associates, Inc., 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pp. 22157–22167. IEEE, 2023.

Edward Milsom, Ben Anson, and Laurence Aitchison. Function-space learning rates. In *ICML*, pp. 44225–44251. PMLR, 2025.

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. In *CVPR*, pp. 4207–4215. IEEE, 2016.

Weizhi Nie, Yan Yan, Dan Song, and Kun Wang. Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition. *Multimedia Tools and Applications*, 80(11):16205–16214, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035. Curran Associates, Inc., 2019.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pp. 8228–8237. IEEE, 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, pp. 3942–3951. AAAI Press, 2018.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *ICLR*. OpenReview.net, 2026.

Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, and Yarin Gal. Continual learning via sequential function-space variational inference. In *ICML*, pp. 18871–18887. PMLR, 2022.

Aidan Scannell, Riccardo Mereu, Paul Edmund Chang, Ella Tamir, Joni Pajarinen, and Arno Solin. Function-space parameterization of neural networks for sequential learning. In *ICLR*. OpenReview.net, 2024.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pp. 12692–12702. IEEE, 2020.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *ICCV*, pp. 2250–2259. IEEE, 2025.

Yake Wei and Di Hu. MMPareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, pp. 52559–52572. PMLR, 2024.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, pp. 27328–27337. IEEE, 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 *ICML*, pp. 24043–24055. PMLR, 2022.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *NeurIPS*, pp. 62108–62122. Curran Associates, Inc., 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.

Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of ACL*, pp. 1824–1834. Association for Computational Linguistics, 2022.

Jianfei Yu and Jing Jiang. Adapting BERT for target-oriented multimodal sentiment classification. In *IJCAI*, pp. 5408–5414. ijcai.org, 2019.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *CVPR*, pp. 27446–27456. IEEE, 2024.

A Algorithmic Overview

Algorithm 1 distinguishes optimization-step index k from adjustment-event index t. The modality tensor sets include both encoder and classifier parameters. At an adjustment event, the provisional optimizer call advances its internal state once. We restore only its parameter displacement, retain the resulting optimizer state, and apply the recovered direction with the new learning rates without a second optimizer call. The response-probing pass must not advance the task optimizer state or permanently change non-parameter model buffers.

Algorithm 1 Function-Space Guided Multimodal Optimization (FGMO)

Require: Training set $\mathcal{D}$; reference trajectories $\{\mathcal{L}^{\mathrm{s}}_m(\tau), \delta^{\mathrm{s}}_{m,l}(\tau)\}$; interval K; controller gain λ; stabilizer ϵ; outer-learning-rate schedule; number of optimization steps N_{opt}
Ensure: Optimized parameters $\mathbf{\Theta}$

Initialize $\mathbf{\Theta}$ and the task optimizer state
Initialize common tensor learning rates $\{\eta^0_{m,l} > 0\}$, positive shares $\mathbf{s}^0 \in \Delta_M$, and $t \leftarrow 0$
for $k = 1, \ldots, N_{\mathrm{opt}}$ **do**
    Sample a mini-batch and compute the task gradients using Eq. (1)
    Compute the current modality-specific losses
    if $k \bmod K = 0$ **then**
        Denote the mini-batch by $\mathcal{B}_t$ and save the pre-update parameters as $\mathbf{\Theta}^t$
        Record the current modality losses as $\{\mathcal{L}^t_m\}$
        Perform one provisional task optimizer update using the current rates $\{\eta^t_{m,l}\}$
        Recover $\{\mathbf{u}^t_{m,l}\}$ using Eq. (2)
        Restore $\mathbf{\Theta}^t$ while retaining the post-step task optimizer state
        Estimate $\{\delta^t_{m,l}\}$ at the restored parameters on $\mathcal{B}_t$ using Appendix B.2
        for $m = 1, \ldots, M$ **do**
            Find τ^t_m using Eq. (4)
            Set $\delta'^{,t}_{m,l} \leftarrow \delta^{\mathrm{s}}_{m,l}(\tau^t_m)$ for $l \in \mathcal{T}_m$
            Compute ρ^t_m using Eq. (5)
        end for
        Update $\mathbf{s}^{t+1}$ using Eq. (7)
        Set $B^{t+1}_m \leftarrow s^{t+1}_m \sum_{j=1}^{M} \sum_{l \in \mathcal{T}_j} \delta^t_{j,l}$ for all modalities
        Compute $\{\eta^{t+1}_{m,l}\}$ using Eq. (8)
        Set $\mathbf{\Theta}_{m,l} \leftarrow \mathbf{\Theta}^t_{m,l} + \eta^{t+1}_{m,l} \mathbf{u}^t_{m,l}$ for every tensor l of each modality m, without another optimizer call
        $t \leftarrow t + 1$
    else
        Perform one task optimizer update using the current learning rates
    end if
end for
return $\mathbf{\Theta}$

B Additional Theoretical Analysis

In this section, we provide additional theoretical analysis of the proposed Function-Space Guided Multimodal Optimization (FGMO). We first derive the function-space response estimator used to construct Relative Functional Progress (RFP) and analyze its scale property. We then derive the closed-form solution of Functional Response Control (FRC) and show how tensor-wise learning-rate adjustment realizes the modality-level function-space budget. Finally, we characterize the one-step

contractive property of the FRC response target, analyze deviations between assigned and realized responses, and extend the result to multiple modalities. Importantly, this analysis does not require assumptions on the neural-network state or reference scale at the next adjustment event.

B.1 Function-Space Response Estimation

Function-space response. Consider a minibatch containing n_b samples and a classification problem with C classes. Let $\mathbf{z}_n^t \in \mathbb{R}^C$ denote the logits of the n-th sample at adjustment event t. Since the softmax prediction is invariant to a constant shift of all logits, we remove this uninformative direction using

$$\mathbf{P} = \mathbf{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top. \tag{11}$$

We then stack the centered logits over the minibatch as

$$\mathbf{f}^t = \mathrm{vec}\left(\mathbf{P}\mathbf{z}_1^t, \ldots, \mathbf{P}\mathbf{z}_{n_b}^t\right) \in \mathbb{R}^D, \qquad D = n_b C. \tag{12}$$

For a vectorized parameter tensor $\mathbf{\Theta}_{m,l}$ of modality m, let $\mathbf{u}_{m,l}^t$ denote its update direction under a unit learning rate. In particular, if a provisional optimizer step is written as

$$\widetilde{\mathbf{\Theta}}_{m,l}^t = \mathbf{\Theta}_{m,l}^t + \eta_{m,l}^t \mathbf{u}_{m,l}^t, \tag{13}$$

we recover

$$\mathbf{u}_{m,l}^t = \frac{\widetilde{\mathbf{\Theta}}_{m,l}^t - \mathbf{\Theta}_{m,l}^t}{\eta_{m,l}^t}. \tag{14}$$

Using the unit-learning-rate update removes the direct dependence of the response estimate on the current learning-rate assignment.

Let

$$\mathbf{J}_{m,l}^t = \frac{\partial \mathbf{f}^t}{\partial \mathbf{\Theta}_{m,l}^t} \tag{15}$$

denote the Jacobian of the centered output with respect to $\mathbf{\Theta}_{m,l}^t$. For a sufficiently small update, a first-order expansion gives

$$f(\mathbf{\Theta}_{m,l}^t + \mathbf{u}_{m,l}^t) - f(\mathbf{\Theta}_{m,l}^t) = \mathbf{J}_{m,l}^t \mathbf{u}_{m,l}^t + \mathcal{O}\left(\|\mathbf{u}_{m,l}^t\|_2^2\right). \tag{16}$$

We therefore define the local function-space response of tensor l of modality m as

$$\delta_{m,l}^t = \frac{1}{\sqrt{D}} \left\|\mathbf{J}_{m,l}^t \mathbf{u}_{m,l}^t\right\|_2. \tag{17}$$

Directly constructing $\mathbf{J}_{m,l}^t$ is expensive for modern neural networks. Following the function-space estimation principle (Milsom et al., 2025), we estimate Eq. (17) through a Gaussian random projection.

Let

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D), \tag{18}$$

and define

$$\phi^t = \frac{1}{\sqrt{D}} \boldsymbol{\varepsilon}^\top \mathbf{f}^t. \tag{19}$$

The gradient of this scalar projection with respect to $\mathbf{\Theta}_{m,l}$ is

$$\mathbf{g}_{m,l}^{\phi,t} = \nabla_{\mathbf{\Theta}_{m,l}} \phi^t = \frac{1}{\sqrt{D}} (\mathbf{J}_{m,l}^t)^\top \boldsymbol{\varepsilon}. \tag{20}$$

Proposition 1 (Random-projection response estimator). For $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$, the projected response satisfies

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\langle \mathbf{u}_{m,l}^t, \mathbf{g}_{m,l}^{\phi,t} \right\rangle^2\right] = \frac{1}{D}\left\|\mathbf{J}_{m,l}^t \mathbf{u}_{m,l}^t\right\|_2^2. \tag{21}$$

Hence,

$$\delta_{m,l}^t = \sqrt{\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\langle \mathbf{u}_{m,l}^t, \mathbf{g}_{m,l}^{\phi,t} \right\rangle^2\right]}. \tag{22}$$

Proof. Substituting Eq. (20) gives

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\langle \mathbf{u}_{m,l}^{t}, \mathbf{g}_{m,l}^{\phi,t}\right\rangle^{2}\right] = \frac{1}{D}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left(\boldsymbol{\varepsilon}^{\top}\mathbf{J}_{m,l}^{t}\mathbf{u}_{m,l}^{t}\right)^{2}\right]$$
$$= \frac{1}{D}(\mathbf{u}_{m,l}^{t})^{\top}(\mathbf{J}_{m,l}^{t})^{\top}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top}\right]\mathbf{J}_{m,l}^{t}\mathbf{u}_{m,l}^{t}. \tag{23}$$

Since

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\top}\right] = \mathbf{I}_{D}, \tag{24}$$

we obtain

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\left\langle \mathbf{u}_{m,l}^{t}, \mathbf{g}_{m,l}^{\phi,t}\right\rangle^{2}\right] = \frac{1}{D}(\mathbf{u}_{m,l}^{t})^{\top}(\mathbf{J}_{m,l}^{t})^{\top}\mathbf{J}_{m,l}^{t}\mathbf{u}_{m,l}^{t}$$
$$= \frac{1}{D}\left\|\mathbf{J}_{m,l}^{t}\mathbf{u}_{m,l}^{t}\right\|_{2}^{2}. \tag{25}$$

Taking the square root proves Eq. (22).

This result is important for two reasons. First, the estimator measures the functional effect of an update rather than its magnitude in parameter space. Second, it requires only the gradient of a scalar projection and therefore avoids explicitly constructing the full Jacobian. In contrast to directly using $\|\mathbf{u}_{m,l}^{t}\|$, the response $\delta_{m,l}^{t}$ accounts for both the optimizer update and the local input–output geometry represented by $\mathbf{J}_{m,l}^{t}$.

B.2 Low-Variance Online Estimation

A single random projection gives an unbiased estimate of the squared function-space response, but its variance can be high. We therefore use the Kronecker-factored covariance approximation introduced for function-space learning-rate estimation (Milsom et al., 2025).

For a matrix parameter $\mathbf{W}_{m,l} \in \mathbb{R}^{d_1 \times d_2}$ with $\boldsymbol{\Theta}_{m,l} = \mathrm{vec}(\mathbf{W}_{m,l})$, let $\mathbf{U}_{m,l}^{t}$ and $\mathbf{G}_{m,l}^{\phi,t}$ reshape $\mathbf{u}_{m,l}^{t}$ and $\mathbf{g}_{m,l}^{\phi,t}$ to its dimensions. Define

$$\mathbf{Z}_{m,l}^{t} = \mathbf{U}_{m,l}^{t} \odot \mathbf{G}_{m,l}^{\phi,t}, \tag{26}$$

where $\odot$ denotes the element-wise product. The projected response can be written as

$$\left\langle \mathbf{u}_{m,l}^{t}, \mathbf{g}_{m,l}^{\phi,t}\right\rangle = \sum_{i,j}[\mathbf{Z}_{m,l}^{t}]_{ij}. \tag{27}$$

With covariance factors $\mathbf{C}^{(1)}$ and $\mathbf{C}^{(2)}$, assume

$$\mathrm{Cov}\left([\mathbf{Z}_{m,l}^{t}]_{ij}, [\mathbf{Z}_{m,l}^{t}]_{i'j'}\right) \approx [\mathbf{C}^{(1)}]_{ii'}[\mathbf{C}^{(2)}]_{jj'}. \tag{28}$$

Under this approximation, the squared response can be estimated from three scalar second-order statistics,

$$A_{0} = \mathbb{E}\left[\|\mathbf{Z}_{m,l}^{t}\|_{F}^{2}\right], \tag{29}$$

$$A_{1} = \mathbb{E}\left[\sum_{j}\left(\sum_{i}[\mathbf{Z}_{m,l}^{t}]_{ij}\right)^{2}\right], \tag{30}$$

and

$$A_{2} = \mathbb{E}\left[\sum_{i}\left(\sum_{j}[\mathbf{Z}_{m,l}^{t}]_{ij}\right)^{2}\right]. \tag{31}$$

This yields

$$(\delta_{m,l}^{t})^{2} \approx \frac{A_{1}A_{2}}{A_{0}}. \tag{32}$$

For an order-$d_{m,l}$ parameter tensor, let $\mathcal{Z}_{m,l}^{t}$ denote the tensor reshape of $\mathbf{u}_{m,l}^{t} \odot \mathbf{g}_{m,l}^{\phi,t}$. Define

$$A_{m,l,0} = \mathbb{E}\left[\|\mathcal{Z}_{m,l}^{t}\|_{F}^{2}\right], \tag{33}$$

and let $A_{m,l,d}$ denote the second-order statistic obtained by summing $\mathcal{Z}^t_{m,l}$ along its d-th dimension before taking the squared Frobenius norm. The resulting estimator takes the form

$$(\delta^t_{m,l})^2 \approx \frac{\prod_{d=1}^{d_{m,l}} A_{m,l,d}}{A_{m,l,0}^{d_{m,l}-1}}. \tag{34}$$

In practice, we maintain exponential moving averages of these scalar statistics. This reduces the variance of the online estimate while requiring only a small number of additional statistics for each parameter tensor. The unit-learning-rate update $\mathbf{u}^t_{m,l}$ is used throughout the estimation process, so the recorded response remains comparable when the actual tensor-wise learning rates change over training.

Empirical Validation of Response Estimation. We further evaluate the fidelity and computational efficiency of the function-space response estimator. We compare the proposed Kronecker + EMA estimator with direct Monte Carlo estimation using one, four, and eight independent random projections. The estimated tensor-wise responses are compared against directly measured function-space responses. We report Pearson and Spearman correlations, mean and median absolute relative error (ARE), and the relative computational cost normalized by a single projection. The results are summarized in Table 3.

Table 3: Fidelity and computational efficiency of different function-space response estimators. Relative cost is normalized by one random projection. Higher Pearson and Spearman correlations indicate better agreement with the directly measured responses, while lower ARE indicates smaller estimation error.

Estimator	Pearson ↑	Spearman ↑	Mean ARE ↓	Median ARE ↓	Cost ↓
1 Projection	0.9042	0.8783	59.44%	63.56%	**1.00**×
4 Projections	0.9415	0.9251	31.44%	23.34%	3.89×
8 Projections	**0.9931**	**0.9501**	**13.37%**	**10.89%**	7.87×
Kronecker + EMA	0.9798	0.9434	15.75%	13.86%	2.33×

As the number of independent projections increases from one to eight, the estimation fidelity consistently improves: both Pearson and Spearman correlations increase, while the mean and median ARE decrease. However, this improvement comes with a nearly proportional increase in computational cost. In particular, eight projections achieve the highest estimation accuracy but require a relative cost of $7.87\times$.

In contrast, Kronecker + EMA provides a more favorable accuracy–efficiency trade-off. Compared with four independent projections, it improves the Pearson correlation from 0.9415 to 0.9798 and the Spearman correlation from 0.9251 to 0.9434, while reducing the mean ARE from 31.44% to 15.75% and the computational cost from $3.89\times$ to $2.33\times$. Moreover, it approaches the correlation achieved by eight projections while using only 29.6% of their computational cost. These results indicate that the adopted estimator preserves the relative structure of tensor-wise function-space responses at substantially lower cost, making it suitable for periodic online response estimation in FGMO.

B.3 Reference-Aware Aggregation and Scale Property of RFP

The tensor-wise function-space response can vary substantially across modalities and training stages. We therefore compare each tensor with its loss-aligned unimodal reference before forming the modality-level control signal.

For tensor $l \in \mathcal{T}_m$, define the tensor-wise relative functional progress

$$r^t_{m,l} = \frac{\delta^t_{m,l}}{\delta'^{,t}_{m,l}}, \tag{35}$$

and the reference-aware weight

$$w^t_{m,l} = \frac{\delta'^{,t}_{m,l}}{\sum_{j\in\mathcal{T}_m} \delta'^{,t}_{m,j}}. \tag{36}$$

The weights satisfy $w_{m,l}^t \geq 0$ and $\sum_{l\in\mathcal{T}_m} w_{m,l}^t = 1$. The modality-level RFP is the weighted aggregation

$$\rho_m^t = \sum_{l\in\mathcal{T}_m} w_{m,l}^t r_{m,l}^t. \tag{37}$$

Proposition 2 (Reference-aware aggregation). Ignoring the numerical stabilizer, Eq. (37) is exactly equivalent to

$$\rho_m^t = \frac{\sum_{l\in\mathcal{T}_m} \delta_{m,l}^t}{\sum_{l\in\mathcal{T}_m} \delta_{m,l}^{\prime,t}}. \tag{38}$$

Proof. Substituting Eqs. (35) and (36) into Eq. (37) gives

$$\begin{aligned}\rho_m^t &= \sum_{l\in\mathcal{T}_m} \frac{\delta_{m,l}^{\prime,t}}{\sum_{j\in\mathcal{T}_m} \delta_{m,j}^{\prime,t}} \frac{\delta_{m,l}^t}{\delta_{m,l}^{\prime,t}} \\ &= \frac{\sum_{l\in\mathcal{T}_m} \delta_{m,l}^t}{\sum_{j\in\mathcal{T}_m} \delta_{m,j}^{\prime,t}}.\end{aligned} \tag{39}$$

This form provides two useful interpretations. First, RFP preserves tensor-level relative progress before aggregation. Second, tensors with larger responses along the unimodal trajectory receive larger weights in the modality-level signal.

Scale cancellation. Suppose all tensor-wise responses of modality m are multiplied by the same positive constant c_m in both joint and loss-aligned unimodal training:

$$\widetilde{\delta}_{m,l}^t = c_m \delta_{m,l}^t, \qquad \widetilde{\delta}_{m,l}^{\prime,t} = c_m \delta_{m,l}^{\prime,t}. \tag{40}$$

Then

$$\widetilde{r}_{m,l}^t = r_{m,l}^t, \qquad \widetilde{w}_{m,l}^t = w_{m,l}^t, \qquad \widetilde{\rho}_m^t = \rho_m^t. \tag{41}$$

Hence, RFP is invariant to a common modality-specific response scale. This property allows the controller to compare relative tensor-wise progress across heterogeneous modalities without directly comparing their raw response magnitudes. Loss alignment further matches the joint and unimodal tensors at similar optimization states rather than at the same epoch.

B.4 Closed-Form Solution of the KL-Regularized Controller

Let

$$\mathbf{s}^t = (s_1^t, \ldots, s_M^t) \tag{42}$$

denote the modality budget shares at adjustment event t, where

$$s_m^t > 0, \qquad \sum_{m=1}^M s_m^t = 1. \tag{43}$$

Given the current RFP values $\boldsymbol{\rho}^t = (\rho_1^t, \ldots, \rho_M^t)$, we update the shares by solving

$$\mathbf{s}^{t+1} = \arg\min_{\mathbf{s}\in\Delta_M} \left[\sum_{m=1}^M s_m \log \rho_m^t + \frac{1}{\lambda} D_{\mathrm{KL}}(\mathbf{s}\|\mathbf{s}^t) \right], \tag{44}$$

where

$$D_{\mathrm{KL}}(\mathbf{s}\|\mathbf{s}^t) = \sum_{m=1}^M s_m \log \frac{s_m}{s_m^t}. \tag{45}$$

The first term assigns a lower cost to modalities with smaller RFP, while the KL term limits abrupt changes in the allocation.

Proposition 3 (Closed-form modality update). The solution of Eq. (44) is

$$s_m^{t+1} = \frac{s_m^t (\rho_m^t)^{-\lambda}}{\sum_{j=1}^M s_j^t (\rho_j^t)^{-\lambda}}. \tag{46}$$

Proof. Introducing a Lagrange multiplier μ for the simplex constraint gives

$$\mathcal{J}(\mathbf{s},\mu) = \sum_m s_m \log \rho_m^t + \frac{1}{\lambda}\sum_m s_m \log \frac{s_m}{s_m^t} + \mu\left(\sum_m s_m - 1\right). \tag{47}$$

Taking the derivative with respect to s_m gives

$$\frac{\partial \mathcal{J}}{\partial s_m} = \log \rho_m^t + \frac{1}{\lambda}\left(\log \frac{s_m}{s_m^t} + 1\right) + \mu. \tag{48}$$

At the optimum,

$$\log \frac{s_m}{s_m^t} = -\lambda \log \rho_m^t - 1 - \lambda\mu. \tag{49}$$

Therefore,

$$s_m = c_{\text{norm}} s_m^t (\rho_m^t)^{-\lambda}, \tag{50}$$

where $c_{\text{norm}} = \exp(-1-\lambda\mu)$ is shared by all modalities. Applying the normalization constraint $\sum_m s_m = 1$ gives

$$c_{\text{norm}} = \left[\sum_j s_j^t (\rho_j^t)^{-\lambda}\right]^{-1}, \tag{51}$$

which leads to Eq. (46).

Eq. (46) has a direct feedback interpretation. If modality m has a smaller RFP, $(\rho_m^t)^{-\lambda}$ becomes larger and the modality receives a larger share at the next adjustment event. A modality with faster relative progress is adjusted in the opposite direction. Therefore, the controller forms a negative-feedback loop over reference-aware functional progress.

B.5 Tensor-Wise Realization of the Modality Budget

After the modality controller determines s_m^{t+1}, we assign

$$B_m^{t+1} = B_{\text{tot}}^t s_m^{t+1}, \tag{52}$$

where

$$B_{\text{tot}}^t = \sum_{m=1}^{M} \sum_{l\in\mathcal{T}_m} \delta_{m,l}^t \tag{53}$$

is the total unit-learning-rate tensor-wise response scale at adjustment event t. Accordingly, B_m^{t+1} denotes the modality-level response budget before scaling by the outer learning rate η_{outer}^t. Under the local first-order approximation and with $\epsilon = 0$, the corresponding aggregate realized tensor-wise response is approximately $\eta_{\text{outer}}^t B_m^{t+1}$, as established in Proposition 4.

Within modality m, the loss-aligned reference defines the loss-aligned reference profile

$$w_{m,l}^t = \frac{\delta_{m,l}'^{,t}}{\sum_{j\in\mathcal{T}_m} \delta_{m,j}'^{,t}}, \qquad l \in \mathcal{T}_m, \tag{54}$$

with $\sum_{l\in\mathcal{T}_m} w_{m,l}^t = 1$. The tensor-wise learning rate is

$$\eta_{m,l}^{t+1} = \eta_{\text{outer}}^t \frac{B_m^{t+1} w_{m,l}^t}{\delta_{m,l}^t + \epsilon}. \tag{55}$$

Proposition 4 (Local tensor-response budget realization). Under the first-order function approximation and with $\epsilon = 0$, let $\widetilde{\delta}_{m,l}^{t,\text{ctrl}}$ denote the functional response produced by applying the controller-assigned learning rate $\eta_{m,l}^{t+1}$ to the unit-learning-rate update measured at event t. Then

$$\widetilde{\delta}_{m,l}^{t,\text{ctrl}} \approx \eta_{\text{outer}}^t B_m^{t+1} w_{m,l}^t. \tag{56}$$

Consequently, the aggregate controller-realized response of modality m satisfies

$$\sum_{l\in\mathcal{T}_m} \widetilde{\delta}_{m,l}^{t,\text{ctrl}} \approx \eta_{\text{outer}}^t B_m^{t+1}. \tag{57}$$

Proof. Since $\delta^t_{m,l}$ measures the function-space response under a unit-learning-rate update, scaling the same update by $\eta^{t+1}_{m,l}$ gives, under the local linear approximation,

$$\widetilde{\delta}^{t,\mathrm{ctrl}}_{m,l} \approx \eta^{t+1}_{m,l}\delta^t_{m,l}. \tag{58}$$

Substituting Eq. (55) with $\epsilon = 0$ yields

$$\widetilde{\delta}^{t,\mathrm{ctrl}}_{m,l} \approx \eta^t_{\mathrm{outer}} B^{t+1}_m w^t_{m,l}. \tag{59}$$

Summing over all tensors in modality m gives

$$\begin{aligned}\sum_{l\in\mathcal{T}_m} \widetilde{\delta}^{t,\mathrm{ctrl}}_{m,l} &\approx \eta^t_{\mathrm{outer}} B^{t+1}_m \sum_{l\in\mathcal{T}_m} w^t_{m,l} \\ &= \eta^t_{\mathrm{outer}} B^{t+1}_m.\end{aligned} \tag{60}$$

Thus, the controller determines how much aggregate tensor-wise functional response is assigned to each modality, while $w^t_{m,l}$ determines how that budget is distributed across tensors. Importantly, the same reference profile is used both to construct modality-level RFP and to realize the tensor-wise update, which keeps the measurement and allocation stages consistent.

With $\epsilon > 0$, the controller-realized aggregate response becomes

$$\sum_{l\in\mathcal{T}_m} \widetilde{\delta}^{t,\mathrm{ctrl}}_{m,l} \approx \eta^t_{\mathrm{outer}} B^{t+1}_m \sum_{l\in\mathcal{T}_m} w^t_{m,l} \frac{\delta^t_{m,l}}{\delta^t_{m,l} + \epsilon}. \tag{61}$$

When $\delta^t_{m,l} \gg \epsilon$, the effect of the stabilizer is negligible.

Empirical Validation of Budget Realization. We further examine whether the tensor-wise learning rates produced by FRC realize the assigned modality-level response budgets in practice. At each adjustment event, we compare the target response implied by Eq. (57) with the aggregate response produced by the controller-scaled tensor updates. As shown in Table 4, the assigned and realized responses show strong agreement, with R^2 values of 0.9736 and 0.9586 for Image and Text, respectively. The corresponding mean relative errors are 2.74% and 3.21%, while the 95th-percentile errors remain below 10%. These results empirically support the local budget-realization property characterized in Proposition 4.

Table 4: Empirical validation of modality-level response-budget realization. The target response is compared with the aggregate response produced by the controller-scaled tensor updates.

Modality	$R^2 \uparrow$	**Mean Error** $\downarrow$	**P95 Error** $\downarrow$
Image	0.9736	2.74%	8.97%
Text	0.9586	3.21%	9.69%

B.6 Contractive Property of Functional Response Control

We next analyze how FRC transforms the RFP discrepancy observed at the current adjustment event into the modality-level response target assigned by the controller. Unlike an analysis of the next measured network state, this one-step analysis uses only quantities available at the current adjustment event and therefore does not require assumptions on future network dynamics. For clarity, we first consider two modalities and omit the numerical stabilizer ϵ.

For each modality, define the current aggregate functional response and the corresponding loss-aligned reference response as

$$R^t_m = \sum_{l\in\mathcal{T}_m} \delta^t_{m,l}, \qquad R'^{,t}_m = \sum_{l\in\mathcal{T}_m} \delta'^{,t}_{m,l}. \tag{62}$$

Using Proposition 2,

$$\rho^t_m = \frac{R^t_m}{R'^{,t}_m}. \tag{63}$$

Define the current RFP ratio

$$Q^t = \frac{\rho_1^t}{\rho_2^t}, \tag{64}$$

where $Q^t = 1$ indicates equal reference-aware relative functional progress.

We further define the normalized functional-response share measured at the current adjustment event as

$$\widehat{s}_m^t = \frac{R_m^t}{R_1^t + R_2^t}. \tag{65}$$

Since

$$\frac{\widehat{s}_1^t}{\widehat{s}_2^t} = \frac{R_1^t}{R_2^t}, \tag{66}$$

the current RFP ratio satisfies the exact identity

$$Q^t = \frac{\widehat{s}_1^t}{\widehat{s}_2^t} \frac{R_2'^{,t}}{R_1'^{,t}}. \tag{67}$$

Let

$$S^t = \frac{s_1^t}{s_2^t} \tag{68}$$

denote the ratio of the controller shares carried into adjustment event t. From Eq. (46), the normalization term cancels, yielding

$$S^{t+1} = S^t (Q^t)^{-\lambda}. \tag{69}$$

Because the modality-level target satisfies $B_m^{t+1} = B_{\text{tot}}^t s_m^{t+1}$, we define the reference-normalized target ratio produced by FRC as

$$Q_{\text{tar}}^{t+1|t} = \frac{B_1^{t+1}/R_1'^{,t}}{B_2^{t+1}/R_2'^{,t}}. \tag{70}$$

The notation $t+1|t$ emphasizes that this is the response target assigned for the next controlled update using information measured at event t, rather than the RFP value measured at a future adjustment event.

Since the common total budget cancels,

$$Q_{\text{tar}}^{t+1|t} = S^{t+1} \frac{R_2'^{,t}}{R_1'^{,t}}. \tag{71}$$

We now define the controller-state mismatch

$$d_t = \log \frac{S^t}{\widehat{s}_1^t / \widehat{s}_2^t}. \tag{72}$$

Here, d_t is a directly defined and observable quantity. It measures the difference between the allocation state carried by the controller and the response share measured from the current function-space responses.

Theorem 1 (One-step target contraction of FRC). For the two-modality case and the RFP definition without the numerical stabilizer, the target ratio generated by FRC satisfies

$$\log Q_{\text{tar}}^{t+1|t} = (1 - \lambda) \log Q^t + d_t. \tag{73}$$

Consequently,

$$\left| \log Q_{\text{tar}}^{t+1|t} \right| \le |1 - \lambda| \left| \log Q^t \right| + |d_t|. \tag{74}$$

If $0 < \lambda < 2$ and

$$|d_t| < (1 - |1 - \lambda|) \left| \log Q^t \right|, \tag{75}$$

then

$$\left| \log Q_{\text{tar}}^{t+1|t} \right| < \left| \log Q^t \right|. \tag{76}$$

Proof. From Eq. (71) and Eq. (69),

$$Q_{\text{tar}}^{t+1|t} = S^t (Q^t)^{-\lambda} \frac{R_2'^{,t}}{R_1'^{,t}}. \tag{77}$$

Meanwhile, Eq. (67) gives

$$\frac{R_2'^{,t}}{R_1'^{,t}} = Q^t \frac{\widehat{s}_2^t}{\widehat{s}_1^t}. \tag{78}$$

Substituting this expression yields

$$\begin{aligned} Q_{\text{tar}}^{t+1|t} &= S^t (Q^t)^{-\lambda} Q^t \frac{\widehat{s}_2^t}{\widehat{s}_1^t} \\ &= (Q^t)^{1-\lambda} \frac{S^t}{\widehat{s}_1^t / \widehat{s}_2^t}. \end{aligned} \tag{79}$$

Taking logarithms and using Eq. (72) gives Eq. (73). Applying the triangle inequality gives Eq. (74). For $0 < \lambda < 2$, $|1 - \lambda| < 1$. If Eq. (75) holds, then

$$\begin{aligned} \left| \log Q_{\text{tar}}^{t+1|t} \right| &< |1 - \lambda| \left| \log Q^t \right| + (1 - |1 - \lambda|) \left| \log Q^t \right| \\ &= \left| \log Q^t \right|, \end{aligned} \tag{80}$$

which proves Eq. (76).

Corollary 1 (Matched controller state). If the carried controller share matches the currently measured response share, i.e.,

$$s_m^t = \widehat{s}_m^t, \qquad m \in \{1, 2\}, \tag{81}$$

then $d_t = 0$, and Eq. (73) reduces to

$$\log Q_{\text{tar}}^{t+1|t} = (1 - \lambda) \log Q^t. \tag{82}$$

Therefore, for $0 < \lambda < 2$,

$$\left| \log Q_{\text{tar}}^{t+1|t} \right| = |1 - \lambda| \left| \log Q^t \right| < \left| \log Q^t \right| \tag{83}$$

whenever $Q^t \neq 1$.

The result provides a direct interpretation of the controller gain. When $0 < \lambda < 1$, the ideal target reduces the log-RFP discrepancy without reversing its sign. When $1 < \lambda < 2$, the ideal target crosses the balanced state but still has a smaller absolute discrepancy. At $\lambda = 1$, the ideal term $(1 - \lambda) \log Q^t$ vanishes, and the remaining target discrepancy is determined entirely by the measurable controller-state mismatch d_t. For $\lambda \geq 2$, the ideal target no longer has a contraction factor below one.

Importantly, Theorem 1 characterizes the response target generated by FRC at the current adjustment event. It does not assume that the future neural-network state exactly follows this target, and therefore does not require a model of the future Jacobian, optimizer state, minibatch, or loss-aligned reference.

B.7 Empirical Diagnostics of Controller-State Consistency

Theorem 1 expresses departures from the ideal target contraction through the observable controller-state mismatch d_t. Since $d_t = 0$ whenever the controller shares s_m^t coincide with the measured response shares $\widehat{s}_m^t$, we examine their agreement empirically on Sarcasm.

Controller–Response Share Agreement. We compute the normalized response share $\widehat{s}_m^t$ using Eq. (65) and compare it with the controller share s_m^t at the same adjustment event. We evaluate this correspondence under controller gains $\lambda \in \{0.5, 1.0, 1.5\}$.

As shown in Table 5, the normalized response shares closely follow the controller shares for all evaluated gains. Pearson correlations remain above 0.989, while the identity-line coefficients $R^2_{y=x}$ remain above 0.974. The mean absolute deviations range from 0.030 to 0.047, and the 95th-percentile deviations remain at or below 0.112. In particular, at $\lambda = 1.0$, Pearson correlation reaches 0.995 and $R^2_{y=x}$ reaches 0.988, with a mean absolute deviation of 0.030. These statistics indicate that the allocation state maintained by FRC generally remains close to the response share measured at the current event. Exact equality between the two quantities is not assumed by Theorem 1.

Table 5: Empirical diagnostics of controller-state consistency and reference dynamics. Controller–response agreement compares the normalized functional-response share $\widehat{s}_m^t$ with the carried controller share s_m^t, where $R^2_{y=x}$ measures agreement with the identity line. Reference drift is reported as a supplementary diagnostic and is not an assumption of Theorem 1.

Controller–Response Agreement					
λ	Pearson ↑	$R^2_{y=x}$ ↑	Mean ↓	Median ↓	P95 ↓
0.5	0.989	0.974	0.047	0.036	0.112
1.0	0.995	0.988	0.030	0.024	0.075
1.5	0.991	0.981	0.039	0.029	0.098

Reference Drift Δ_{ref}			
Mean ↓	Median ↓	P95 ↓	≤ 0.1 ↑
0.0664	0.0652	0.1041	86.9%

Table 6: Empirical evaluation of the sufficient condition and one-step target contraction in Theorem 1.

λ	Events	Condition satisfaction rate (%) ↑	Target contraction rate (%) ↑
0.5	1,000	92.2	94.3
1.0	1,000	96.2	96.2
1.5	1,000	90.1	93.4

Reference Dynamics. For completeness, we also report how the relative loss-aligned reference normalization changes across consecutive adjustment events. Define

$$\Delta_{\mathrm{ref}}^t = \left| \log \frac{R_1'^{,t+1}/R_2'^{,t+1}}{R_1'^{,t}/R_2'^{,t}} \right|. \tag{84}$$

A smaller Δ_{ref}^t indicates a smoother change in the relative reference scale. As shown in Table 5, the mean and median reference drifts are 0.0664 and 0.0652, respectively, and 86.9% of adjustment events satisfy $\Delta_{\mathrm{ref}}^t \leq 0.1$. The 95th-percentile drift is 0.1041. These statistics describe the empirical evolution of the loss-aligned reference but are not required by the one-step target contraction result in Theorem 1.

Empirical Validation of Target Contraction. We further directly evaluate the sufficient condition in Theorem 1 and the contraction of the assigned response targets. For each controller gain $\lambda \in \{0.5, 1.0, 1.5\}$, we evaluate 1,000 adjustment events using the stabilizer-free definitions in the theorem. At each event, the controller-state mismatch d_t is computed using the controller shares before the allocation update. The condition satisfaction rate is the percentage of events satisfying Eq. (75), while the target contraction rate is the percentage satisfying Eq. (76). Both rates use the same 1,000 events as their denominator.

As shown in Table 6, the sufficient condition holds for 90.1–96.2% of the evaluated events, while the assigned targets exhibit contraction in 93.4–96.2% of events. At $\lambda = 1$, the two rates coincide at 96.2%, consistent with $\log Q_{\mathrm{tar}}^{t+1|t} = d_t$, which makes the sufficient condition equivalent to target contraction. For $\lambda = 0.5$ and $\lambda = 1.5$, the target contraction rates exceed the condition satisfaction rates, consistent with the condition being sufficient but not necessary. These results show that the theoretical condition holds at most evaluated adjustment events and that FRC typically assigns response targets with smaller log-RFP discrepancies.

B.8 Robustness to Target-Realization Error

Theorem 1 characterizes the modality-level target produced by FRC. Proposition 4 further shows that the tensor-wise learning-rate adjustment approximately realizes this target in function space. We now explicitly separate the target contraction from the remaining realization error.

Define the aggregate functional response produced by the controller-scaled update as

$$\widetilde{R}_m^{t,\mathrm{ctrl}} = \sum_{l \in \mathcal{T}_m} \widetilde{\delta}_{m,l}^{t,\mathrm{ctrl}}. \tag{85}$$

We introduce the positive realization factor

$$\gamma_m^t = \frac{\widetilde{R}_m^{t,\mathrm{ctrl}}}{\eta_{\mathrm{outer}}^t B_m^{t+1}}. \tag{86}$$

Under exact budget realization, $\gamma_m^t = 1$. Proposition 4 implies that γ_m^t is locally close to one when the first-order approximation is accurate and the numerical stabilizer has negligible effect.

Define the reference-normalized ratio of the realized controller response as

$$\widetilde{Q}_{\mathrm{ctrl}}^{t+1|t} = \frac{\widetilde{R}_1^{t,\mathrm{ctrl}}/R_1'^{,t}}{\widetilde{R}_2^{t,\mathrm{ctrl}}/R_2'^{,t}}. \tag{87}$$

Using Eq. (86) and the fact that the common outer learning rate cancels,

$$\widetilde{Q}_{\mathrm{ctrl}}^{t+1|t} = Q_{\mathrm{tar}}^{t+1|t}\frac{\gamma_1^t}{\gamma_2^t}. \tag{88}$$

Let

$$\zeta_t = \log\frac{\gamma_1^t}{\gamma_2^t} \tag{89}$$

denote the relative target-realization error. Combining this expression with Theorem 1 gives

$$\log\widetilde{Q}_{\mathrm{ctrl}}^{t+1|t} = (1-\lambda)\log Q^t + d_t + \zeta_t. \tag{90}$$

Therefore,

$$\left|\log\widetilde{Q}_{\mathrm{ctrl}}^{t+1|t}\right| \le |1-\lambda|\left|\log Q^t\right| + |d_t| + |\zeta_t|. \tag{91}$$

Consequently, if $0 < \lambda < 2$ and

$$|d_t| + |\zeta_t| < (1 - |1-\lambda|)\left|\log Q^t\right|, \tag{92}$$

then the response actually produced by the controller-scaled update satisfies

$$\left|\log\widetilde{Q}_{\mathrm{ctrl}}^{t+1|t}\right| < \left|\log Q^t\right|. \tag{93}$$

This result separates two distinct effects. The term d_t captures the difference between the controller state and the response share measured at the current event, while ζ_t captures the error in realizing the newly assigned response target through tensor-wise learning rates. Both quantities are defined at the current adjustment event and do not require assumptions about the future neural-network state. The strong agreement between assigned and realized responses in Table 4 empirically supports the small-realization-error regime characterized by Proposition 4.

B.9 Extension to Multiple Modalities

The same one-step target analysis extends directly to $M > 2$ modalities. For any pair of modalities i and j, define

$$Q_{ij}^t = \frac{\rho_i^t}{\rho_j^t}, \qquad S_{ij}^t = \frac{s_i^t}{s_j^t}, \qquad \widehat{S}_{ij}^t = \frac{\widehat{s}_i^t}{\widehat{s}_j^t}. \tag{94}$$

From Eq. (46),

$$S_{ij}^{t+1} = S_{ij}^t (Q_{ij}^t)^{-\lambda}. \tag{95}$$

The measured RFP ratio satisfies

$$Q_{ij}^t = \widehat{S}_{ij}^t \frac{R_j'^{,t}}{R_i'^{,t}}. \tag{96}$$

Define the pairwise target ratio as

$$Q_{\mathrm{tar},ij}^{t+1|t} = \frac{B_i^{t+1}/R_i'^{,t}}{B_j^{t+1}/R_j'^{,t}}. \tag{97}$$

Since $B_m^{t+1} = B_{\text{tot}}^t s_m^{t+1}$,

$$Q_{\text{tar},ij}^{t+1|t} = S_{ij}^{t+1} \frac{R_j'^{,t}}{R_i'^{,t}}. \tag{98}$$

Define the pairwise controller-state mismatch

$$d_{ij}^t = \log \frac{S_{ij}^t}{\widehat{S}_{ij}^t}. \tag{99}$$

Following the same derivation as Theorem 1,

$$\log Q_{\text{tar},ij}^{t+1|t} = (1-\lambda) \log Q_{ij}^t + d_{ij}^t. \tag{100}$$

Hence,

$$\left|\log Q_{\text{tar},ij}^{t+1|t}\right| \le |1-\lambda| \left|\log Q_{ij}^t\right| + |d_{ij}^t|. \tag{101}$$

Therefore, for every modality pair (i, j), if $0 < \lambda < 2$ and

$$|d_{ij}^t| < (1 - |1-\lambda|) \left|\log Q_{ij}^t\right|, \tag{102}$$

then

$$\left|\log Q_{\text{tar},ij}^{t+1|t}\right| < \left|\log Q_{ij}^t\right|. \tag{103}$$

Thus, the same FRC update produces pairwise contractive response targets for multiple modalities whenever the corresponding controller-state mismatch is sufficiently small. The allocation rule itself does not change with the number of modalities, and no additional assumption on future neural-network dynamics or inter-event reference stability is required.

C Additional Empirical Analysis

C.1 Datasets

We provide detailed descriptions of the six multimodal benchmarks used in our experiments.

CREMA-D. CREMA-D (Cao et al., 2014) is an audio-visual dataset for speech emotion recognition. It contains 7,442 video clips of approximately 2–3 seconds from 91 actors speaking a collection of short sentences. Following the standard split used in previous multimodal learning studies, we use 6,698 samples for training and 744 samples for testing. The dataset covers six emotion categories: anger, disgust, fear, happiness, neutral, and sadness. Each sample contains synchronized audio and visual modalities.

KSounds. KSounds (Arandjelović & Zisserman, 2017) is an audio-visual action recognition dataset containing 31 action categories that can be recognized from both visual and auditory information. It contains approximately 19,000 10-second video clips collected from YouTube. Following the standard experimental protocol, the dataset is divided into approximately 15,000 clips for training, 1,900 clips for validation, and 1,900 clips for testing. Each video provides synchronized audio and visual modalities.

VGGSound. VGGSound (Chen et al., 2020) is a large-scale audio-visual dataset collected from videos in the wild. It originally contains nearly 200,000 10-second video clips, where the sound-producing object or event is typically visible in the corresponding video. After filtering unavailable videos, we use 168,618 videos for training and validation and 13,954 videos for testing, following the experimental setting adopted in prior multimodal learning studies. Each sample contains paired audio and visual streams.

Twitter. Twitter (Yu & Jiang, 2019) is an image-text dataset for multimodal sentiment classification. It contains paired tweets and associated images with three sentiment labels: positive, negative, and neutral. Following the standard split, we use 3,197 image-text pairs for training, 1,122 pairs for validation, and 1,037 pairs for testing. The two modalities are the tweet text and its corresponding image.

Sarcasm. Sarcasm (Cai et al., 2019) is a multimodal dataset designed for sarcasm detection from paired textual and visual information. Each sample consists of a text-image pair labeled as sarcastic or non-sarcastic. The dataset contains 19,816 pairs for training, 2,410 pairs for validation, and 2,409 pairs for testing, resulting in 24,635 samples in total. The two modalities are the textual content and its associated image.

NVGesture. NVGesture (Molchanov et al., 2016) is a trimodal dataset for dynamic hand-gesture recognition. It contains three input modalities: RGB, optical flow (OF), and depth. The dataset covers 25 gesture categories and contains 1,532 gesture samples in total, including 1,050 training samples and 482 testing samples. Unlike the other benchmarks considered in this work, NVGesture allows us to evaluate FGMO in a setting involving more than two modalities.

C.2 Implementation Details

Response Estimation. We initialize the audio and video response shares to 0.5 and use an adjustment interval of $K = 20$ optimization steps with controller gain $\lambda = 1.0$. The first adjustment occurs at step 20. At each event, we use one Gaussian random projection and the Kronecker-factored estimator with EMA decay $\beta = 0.9$. The estimator statistics are initialized to zero and bias-corrected using $1 - \beta^j$, where $j \geq 1$ counts estimation events. We set $\epsilon = 10^{-8}$ and use a floor of 10^{-30} for the estimator statistics entering logarithmic computations. The probe uses the current training minibatch and centered logits, with the model temporarily in evaluation mode to prevent additional BatchNorm running-statistic updates. The provisional optimizer step advances the optimizer state once; we restore its parameter displacement and apply the recovered update with the assigned learning rates without a second optimizer step.

Reference Construction and Loss Alignment. We train each modality independently for 100 epochs and record its minibatch cross-entropy loss together with the tensor-wise response estimates every 20 optimization steps. The default reference seed is 42, and the resulting trajectories remain fixed during joint training. Matching uses the current modality-specific minibatch loss without additional temporal smoothing. We select the recorded event with the smallest absolute log-loss difference, flooring losses at 10^{-8} before taking logarithms. Matching is performed over the complete recorded trajectory without interpolation or a monotonicity constraint. A loss outside the recorded range is matched to the event with the nearest boundary loss value.

Stability Bounds. After the closed-form share update, we clip the audio share to $[0.2, 0.8]$ and set the video share to its complement. For each tensor, we first clip the proposed learning-rate multiplier relative to the current outer learning rate to $[10^{-4}, 10]$. After multiplication by the outer learning rate, the proposed tensor learning rate is further clipped to between one half and twice its value immediately before the adjustment. These operations are applied in this order. They are additional implementation safeguards: share clipping modifies the unprojected controller update, while learning-rate clipping can introduce deviations from the assigned response targets.

C.3 Controlled Intervention Experiments

We provide the protocol for the stage-wise analysis in Figure 1 and the cross-dataset reliability comparison in Figure 5 (a). Both experiments evaluate whether a modality-level signal can guide beneficial rebalancing from a given training state. We extract the signals from a common vanilla checkpoint, convert them into learning-rate multipliers using one shared mapping, and compare their continuations with a paired vanilla continuation. Only the extracted signals differ; method-specific resampling, auxiliary losses, gradient-direction modifications, and FGMO's tensor-wise allocation are excluded.

Checkpoint Selection. We use the same architectures, data splits, preprocessing, training objective, and optimizer configurations as in the main experiments. For these diagnostic experiments, the vanilla parent trajectories run for 60 epochs, with the base learning rate multiplied by 0.1 after epochs 30 and 45. Checkpoints are selected from the ends of epochs 1–50, leaving a complete ten-epoch continuation horizon for every checkpoint.

For Figure 1, we use one CREMA-D trajectory with seed 1024 and retain all 50 checkpoints. Each checkpoint generates four branches: vanilla and continuations guided by the RGM-derived score, SMV-derived score, and RFP. Each experiment starts independently from its own vanilla checkpoint; interventions are never carried over to the next checkpoint.

Table 7: Checkpoint allocation for the cross-dataset intervention experiment. Counts within each row follow the five parent-training seeds listed in the text.

Dataset	Checkpoints per parent trajectory	Total
CREMA-D	34 / 34 / 33 / 33 / 33	167
KSounds	34 / 34 / 33 / 33 / 33	167
VGGSound	34 / 34 / 33 / 33 / 33	167
Twitter	34 / 34 / 33 / 33 / 33	167
Sarcasm	34 / 33 / 33 / 33 / 33	166
NVGesture	34 / 33 / 33 / 33 / 33	166
Total		1,000

For Figure 5 (a), we evaluate 1,000 checkpoints across the six datasets using five parent-training seeds, $\{2027, 3407, 4096, 5678, 6789\}$. Table 7 specifies the allocation. Within each trajectory, checkpoints are sampled without replacement from early (epochs 1–16), middle (17–33), and late (34–50) stages. A trajectory contributing 33 checkpoints supplies 11 per stage; one contributing 34 supplies 11, 12, and 11, respectively. Sampling uses seed 9001, following the dataset order in the table and the listed seed order. The checkpoint set is fixed before running any interventions. Each checkpoint generates seven branches: vanilla and one branch for each of OGM, PMR, AGM, SMV, RGM, and RFP. All signals use exactly the same checkpoint set, which is separate from the introduction experiment.

Signal Extraction. For each dataset, we fix 256 training examples by class-proportional stratified sampling without replacement using seed 9002. The same examples are used across checkpoints, parent seeds, and signals, and are processed with the dataset's original batch size. Sample-level quantities are averaged before coefficient computation. Validation and test data are not used to construct the signals.

The extracted quantities are the mean ground-truth-class confidence for OGM, mean ground-truth-class prototype probability for PMR, current relative information strength minus its historical reference for AGM, mean sample-level Shapley modality contribution for SMV, cumulative gradient energy (CGE) for RGM, and the response-to-reference ratio in Eq. (5) for RFP. These are signal-derived variants evaluated under the common protocol.

PMR prototypes are recomputed at each checkpoint from a fixed class-stratified 10% training subset, sampled with seed 9003 and containing at least one example per class. AGM's historical statistics and RGM's CGE are maintained along the vanilla parent trajectory and restored at each checkpoint, following their original update and gradient definitions. SMV uses exact subset enumeration for both two- and three-modality datasets. RFP uses the FPE estimator and loss-aligned reference matching described in the method; one unimodal reference per dataset and modality, constructed with seed 9999, is shared across parent-training seeds.

Signal extraction is performed on checkpoint copies. Provisional updates used by FPE leave the continuation's parameters, optimizer states, model buffers, and random-number states unchanged.

Shared Score-to-Coefficient Mapping. Let $q_{m,c}^{(j)}$ denote the score of modality m from signal j at checkpoint c, with M modalities. Scores are oriented so that larger values receive smaller learning-rate multipliers. To remove differences in overall numerical scale, we compute

$$a_c^{(j)} = \frac{1}{M}\sum_{m=1}^{M}\left|q_{m,c}^{(j)}\right|, \qquad \widetilde{q}_{m,c}^{(j)} = \begin{cases} q_{m,c}^{(j)}/a_c^{(j)}, & a_c^{(j)} > 0, \\ 0, & a_c^{(j)} = 0. \end{cases} \tag{104}$$

The common mapping is

$$\alpha_{m,c}^{(j)} = M\frac{\exp\left(-\widetilde{q}_{m,c}^{(j)}/T\right)}{\sum_{n=1}^{M}\exp\left(-\widetilde{q}_{n,c}^{(j)}/T\right)}, \qquad T = 1. \tag{105}$$

The temperature is fixed across signals, datasets, and checkpoints. This rule satisfies $\sum_m \alpha_{m,c}^{(j)} = M$, fixing the average modality multiplier at one. It does not preserve the total gradient norm or

function-space response. Equal modality scores yield unit multipliers, while reversing the score ordering reverses the intervention direction. For NVGesture, the rule is applied jointly to all three modalities.

Paired Continuations. Coefficients are computed once at the starting checkpoint and held fixed for ten continuation epochs. For every encoder or classifier tensor $l \in \mathcal{T}_m$, we set

$$\eta_{m,l,c}^{(j)}(u) = \alpha_{m,c}^{(j)} \eta_{m,l}^{\text{base}}(k_c + u), \tag{106}$$

where k_c is the parent checkpoint's optimization step, u indexes continuation updates, and η^{base} is the inherited learning-rate schedule. The multiplier is applied through optimizer parameter-group learning rates and is shared by all tensors within a modality. Shared parameters, if present, retain their base learning rates. Vanilla uses unit multipliers.

All branches restore the same model parameters and buffers, optimizer moments and counters, scheduler state, and AMP scaler state when applicable. The scheduler continues from the parent training position. With checkpoint identifier c, paired branches use continuation seed $100000 + c$ and matched minibatch order, data augmentation, and dropout streams. Diagnostic computations use separate random-number streams. Each branch is run once, with no early stopping or intermediate model selection.

Consequently, identical coefficient vectors at the same checkpoint define identical interventions under matched stochastic inputs. In particular, unit multipliers reproduce vanilla, up to implementation nondeterminism. Coefficients are not refreshed during continuation, so this experiment evaluates the guidance available at the starting checkpoint.

Evaluation and Beneficial Rate. We evaluate the fused model after exactly ten continuation epochs on the dataset's fixed test split. Let $A_c^{(j)} \in [0, 1]$ and $A_c^{(0)}$ denote the endpoint accuracies of the signal-guided and paired vanilla branches. The intervention gain, reported as ΔAcc in the figures, is

$$\Delta A_c^{(j)} = 100 \left(A_c^{(j)} - A_c^{(0)} \right). \tag{107}$$

Gains are measured in percentage points. An intervention is beneficial if and only if $\Delta A_c^{(j)} > 0$, using unrounded accuracies. Zero and negative gains remain in the denominator and are counted as non-beneficial.

For the cross-dataset experiment, the pooled Beneficial Rate is

$$\text{BR}^{(j)} = \frac{100}{N} \sum_{c=1}^{N} \mathbf{1}\left[\Delta A_c^{(j)} > 0 \right], \qquad N = 1{,}000. \tag{108}$$

Thus, BR measures the frequency of positive intervention outcomes under the specified common mapping. All signals share the same denominator. In Figure 1, score curves are measured on the parent vanilla trajectory, and each bar reports the gain from the corresponding checkpoint's paired continuation. Checkpoints from the same trajectory are treated as correlated training states rather than independent training seeds.

C.4 Cross-Seed Stability of Unimodal References

We further evaluate the sensitivity of FGMO to the random seed used for constructing the unimodal reference. As shown in Table 8, RFP estimates and tensor-wise allocations remain highly consistent across different reference seeds, indicating that the loss-aligned unimodal reference is stable across random initializations. Notably, all multi-seed experiments in this paper use the same fixed reference trajectory and achieve stable performance across training seeds. These results suggest that FGMO does not rely on a particular unimodal reference seed.

C.5 Dynamic Modality Allocation

We further examine how FRC dynamically adjusts the modality response shares during joint training. As shown in Figure 6, we report the evolution of s_1 at consecutive RFP measurement events for the Image modality on Sarcasm and the Audio modality on CREMA-D. Rather than converging to a fixed allocation or following a monotonic schedule, both trajectories are continuously readjusted

throughout training. On Sarcasm, the Image share stays close to the initial balanced allocation in the early stage, then increases substantially at several intermediate and later measurement events before being reduced again as training progresses. A similar adaptive behavior can be observed on CREMA-D, where the Audio share first decreases below the balanced allocation, subsequently recovers, and continues to fluctuate in response to the evolving optimization state. Importantly, the two datasets exhibit clearly different allocation trajectories, indicating that FRC does not impose a predefined modality preference or a fixed reweighting schedule.

These variations directly reflect the feedback mechanism of FRC. At each measurement event, RFP characterizes the relative functional progress of each modality with respect to its loss-aligned unimodal reference. FRC then updates the response shares according to the current RFP values, assigning a larger share to modalities with relatively slower functional progress while reducing the share of those progressing relatively faster. Meanwhile, the KL regularization constrains abrupt changes from the previous allocation, allowing the controller to respond to evolving modality states without making excessively aggressive corrections. Consequently, the repeated rises, decreases, and reversals show that the preferred allocation can change substantially across training stages. This behavior further supports the motivation of FGMO: modality imbalance is not a static property that can be addressed by permanently strengthening or suppressing one modality, but an evolving optimization state that requires stage-dependent feedback and continuous reallocation of the function-space response budget.

Table 8: Stability analysis of cross-seed unimodal references.

Modality	Seed Pair (Eval / Ref)	Progress Shift		Allocation Cosine Similarity		
		Mean	95% Pct.	Median	5% Pct.	≥ 0.90
Audio	2027 / 1024	0.84%	2.86%	0.982	0.958	100.0%
Audio	3407 / 1024	0.71%	1.76%	0.974	0.936	99.4%
Video	2027 / 1024	0.75%	2.43%	0.992	0.968	100.0%
Video	3407 / 1024	2.10%	5.72%	0.989	0.963	99.4%
Overall (4,000 queries)		1.10%	3.34%	0.986	0.954	99.7%

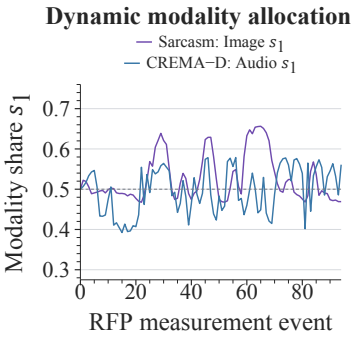


Figure 6: Dynamic modality allocation trajectories on Sarcasm and CREMA-D datasets.

C.6 Comparison of Progress Measurement Spaces

To isolate the effect of the measurement space from that of reference calibration, we compare four relative progress signals constructed from prediction, gradient, optimizer-update, and function-space measurements. For all variants, the current modality-wise quantity is normalized by its corresponding unimodal quantity at the same loss-aligned reference state. We use the same checkpoint set, score-to-coefficient mapping, paired continuation protocol, and Beneficial Rate criterion described in Appendix C.3. Thus, the only difference among the four variants is the quantity used to characterize modality progress.

Relative Progress Signals. For each modality m at checkpoint c, we first determine the loss-aligned unimodal state τ_m^c using Eq. (4). We then construct the relative progress signals in different measurement spaces using the current quantity and its corresponding unimodal reference.

For the prediction-space variant, we use the mean ground-truth-class probability over the fixed signal-extraction subset:

$$P_m^c = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} p_{i,y_i}^{m,c}, \qquad q_{m,c}^{\text{pred}} = \frac{P_m^c}{P_m^{\text{s}}(\tau_m^c)}. \tag{109}$$

Table 9: Comparison of progress measurement spaces under the same loss-aligned reference calibration and controlled intervention protocol. Each entry reports the Beneficial Rate (%), i.e., the percentage of checkpoints where the guided continuation improves over its paired vanilla continuation.

Measurement Space	CREMA-D	KSounds	VGGSound	Twitter	Sarcasm	NVGesture	Overall BR ↑
Prediction-relative	71.3	68.9	66.5	73.1	74.1	69.9	70.6
Gradient-relative	74.9	72.5	70.1	73.7	75.9	72.3	73.2
Update-relative	77.8	75.4	73.7	**78.4**	77.1	75.3	76.3
Function-relative	**85.0**	**83.8**	**82.6**	77.2	**86.7**	**86.7**	**83.7**

For the gradient-space variant, we aggregate the Euclidean norms of the task gradients over all tensors belonging to modality m:

$$G_m^c = \sum_{l \in \mathcal{T}_m} \left\| \nabla_{\mathbf{\Theta}_{m,l}} \mathcal{L} \right\|_2, \qquad q_{m,c}^{\text{grad}} = \frac{G_m^c}{G_m^{\text{s}}(\tau_m^c)}. \tag{110}$$

For the update-space variant, we use the magnitude of the unit-learning-rate optimizer update recovered from the provisional step:

$$U_m^c = \sum_{l \in \mathcal{T}_m} \left\| \mathbf{u}_{m,l}^c \right\|_2, \qquad q_{m,c}^{\text{update}} = \frac{U_m^c}{U_m^{\text{s}}(\tau_m^c)}. \tag{111}$$

For the function-space variant, we aggregate the functional responses defined in Eq. (3):

$$R_m^c = \sum_{l \in \mathcal{T}_m} \delta_{m,l}^c, \qquad q_{m,c}^{\text{func}} = \frac{R_m^c}{R_m^{\text{s}}(\tau_m^c)} = \rho_m^c. \tag{112}$$

Therefore, all four variants use the same loss-aligned unimodal reference and the same relative normalization, while differing only in the space used to measure modality progress. The resulting scores are converted into modality-wise learning-rate multipliers using the shared mapping in Eq. (105). We then apply the same paired ten-epoch continuations and compute Beneficial Rate according to Eq. (108).

As shown in Table 9, prediction-relative and gradient-relative measurements achieve overall Beneficial Rates of 70.6% and 73.2%, respectively. Using the actual optimizer update further increases the overall Beneficial Rate to 76.3%, suggesting that incorporating optimizer dynamics provides a more informative estimate of the current optimization state than prediction or gradient statistics alone.

Function-relative progress achieves the highest overall Beneficial Rate of 83.7%, outperforming update-relative, gradient-relative, and prediction-relative measurements by 7.4, 10.5, and 13.1 percentage points, respectively. It achieves the highest Beneficial Rate on five of the six datasets, while update-relative performs slightly better on Twitter.

Since all variants use the same loss-aligned reference, relative normalization, score-to-coefficient mapping, checkpoint set, and continuation protocol, the observed differences cannot be attributed to reference calibration or different rebalancing rules. Instead, they isolate the effect of the measurement space. Prediction statistics characterize the current model outputs, while gradient and optimizer-update magnitudes characterize changes in parameter space. In contrast, the function-space response measures how the optimizer update changes the model output, thereby incorporating both the update and the local input–output geometry. These results support function space as a more reliable measurement space for estimating relative modality progress.