CAT-GS: Balanced Multimodal Learning via Calibrated Gating and Fusion Surgery

Mahir Shahriar Tamim[a,*], Sharjil Khan[a], Md. Samiul Alim[a], Tanvir Ahmed Khan[a], Shafin Rahman[a], Nabeel Mohammed[a]

[a]*Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh*

Abstract

End-to-end training of multimodal neural networks often exhibits unstable neural dynamics characterized by three coupled failure modes that degrade learning: (i) *modality imbalance*, where one branch dominates gradient-based optimization; (ii) *unstable gating*, where noisy confidence cues induce erratic modality selection; and (iii) *fusion interference*, where modality-specific gradients conflict at the shared fusion layer. We propose **CAT-GS** (*Calibrated, Adaptive, Thresholded Gating with Fusion Surgery*), a neural dynamics-based optimization controller for intelligent computing applications. CAT-GS operates during backpropagation without modifying model architectures, fusion modules, or task losses. Through calibration of teacher-derived reliability via temperature scaling and EMA smoothing, CAT-GS stabilizes neural dynamics using a margin-thresholded policy to switch between warm-up dropout, weak-modality prioritization, and weak-biased blending, stabilizes gradient magnitudes under aggressive gating via capped gradient-budget renormalization, and applies *fusion-only* PCGrad to reduce destructive cross-modal interference at the primary shared bottleneck. We evaluate CAT-GS on audio–visual multimodal pattern recognition benchmarks (CREMA-D, AV-MNIST, and VGGSound), a tri-modal setting (UR-FUNNY), controlled synthetic data (CG-MNIST), and additional cross-domain benchmarks (AVE and CMU-MOSI). CAT-GS improves or matches fused multimodal accuracy against strong imbalance-aware baselines (including OGM-GE, G^{2}D, and UMT) across settings, and yields smoother gating behavior with fewer conflicting fusion gradients.

Keywords: Neural dynamics, Multimodal learning, Intelligent computing, Gradient-based optimization, Learning dynamics, Knowledge distillation, Calibration, Pattern recognition

1. Introduction

Deep multimodal learning integrates heterogeneous signals (e.g., audio/speech and vision) within neural network architectures to improve pattern recognition and robustness beyond any single modality [1]. From a neural dynamics perspective, intelligent computing systems require stable temporal evolution of gradient flows and adaptive control mechanisms throughout training. CAT-GS is a neural dynamics-based optimization controller that operates during backpropagation to regulate learning dynamics without modifying model architectures, fusion modules, or task losses. In an ideal system, the model dynamically relies on the most reliable modality *for the current sample and training stage*. In practice, however, joint optimization is frequently derailed by the *modality imbalance phenomenon* [2]: one branch (often the easier or higher-SNR modality) quickly dominates gradient flow, while the weaker branch receives progressively smaller updates. The result is brittle fusion, under-trained modality-specific encoders, and reduced multimodal gains. Figure 1 illustrates these failure modes on CREMA-D, highlighting how dominant-modality imbalance and fusion-gradient conflicts can derail convergence under standard joint training.

Existing approaches mitigate imbalance primarily through two families of techniques. *Gradient-modulation methods* attenuate the dominant modality's gradients to amplify weaker ones [2, 3], while *confidence-based gating methods* down-weight or mask unreliable modalities based on calibration-aware confidence estimates [4]. Although effective in isolation, these strategies fail to address three tightly-coupled challenges that arise in the learning dynamics of end-to-end multimodal optimization. First, confidence signals used for modality gating are typically derived from raw logits, which are frequently miscalibrated and noisy [5]. Without temporal stabilization, these unreliable estimates induce sharp batch-to-batch fluctuations in gating decisions, leading to "gate thrashing" rather than reliability-aware control. Second, aggressive or persistent suppression of a modality can result in *gradient starvation*: when a branch receives near-zero gradients over extended periods, its representations stagnate relative to the dominant modality, reinforcing the imbalance and making recovery increasingly difficult [6]. Third, even when modality-specific gradients are individually balanced, their interaction in the shared fusion layer can remain destructive; gradients with negative cosine similarity [7] interfere at the bottleneck, impeding joint convergence despite otherwise well-behaved unimodal updates.

To address these limitations, we introduce **CAT-GS**, a neural dynamics-based optimization framework for intelligent computing applications that provides a unified treatment of these coupled failure modes for calibrated, balanced, and conflict-aware multimodal learning. CAT-GS wraps an existing multimodal training loop and modifies *only* the backward-stage optimization signals. It uses a *margin-thresholded controller* driven by stabilized teacher reliability scores to switch between warm-

*Corresponding author

Email address: mahir.tamim@northsouth.edu (Mahir Shahriar Tamim)

up dropout, dominance suppression, and weak-biased blending. This produces interpretable regime transitions and avoids oscillatory gating behavior. Empirically, CAT-GS improves fused multimodal accuracy in diverse settings, achieving gains over strong imbalance-aware baselines (e.g., in the UR-FUNNY A–V–TXT task; Table 2), while also exhibiting smoother gating dynamics and fewer negative fusion-gradient cosine events (Figure 4). Unlike prior work that addresses imbalance, gating, or conflict resolution in isolation, CAT-GS couples calibrated reliability estimation via knowledge distillation, regime-aware gating, budget-preserving gradient renormalization, and localized conflict resolution within a single optimization step.

Novelty and contributions. We introduce CAT-GS, an optimization-stage multimodal backward-pass controller that jointly addresses gate instability, modality starvation, and fusion interference within a single optimization step. The building blocks draw on established techniques; the two new elements are restricting PCGrad to the fusion bottleneck rather than the whole network, and a budget-preserving renormalization that prevents a gated modality from starving. Our contributions are as follows:

- We cast adaptive modality gating as a *regime-based gating controller* driven by calibrated reliability margins from knowledge distillation, enabling interpretable transitions between warm-up dropout, dominance suppression, and weak-biased blending, without modifying model architectures or task losses.
- We couple gating with a *budget-preserving gradient renormalization mechanism* that stabilizes update magnitudes under aggressive (hard) gating using an EMA target magnitude with a hard cap, mitigating failure modes such as prolonged imbalance and brittle convergence.
- We adopt a *fusion-only gradient surgery strategy*, applying PCGrad exclusively at the shared fusion bottleneck where cross-modal gradients first interact, thereby reducing destructive interference while preserving unimodal encoder dynamics at substantially lower cost than global projection.
- We provide extensive empirical validation across audio–visual, tri-modal, and controlled synthetic benchmarks, plus additional AVE and CMU-MOSI evaluations, demonstrating consistent improvements in fused accuracy, optimization stability, and robustness to modality imbalance.

2. Related Work

Modality Imbalance in Multimodal Learning. The dominance of specific modalities during joint training is a pervasive issue in multimodal learning. Early efforts address this through auxiliary losses or feature norm regularization [2]. More recent gradient-modulation techniques, such as on-the-fly rescaling [2] and multi-loss balancing [8], dynamically adjust updates to prevent one modality from overpowering others. Recent balancing-oriented methods include MMCosine [9], which uses

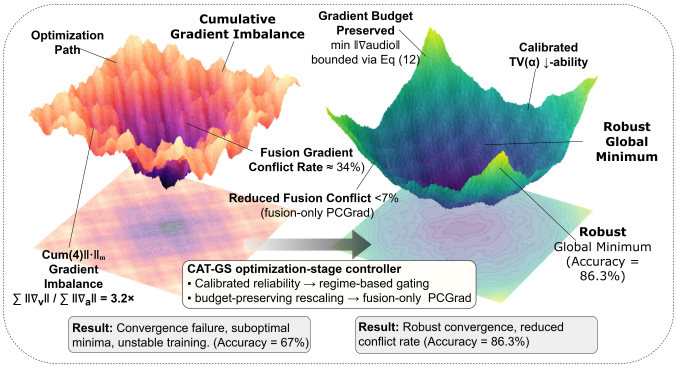

Figure 1: **Optimization dynamics on CREMA-D.** Shown are experimentally derived visualizations of training behavior under standard joint training (left) and CAT-GS (right). Annotations are illustrative summaries of observed gradient statistics. Joint training exhibits unstable dynamics with dominant-modality imbalance and frequent fusion-gradient conflicts, converging to a suboptimal solution (≈67% accuracy). CAT-GS stabilizes optimization via calibrated gating, gradient-budget preservation, and fusion-only gradient surgery, yielding smooth convergence to a robust solution (86.3% accuracy).

cosine-based feature normalization for improved fine-grained discriminability, DRL [10], which diagnoses modality learning states and applies soft re-learning, and MCR [11], which introduces game-theoretic regularization to mitigate modality competition. While these methods improve balance, many still rely on predefined or coarse dominance cues and may not explicitly prevent long-horizon gradient starvation in suppressed branches [12].

In contrast, CAT-GS leverages calibrated confidence margins for context-aware modulation, ensuring suppression is temporary and paired with budget-preserving rescaling.

Confidence-Based Gating and Dynamic Fusion. Multimodal fusion has been studied across a wide range of tasks, from sentiment analysis that combines audio, visual, and textual cues [13] to video captioning via hierarchical attention-based fusion [14]. Gating mechanisms have long enabled adaptive fusion in multimodal systems [15]. Contemporary approaches use soft attention [16] or entropy-driven gates to weigh modalities by reliability. Teacher-guided fusion further enhances this by distilling unimodal priors [17]. However, reliance on uncalibrated teacher logits leads to unstable decisions [5], exacerbating gate thrashing. Few works integrate direct calibration into the gating loop; CAT-GS advances this by combining temperature scaling, EMA smoothing, and margin thresholding for robust, regime-aware control.

Gradient Conflict Resolution. Gradient conflicts hinder multitask and multimodal optimization [18]. Projection-based methods like PCGrad and CAGrad excise destructive components, while others normalize via variance reduction. In multimodal settings, conflicts are acute at fusion layers [18], yet prior applications are global rather than localized. To the best of our knowledge, prior work has not systematically explored applying PCGrad selectively at the fusion head; CAT-GS adopts this localized strategy to preserve unimodal encoder dynamics while resolving bottleneck interference.

Knowledge Distillation in Multimodal Settings. Distillation

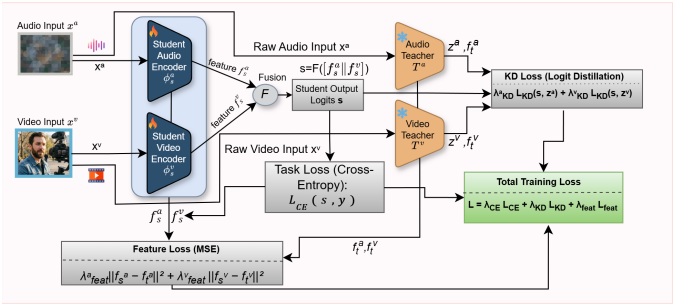

Figure 2: **High-level forward architecture and supervision flow.** Audio and video inputs are processed by trainable student encoders and fused to produce predictions. Frozen unimodal teachers receive the same raw modality inputs (x^a, x^v) directly (not student-derived features) and remain frozen throughout training; they provide auxiliary supervision via knowledge distillation and per-modality reliability signals utilized by CAT-GS. Optimization-stage control (CAT-GS) operates exclusively during backpropagation and is shown separately in Figure 3.

from unimodal teachers is standard for multimodal students [17, 19]. Methods like UMT [17] match logits and features but treat teacher outputs as infallible, ignoring miscalibration [5]. Recent calibration-aware variants focus on post-hoc adjustments, not online integration. CAT-GS embeds temperature scaling and EMA stabilization within the distillation-guided gating process, yielding reliable signals for adaptive optimization.

In summary, prior work typically tackles imbalance, confidence-based fusion, and gradient conflicts in isolation. CAT-GS integrates these ideas into a single optimization-stage controller, yielding a practical recipe for stable and balanced multimodal training across fusion designs. The contribution is therefore the optimization-stage integration itself, together with the fusion-localized projection and budget-preserving renormalization it introduces.

3. Methodology

Figures 2 and 3 summarize the overall model and the CAT-GS training iteration. Figure 2 shows the forward architecture and supervision flow (trainable student encoders and fusion head with frozen unimodal teachers), while Figure 3 details the optimization-stage controller: calibrated/EMA-stabilized teacher reliabilities drive regime-based gating, followed by gradient-budget stabilization and fusion-only surgery before the optimizer update. For readability, we present CAT-GS in a strict control flow: reliability estimation (Section 3.3), regime selection and gating (Section 3.4), magnitude stabilization (Section 3.5), and fusion-conflict handling (Section 3.6), followed by the unified training step in Algorithm 1.

In this section we first formalize the multimodal learning setup and training objective, and then introduce CAT-GS as an optimization-level controller that operates on the resulting gradients.

3.1. Problem Formulation

We consider a multimodal dataset $\mathcal{D} = \{(x_i^a, x_i^v, y_i)\}_{i=1}^B$ containing paired audio–visual samples and labels. The student model follows a standard multimodal fusion design and comprises two unimodal encoders $E_a(\cdot;\theta_a)$ and $E_v(\cdot;\theta_v)$, which extract modality-specific features, and a fusion network $F(\cdot;\theta_f)$ that maps the concatenated representations to final logits. Two unimodal teacher models T^a and T^v provide auxiliary guidance in the form of logits and intermediate features. This setup is intentionally generic: the encoders may be CNNs or transformers, and the fusion head may be an MLP or attention module, since CAT-GS assumes no architectural constraints.

Given a batch, the student produces

$$f_s^a = E_a(x^a), \qquad f_s^v = E_v(x^v), \qquad \mathbf{s} = F([f_s^a \| f_s^v]). \tag{1}$$

Training objective. CAT-GS is objective-agnostic, but in our experiments we optimize a standard supervised loss augmented with unimodal-teacher distillation terms. Let $\mathcal{M}$ denote the set of available modalities (e.g., $\{a, v\}$ or $\{a, v, t\}$). For a labeled batch, the student is trained with

$$\begin{aligned}\mathcal{L} = \lambda_{\text{CE}}\,\mathcal{L}_{\text{CE}}(\mathbf{s}, y) + \sum_{m\in\mathcal{M}} \lambda_{\text{KD}}^m\,\mathcal{L}_{\text{KD}}(\mathbf{s}, \mathbf{z}^m, \tau_m) \\ + \sum_{m\in\mathcal{M}} \lambda_{\text{feat}}^m\,\mathcal{L}_{\text{feat}}(f_s^m, f_t^m).\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy, f_t^m and $\mathbf{z}^m$ are the unimodal teacher features and logits, and

$$\mathcal{L}_{\text{KD}}(\mathbf{s}, \mathbf{z}^m, \tau_m) = \tau_m^2\,\text{KL}(\text{Softmax}(\mathbf{z}^m/\tau_m) \,\|\, \text{Softmax}(\mathbf{s}/\tau_m)), \tag{3}$$

with $\mathcal{L}_{\text{feat}}$ implemented as MSE. The coefficients $\{\lambda\}$ and temperatures $\{\tau_m\}$ match the training code and are set per experiment; when a distillation term is not used, its weight is set to zero.

Any differentiable training objective defined over $\mathbf{s}$, the labels y, and optionally the teacher outputs $\{T^m(x^m)\}_{m\in\{a,v\}}$ induces gradients with respect to the modality-specific parameters θ_a, θ_v and the fusion parameters θ_f. In a conventional setting, these gradients are aggregated and applied uniformly across modalities during backpropagation, which is precisely where modality imbalance can arise: one branch may dominate the updates, while the other receives progressively weaker signals. CAT-GS intervenes at this optimization stage. Rather than changing the architecture or the underlying training objective, it modulates and stabilizes modality-specific and fusion-layer gradients, aiming to ensure balanced, reliability-aware updates throughout training.

3.2. CAT-GS Overview

Multimodal models frequently suffer from *modality imbalance*, where one modality converges faster or produces cleaner gradients and consequently dominates optimization. As training progresses, gradients for the weaker modality diminish, feature representations diverge, and the fusion layer receives conflicting update signals, ultimately leading to brittle and suboptimal fusion. CAT-GS addresses this issue at the *optimization level*. Rather than modifying architectures or introducing task-specific losses, CAT-GS restructures how gradients are weighted, stabilized, and combined during training. It operates transparently between `loss.backward()` and `optimizer.step()`, and can

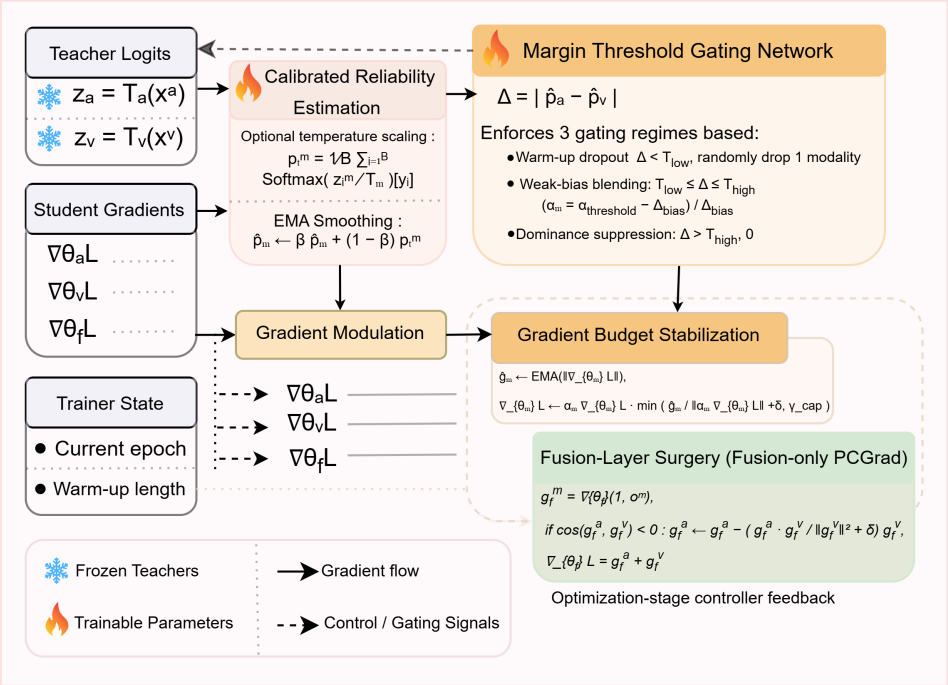

Figure 3: **Overview of CAT-GS.** The student model contains trainable unimodal encoders and a fusion head, while unimodal teachers are frozen and used only to estimate per-modality reliability. Teacher confidences are calibrated and stabilized to drive margin-threshold gating, followed by gradient-budget stabilization and fusion-only PCGrad.

be integrated into any multimodal model with modality-specific parameters. The framework consists of four complementary components:

1. **Calibrated reliability estimation** to produce stable modality confidence signals;
2. **Margin-thresholded adaptive gating** to dynamically prioritize modalities;
3. **Gradient-budget stabilization** to control update magnitudes under aggressive gating;
4. **Fusion-layer gradient surgery** to remove destructive cross-modal conflicts.

These components are jointly necessary. Calibration stabilizes gating decisions, adaptive gating prevents runaway dominance, budget reallocation stabilizes update magnitudes under aggressive gating, and fusion surgery resolves residual gradient interference at the bottleneck. Together, they yield a stable optimization dynamic that adapts to evolving modality reliability without altering the forward computation.

3.3. Calibrated Reliability Estimation

Teacher logits encode useful task knowledge but can be noisy and overconfident, especially early in training. Since CAT-GS relies on teacher reliability as a control signal, we stabilize these estimates to reduce batch-to-batch volatility.

Temperature scaling. For the teacher T_m with logits z^m, we compute the average probability assigned to the ground-truth label under a temperature T_m:

$$p_t^m = \frac{1}{B} \sum_{i=1}^{B} \text{Softmax}(z_i^m / T_m)[y_i], \tag{4}$$

where T_m controls the sharpness of the confidence signal. In our main experiments, we set T_m=1 for all modalities (i.e., no temperature scaling) unless explicitly stated. We include temperature scaling as an optional extension when a calibrated temperature is available, but it is not required for CAT-GS. Thus, temperature scaling is not a required performance driver in our default setting; it is an optional calibration switch and a stress-test factor in Section 7. The resulting scalar p_t^m is a *batch-level* (step-level) proxy for modality reliability; we use the batch mean for stability and to avoid noisy per-sample gating signals.

EMA smoothing. To prevent noisy fluctuations from batch to batch, we maintain:

$$\hat{p}^m \leftarrow \beta\, \hat{p}^m + (1 - \beta)\, p_t^m. \tag{5}$$

The smoothed confidences $\hat{p}^a, \hat{p}^v$ are used as the continuous control signal for gating, while temperature scaling remains an optional calibration step (default T_m=1 in our main experiments). This reduces "gate thrashing" and enables regime decisions based on reliability trends rather than instantaneous noise. In practice, we use a high momentum (e.g., $\beta = 0.9$) so that the controller reacts gradually rather than abruptly.

3.4. Margin-Thresholded Adaptive Gating

CAT-GS uses the smoothed reliabilities to compute a *reliability margin*:

$$\Delta = |\hat{p}^a - \hat{p}^v|. \tag{6}$$

This margin measures how far apart the modalities are in terms of calibrated confidence. Two thresholds τ_{low} and τ_{high} define three regimes reflecting distinct optimization needs. Crucially, these regimes determine the gating coefficients α_m that will later be used to scale gradients flowing through each modality. For a compact view, the gating controller in this subsection can be summarized as one regime map:

$$\alpha_m(\Delta) = \begin{cases} \alpha_m^{\text{drop}}, & \Delta < \tau_{\text{low}} \text{ and epoch} < E_w, \\ \alpha_m^{\text{hard}}, & \Delta > \tau_{\text{high}}, \\ \alpha_m^{\text{soft}}, & \tau_{\text{low}} \le \Delta \le \tau_{\text{high}}, \end{cases} \tag{7}$$

1. Warm-up Dropout ($\Delta < \tau_{low}$, early epochs)

Early teacher signals are noisy, so CAT-GS uses stochastic modality dropout (probability p_{drop}) for a short warm-up to prevent premature specialization and to strengthen unimodal representations.

2. Dominance Suppression ($\Delta > \tau_{high}$)

When one modality becomes clearly more reliable, CAT-GS protects the weaker one by performing hard gating:

$$\alpha_m = \begin{cases} 1, & \hat{p}^m < \hat{p}^{\neg m}, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The weaker branch receives the full update emphasis during this regime, counteracting runaway dominance. This hard regime is used only when the margin is clearly large; otherwise the controller uses soft blending.

3. Weak-Bias Blending ($\tau_{low} \le \Delta \le \tau_{high}$)

When reliabilities are close, CAT-GS uses smooth weak-biased blending rather than abrupt switching:

$$\tilde{\alpha}_m = \frac{\hat{p}^m}{\hat{p}^a + \hat{p}^v}, \tag{9}$$

then boosts the weaker modality using a bias factor λ_{bias}:

$$\alpha_{weak} = \tilde{\alpha}_{weak}(1 + \lambda_{bias}), \quad \alpha_{strong} = \tilde{\alpha}_{strong}(1 - \lambda_{bias}), \tag{10}$$

Weights are then normalized with floor ε so both modalities remain active while the weaker branch receives a controlled advantage.

Together, the three regimes define an interpretable controller that outputs $\alpha_a, \alpha_v \in [0, 1]$ for gradient modulation.

$$\nabla_{\theta_m} \leftarrow \alpha_m \nabla_{\theta_m}. \tag{11}$$

3.5. Gradient-Budget Reallocation

Directly applying α_m scales gradients but does not guarantee stable optimization dynamics under aggressive gating. A fundamental risk in gating mechanisms is *gradient starvation*: if a modality is consistently down-weighted (or hard-gated), its encoder can receive near-zero gradients over extended periods. This causes a vicious cycle where the suppressed modality's features stagnate, making it even less reliable in future epochs, effectively permanently disabling a branch of the network.

CAT-GS mitigates these dynamics by (i) enforcing a nonzero floor ε in soft regimes (Section 3.4), and (ii) stabilizing gradient magnitudes during hard-gating events via *gradient-budget reallocation*. We track the historical gradient norm (per modality) as a running reference scale:

$$\hat{g}^m \leftarrow \beta_g \hat{g}^m + (1 - \beta_g)\|\nabla_{\theta_m}\|. \tag{12}$$

After gating, we rescale gradients:

$$\nabla_{\theta_m} \leftarrow \alpha_m \nabla_{\theta_m} \ \min\left(\frac{\hat{g}^m}{\|\alpha_m \nabla_{\theta_m}\| + \delta}, \ \gamma_{\text{cap}}\right). \tag{13}$$

Optimization view: Gating alone makes encoder updates scale linearly with α_m, so when $\alpha_m \approx 0$ for many steps, the expected update magnitude collapses and the modality may not recover. CAT-GS instead uses $\hat{g}^m$ as a running reference scale and (when a modality is hard-gated) renormalizes the *active* modality's encoder gradients toward its EMA magnitude, capped by γ_{cap}, to avoid brittle update collapse or instability. Equivalently, Step 5 approximately enforces $\|\nabla_{\theta_m}\| \approx \min(\hat{g}^m, \gamma_{\text{cap}} \|\tilde{\nabla}_{\theta_m}\|)$, which stabilizes the magnitude of the applied update while avoiding exploding rescaling when the current gradient is extremely small.

This stabilizes the overall optimization dynamics under aggressive gating. When a modality is fully gated ($\alpha_m = 0$), its encoder receives no update by design; CAT-GS focuses on keeping the remaining active branch well-conditioned while soft regimes maintain nonzero updates via the ε floor. The cap γ_{cap} prevents excessively large rescaling when the current gradient is extremely small, and δ avoids numerical instability.

3.6. Fusion-Layer Gradient Surgery

Even when unimodal gradients are balanced, fusion parameters often receive *conflicting* update directions. Audio may encourage temporal invariances while visual gradients push toward spatial distinctions, producing destructive interference. While methods like PCGrad [7] address this globally, applying projection to the entire network is computationally expensive ($O(N^2)$ in the number of tasks/modalities) and may interfere with the specialized feature extraction of unimodal encoders.

We identify the fusion output layer as the critical *informational bottleneck* where these semantic conflicts are most destructive. Therefore, CAT-GS applies a surgical, localized PCGrad-like projection only to the fusion output parameters (in our implementation, the final fusion classifier layer, e.g., `fc_out`).

Let θ_f denote the parameters of the fusion output layer, and let $\mathbf{o}^a$ and $\mathbf{o}^v$ be the modality-specific fusion outputs (logits) obtained by passing only modality-m features through the fusion

head. In the current implementation, we form modality-specific fusion gradients via a Jacobian–vector product using an all-ones vector (equivalently, the gradient of the sum of logits):

$$\mathbf{g}_f^m = \nabla_{\theta_f} \langle \mathbf{1}, \mathbf{o}^m \rangle, \qquad m \in \{a, v\}. \tag{14}$$

$$\text{If} \quad \cos(\mathbf{g}_f^a, \mathbf{g}_f^v) < 0, \tag{15}$$

the gradients point in conflicting directions. CAT-GS applies a PCGrad-style projection (single-sided in the current implementation):

$$\mathbf{g}_f^a \leftarrow \mathbf{g}_f^a - \frac{\mathbf{g}_f^a \cdot \mathbf{g}_f^v}{\|\mathbf{g}_f^v\|^2 + \delta} \mathbf{g}_f^v, \tag{16}$$

The filtered gradients are then summed to form the fusion-layer update, and the fusion layer's gradient is overwritten with $\mathbf{g}_f^a + \mathbf{g}_f^v$. This approach resolves the "trilemma" component of gradient interference efficiently, ensuring that the shared decision boundary respects the geometric requirements of both modalities without the overhead of global gradient projection.

Why fusion-only PCGrad. Gradient conflicts in multimodal networks arise predominantly at shared representational bottlenecks, where modality-specific signals compete to shape a common decision boundary. Applying projection-based methods such as PCGrad globally treats unimodal encoders as competing tasks, which can inadvertently suppress modality-specific feature learning and increase computational overhead. CAT-GS instead applies PCGrad-like projection only at the fusion output layer, where cross-modal gradients first interact. This localized projection preserves unimodal encoder gradients while explicitly enforcing cooperative geometry at the shared fusion parameters, yielding stable convergence at substantially lower cost than full-network gradient projection.

Cost perspective. For M modalities, projection-based conflict handling is pairwise ($O(M^2)$). Applying it globally scales with the full parameter count $|\theta|$, whereas fusion-only surgery scales with the fusion head size $|\theta_f|$ (typically $|\theta_f| \ll |\theta|$): $O(M^2|\theta|)$ vs. $O(M^2|\theta_f|)$. This substantially reduces overhead and avoids unnecessarily constraining unimodal encoder learning.

3.7. Extension to M Modalities and Unified Training Pipeline

CAT-GS extends to M>2 modalities by computing per-modality stabilized reliabilities $\{\hat{p}^m\}_{m\in\mathcal{M}}$ and using their spread to trigger the same regime policy. For the general M-modality setting, we replace the two-modality reliability gap of Eq. (6) with the max–min confidence margin over all available modalities:

$$\Delta_M = \max_{m\in\mathcal{M}} \hat{p}^m - \min_{m\in\mathcal{M}} \hat{p}^m, \tag{17}$$

and identify $m^- = \arg\min_m \hat{p}^m$ and $m^+ = \arg\max_m \hat{p}^m$. The same CAT-GS controller logic is used across datasets; only the number of modalities and corresponding branch/group definitions are adapted. The tri-modal UR-FUNNY benchmark (M=3, with $\mathcal{M} = \{a, v, t\}$) is one instance of this general formulation; the controller itself is not dataset-specific. The gating coefficients $\{\alpha_m\}$ are computed by: (i) warm-up dropout (optionally keep one random modality), (ii) dominance mode if $\Delta_M > \tau_{\text{high}}$ (set α_{m^-}=1, others 0), else (iii) a floored proportional split with weak bias applied to m^- and m^+:

$$\alpha_m \propto \max\left(\frac{\hat{p}^m}{\sum_{j\in\mathcal{M}} \hat{p}^j}, \varepsilon\right) \cdot \begin{cases} 1 + \lambda_{\text{bias}}, & m = m^-, \\ 1 - \lambda_{\text{bias}}, & m = m^+, \\ 1, & \text{otherwise}, \end{cases} \tag{18}$$

followed by renormalization. Encoder gradients are then scaled as $\nabla_{\theta_m} \leftarrow \alpha_m \nabla_{\theta_m}$ with the same budget-preserving renormalization from Section 3.5 to stabilize update magnitudes under aggressive gating.

Fusion-only PCGrad generalizes by computing modality-specific fusion gradients $\{\mathbf{g}_f^m\}$ and applying standard pairwise PCGrad projections sequentially across conflicting pairs, then updating the fusion head with the summed projected gradients.

Unified training pipeline. The previous sections introduced the CAT-GS components individually. We now summarize how these components operate together during a single training iteration. As shown in Algorithm 1, teacher reliability estimation generates the control signal for adaptive gating, which determines modality-wise gradient scaling. Gradient-budget reallocation then stabilizes the magnitude of gated gradients, and fusion-layer surgery removes cross-modal conflicts before parameters are updated.

Algorithm 1 CAT-GS Training Step

Input: (x^a, x^v, y), teachers T^a, T^v, student S
Hyperparameters: $\tau_{\text{low}}, \tau_{\text{high}}, E_w, p_{\text{drop}}, \beta, \beta_g, \lambda_{bias}, \gamma_{\text{cap}}, \varepsilon, \delta$
1. Teacher Reliability
$z^a \leftarrow T^a(x^a)$, $z^v \leftarrow T^v(x^v)$
Compute calibrated confidences p_t^a, p_t^v (Eq. (4))
Update EMA reliabilities $\hat{p}^a, \hat{p}^v$ (Eq. (5))
2. Student Forward & Backward
$f_s^a, f_s^v, \mathbf{s} \leftarrow S(x^a, x^v)$
Compute $\mathcal{L}$ and gradients ∇_θ
3. Gating Regime Selection
$\Delta \leftarrow |\hat{p}^a - \hat{p}^v|$ (Eq. (6))
if $\Delta < \tau_{\text{low}}$ and epoch $< E_w$ **then**
    Apply stochastic modality dropout with probability p_{drop}
else if $\Delta > \tau_{\text{high}}$ **then**
    Apply dominance gating (Eq. (8))
else
    Compute weak-bias weights α_a, α_v (Eqs. (9)–(10))
end if
4. Gradient Modulation
$\nabla_{\theta_m} \leftarrow \alpha_m \nabla_{\theta_m}$ (Eq. (11))
5. Budget Reallocation
Update $\hat{g}^m$ and rescale gradients (Eqs. (12)–(13))
6. Fusion-Layer Surgery
Compute fusion-output-layer gradients $\mathbf{g}_f^a, \mathbf{g}_f^v$ via Eq. (14); if $\cos(\mathbf{g}_f^a, \mathbf{g}_f^v) < 0$, apply PCGrad-style projection (Eqs. (15)–(16)) and overwrite fusion-layer gradient with $\mathbf{g}_f^a + \mathbf{g}_f^v$
7. Optimizer Update
$\theta \leftarrow \theta - \eta \nabla_\theta$

Table 1: Accuracy (%) of unimodal teachers (T^a, T^v) and multimodal students on audio–visual benchmarks. For each dataset, the "Audio" and "Video" rows present the performance of the student's modality-specific encoders, and the "Multi" row reports the fused multimodal output. The best and second-best multimodal scores are highlighted in bold and underlined, respectively (where applicable). For CAT-GS, we report mean±std when repeated runs are available.

Dataset		T^a	T^v	Joint-Train	MSES [20]	MSLR [3]	AGM [21]	PMR [22]	OGM-GE [2]	MLA [23]	Recon Boost [24]	MM Pareto [25]	DLMG [26]	UMT [17]	G^{2}D [6]	CAT-GS (Ours)
CREMA-D	Audio	61.96	-	59.95	54.86	54.86	48.58	49.19	58.60	59.27	65.46	57.71	54.37	61.02	56.45	55.38
	Video	-	76.48	27.42	22.57	26.31	57.85	23.25	49.06	64.91	55.24	65.21	70.89	25.40	72.72	75.23
	Multi	-	-	67.47	60.99	64.42	78.48	59.13	72.18	79.70	75.13	79.82	83.62	67.61	<u>85.89</u>	**86.29±0.15**
AV-MNIST	Audio	42.70	-	16.05	27.50	22.72	38.90	37.60	24.53	42.26	42.11	41.50	41.99	31.55	39.10	41.28
	Video	-	65.34	55.83	63.34	62.92	63.65	58.50	55.85	65.30	65.26	64.28	65.04	64.08	65.09	64.89
	Multi	-	-	69.77	70.68	70.62	72.14	71.82	71.08	65.32	72.63	72.47	72.14	72.33	<u>73.03</u>	**73.21±0.08**
VGGSound	Audio	43.39	-	39.22	39.57	39.10	38.15	26.30	37.96	37.56	42.44	42.35	41.54	42.12	39.43	39.30
	Video	-	32.32	18.70	17.85	18.66	25.65	7.12	22.64	32.02	17.94	18.12	23.65	23.77	29.88	28.84
	Multi	-	-	50.97	50.76	50.98	47.11	33.07	51.45	51.65	49.69	50.97	52.74	<u>53.78</u>	**53.82**	53.24±0.24

This pipeline consolidates CAT-GS into a lightweight controller applied during optimization. It introduces no changes to the forward computation and only modest overhead in the backward pass, making it easily applicable to a wide range of multimodal architectures.

Local Stability Analysis. CAT-GS is an optimization-stage controller that shapes the effective update direction without altering the objective. We provide a local descent guarantee for the induced update.

Let $\mathcal{L}(\theta)$ denote the training objective and $\mathbf{u}_t$ the CAT-GS update after gating, budget reallocation, and fusion-only surgery:

$$\theta_{t+1} = \theta_t - \eta \mathbf{u}_t. \tag{19}$$

Assume: (A1) $\mathcal{L}$ is L-smooth; (A2) $\|\mathbf{u}_t\| \le C\|\nabla\mathcal{L}(\theta_t)\|$; and (A3)

$$\langle \nabla\mathcal{L}(\theta_t), \mathbf{u}_t \rangle \ge \rho \, \|\nabla\mathcal{L}(\theta_t)\| \, \|\mathbf{u}_t\|, \quad \rho \in (0, 1]. \tag{20}$$

Then

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) - \eta\left(\rho - \frac{L\eta C}{2}\right)\|\nabla\mathcal{L}(\theta_t)\|^2. \tag{21}$$

Hence, for $0 < \eta < \frac{2\rho}{LC}$, CAT-GS guarantees one-step monotone descent. This reflects: (i) EMA smoothing stabilizes regime transitions, (ii) budget caps control C, and (iii) fusion-only surgery improves alignment (larger ρ). This is a local, one-step descent guarantee for the CAT-GS update, not a global convergence result; the broader stabilization behavior is supported empirically in Section 4.

4. Experiments

We evaluate CAT-GS on audio–visual, tri-modal, and controlled synthetic settings to answer the following questions: (Q1) Does CAT-GS improve multimodal accuracy compared to state-of-the-art imbalance-aware training methods? (Q2) How effectively does CAT-GS stabilize unimodal branches and mitigate modality starvation? (Q3) Does CAT-GS reduce gradient conflicts and gate thrashing during optimization? (Q4) How do the individual components of CAT-GS contribute to overall performance?

4.1. Experimental Setup

Datasets. Following prior work on multimodal imbalance and gradient modulation, we evaluate CAT-GS on a representative mix of audio–visual, tri-modal, and controlled synthetic benchmarks. **CREMA-D** [27] is an audio–visual emotion recognition dataset with 7,442 clips, 6 emotion classes, and paired speech–face recordings. **AV-MNIST** [28] combines spoken-digit audio with PCA-compressed MNIST images for 10-class classification, and is widely used to study modality dominance under noise. **VGGSound** [29] is a large-scale audio–visual benchmark with diverse in-the-wild categories. **UR-FUNNY** [30] is a tri-modal (audio, visual, text) humor detection dataset with 16k labeled segments and speaker-independent splits, enabling evaluation under stronger cross-modal interactions. **CG-MNIST** [31] is a controlled synthetic variant of MNIST that allows explicit manipulation of modality correlations, making it suitable for analyzing imbalance and gradient starvation effects. **Audio-Visual Event (AVE)** [32] is an additional benchmark with 4,143 videos across 28 event categories, where videos are temporally labeled with audio–visual event boundaries. **CMU-MOSI** [33] is an additional benchmark containing 2,199 opinion video clips, each annotated with sentiment. We use AVE and CMU-MOSI as additional benchmarks to evaluate generalization beyond the core benchmark set. All datasets in this work are established public benchmarks obtained from official online sources.

Baselines. We compare CAT-GS with ten representative multimodal imbalance-mitigation or gradient-modulation methods: MSES [20], MSLR [3], AGM [21], PMR [22], OGM-GE [2], MLA [23], MM-Pareto [25], ReconBoost [24], DLMG [26], and UMT [17]. These baselines span gradient-based rebalancing, multi-objective optimization, reconstruction-based regularization, and unimodal-teacher-guided training.

Backbone Architectures and Hyperparameters. For audio–visual datasets (CREMA-D and AV-MNIST), we employ **ResNet-18** encoders [34] for both audio and video modalities across student and teacher models. For the tri-modal

Table 2: Accuracy (%) on the UR-FUNNY dataset for bi-modal and tri-modal combinations (A–V, A–TXT, V–TXT, and A–V–TXT). Unimodal teacher performance for audio, visual, and text are 58.67%, 56.84%, and 63.48%, respectively.

Type		Joint-Train	OGM-GE [2]	MM Pareto [25]	Recon Boost [24]	UMT [17]	G^{2}D [6]	CAT-GS
A-V	Audio	57.34	59.76	61.77	60.53	54.63	59.05	58.25
	Visual	53.92	53.82	55.73	57.87	56.44	58.05	54.37
	Multi	61.57	61.87	61.27	62.07	60.46	**62.98**	62.12±0.18
A-TXT	Audio	50.30	54.12	58.15	50.18	55.63	59.86	59.42
	Text	57.44	58.35	58.45	56.98	57.75	58.85	59.30
	Multi	62.17	62.47	62.80	61.06	62.47	63.28	**64.49±1.05**
V-TXT	Visual	49.30	55.33	56.04	55.41	56.34	56.34	56.91
	Text	51.21	58.95	59.15	50.94	53.82	56.04	57.60
	Multi	62.07	62.98	61.27	60.07	63.18	63.48	**65.25±0.97**
A-V-TXT	Audio	55.03	50.30	58.05	51.65	50.70	59.15	59.60
	Visual	54.93	55.73	56.14	55.26	54.93	55.94	55.41
	Text	58.25	55.71	58.55	56.25	52.72	58.15	60.72
	Multi	62.58	63.68	62.88	61.37	63.38	65.49	**67.55±0.84**

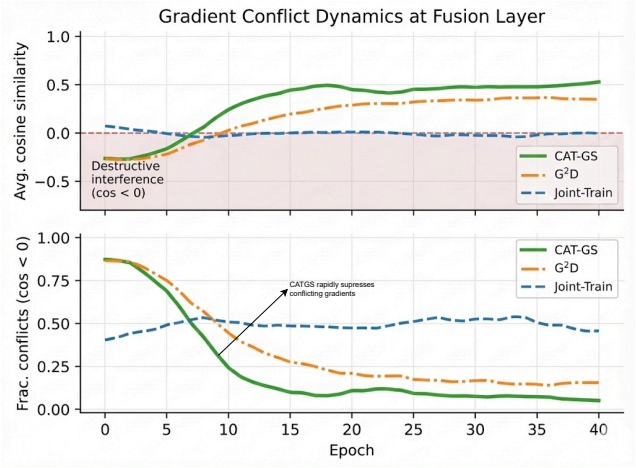

Figure 4: Gradient conflict dynamics at the fusion layer. CAT-GS achieves higher gradient alignment and fewer conflicting updates than Joint-Train and G^{2}D, indicating effective mitigation of destructive interference.

UR-FUNNY dataset, we use lightweight **Transformer encoders** [35] for text, audio, and visual streams to maintain consistent modality capacity. **Teacher Protocol.** To ensure high-quality guidance, all teacher models are pre-trained individually on their respective unimodal datasets until convergence. During student training, these teachers are **frozen** and used solely for inference to provide logits for the calibration module. This setup ensures that the reliability estimates p_t^m are derived from stable, converged experts rather than evolving peers.

Training Details. Unless otherwise noted, we train for 400 epochs with batch size 16 using SGD (momentum 0.9, weight decay 10^{-4}) and an initial learning rate of 10^{-3}, decayed by a factor of 0.1 at epoch 200. We fix the CAT-GS hyperparameters $\beta = 0.9$, $\beta_g = 0.9$, $\gamma_{\text{cap}} = 1.5$, $\varepsilon = 0.1$, $\lambda_{\text{bias}} = 0.2$, warm-up epochs $E_w = 5$, and dropout probability $p_{\text{drop}} = 0.6$ across datasets, while tuning the regime thresholds τ_{low} and τ_{high} once per dataset. In all experiments except the explicit sensitivity analysis, we use identical CAT-GS hyperparameters across datasets.

Evaluation Protocol and Reproducibility. We report top-1 classification accuracy (%) on the held-out test split for each benchmark. During training, we evaluate once per epoch on a validation split and select the checkpoint with the best validation accuracy; we then report test accuracy once using this selected checkpoint. Unless otherwise noted, we report the average test accuracy (%) over three random seeds (42, 123, and 999). For CAT-GS, we additionally report run-to-run variability as mean±std in Tables 1, 2, and 4. With n=3 runs, 95% confidence intervals are computed as $\bar{x} \pm t_{0.975,2}\, s/\sqrt{3}$. CG-MNIST experiments use small CNN encoders to isolate optimization behavior from representational capacity. All methods use identical backbones and fusion heads to ensure fair comparison. Experiments are conducted on NVIDIA RTX 3060 and RTX 4050 GPUs.

Table 3: Additional benchmark comparison on AVE and CMU-MOSI. We report accuracy (%) for representative balancing baselines and CAT-GS under the corresponding setup family (AVE: ResNet V–A; MOSI: Transformer V–T and V–A–T).

Method	AVE (ResNet V–A)	CMU-MOSI (Transformer V–T)	CMU-MOSI (Transformer V–A–T)
MSLR [3]	67.3	–	–
MMCosine [9]	65.0	–	–
OGM [2]	67.3	73.9	–
AGM [21]	68.4	74.0	73.9
ReconBoost [24]	68.4	–	–
MM-Pareto [25]	73.0	73.4	73.7
MCR [11]	73.4	75.2	76.5
DRL [10]	–	–	77.99
CAT-GS (Ours)	**74.2±0.2**	**76.0±1.2**	**78.3±1.0**

Table 4: Performance comparison on the CG-MNIST dataset. Modality 1: monochromatic image; Modality 2: gray-scale image.

Modality	PMR [22]	OGM-GE [2]	MSES [20]	MSLR [3]	G^{2}D [6]	CAT-GS
Mono	99.30	99.38	99.26	99.30	99.26	99.32
Gray	60.40	67.21	60.85	59.96	63.36	67.85
Multi	78.50	97.35	93.46	95.04	97.08	**97.42±0.23**

4.2. Results

Audio–Visual Classification Benchmarks. Table 1 reports results on CREMA-D, AV-MNIST, and VGGSound. In Table 1, **Joint-Train** denotes the original model trained with standard end-to-end optimization (CAT-GS disabled) under the same backbone, fusion head, optimizer schedule, and data split as CAT-GS for fair attribution. Across datasets, CAT-GS consistently remains competitive with imbalance-aware baselines spanning gradient modulation (e.g., OGM-GE [2], AGM [21], PMR [22]) and teacher-guided training (UMT [17]).

On CREMA-D, where the video modality is substantially stronger than audio, CAT-GS improves the fused prediction to 86.29%±0.15 (vs. 85.89% for G^{2}D), indicating that stabilizing gating and resolving fusion conflicts can yield gains beyond confidence-only suppression. On AV-MNIST, where modalities are closer in strength, CAT-GS remains competitive and achieves the best fused accuracy (73.21%±0.08), suggesting the controller does not over-correct when imbalance is mild.

On VGGSound, CAT-GS reaches 53.24%±0.24 fused accuracy and does not surpass the strongest baselines, falling below

G^2D (53.82%) and UMT (53.78%) by roughly 0.5–0.6 points. This reflects the large-scale in-the-wild regime rather than the training setup: the large label space (∼300 classes) and noisy clips yield low, unstable teacher reliabilities, so the margin Δ that drives gating is weakly informative, and accuracy at this scale is dominated by encoder capacity and data volume rather than by gradient-level control. CAT-GS therefore offers limited benefit in this regime. The unimodal branches on VGGSound follow the same pattern and remain close to G^2D, with slightly lower Audio/Video accuracies (39.30 vs. 39.43 and 28.84 vs. 29.88).

For completeness, we also include a no-teacher standard-training reference on CREMA-D: training the student with only the supervised classification objective (i.e., without distillation and without CAT-GS modulation) reaches 64.78% best fused accuracy under the same backbone, fusion setup, and training schedule, and remains substantially below CAT-GS, consistent with the relative ordering reported in Table 1.

The corresponding 95% confidence intervals for CAT-GS are [85.92, 86.66] (CREMA-D), [73.01, 73.41] (AV-MNIST), and [52.64, 53.84] (VGGSound). We note that fused accuracy is the primary objective; unimodal branch accuracies can change in either direction depending on how much specialization toward fusion is beneficial.

Tri-Modal Humor Understanding (UR-FUNNY). Table 2 reports results on UR-FUNNY across increasing modality combinations, from audio–visual (A–V) to text-inclusive settings. In the A–V configuration, CAT-GS performs competitively but does not outperform the strongest baseline (G^2D), which is expected given the reduced modality interactions and weaker gradient conflicts in the bi-modal setting. In contrast, when textual modality is introduced, CAT-GS consistently achieves the best multimodal accuracy in both A–TXT and V–TXT settings, and yields the largest gains in the full A–V–TXT configuration. Specifically, CAT-GS improves over multi-objective optimization (MM-Pareto [25]) and reconstruction-based regularization (ReconBoost [24]), highlighting the effectiveness of calibrated gating, gradient-budget stabilization, and fusion-layer conflict mitigation under more challenging tri-modal fusion. For CAT-GS on UR-FUNNY, the 95% confidence intervals are [61.67, 62.57] (A–V), [61.88, 67.10] (A–TXT), [62.84, 67.66] (V–TXT), and [65.46, 69.64] (A–V–TXT).

Additional Benchmarks (AVE and CMU-MOSI). Table 3 extends evaluation to additional benchmark families and includes recent baselines such as MCR and DRL. CAT-GS remains strongest on AVE and on the MOSI two-modality setting, and reaches 78.3%±1.0 on MOSI three-modality classification, slightly above DRL (77.99) and clearly above MCR (76.5). The corresponding 95% confidence intervals for CAT-GS are [73.70, 74.70] on AVE, [73.02, 78.98] on CMU-MOSI (V–T), and [75.82, 80.78] on CMU-MOSI (V–A–T). Compared with earlier methods (e.g., MSLR/OGM/AGM/MM-Pareto), the margins are larger, while against stronger recent methods they become tighter but remain positive. Overall, these results indicate that CAT-GS transfers to both event-centric and sentiment-centric benchmarks instead of being limited to the original dataset family.

Controlled Synthetic Benchmark (CG-MNIST). Results on CG-MNIST are shown in Table 4. This benchmark allows explicit control over label correlations between the monochromatic and grayscale modalities. Even when one modality is made spuriously more predictive early in training, CAT-GS maintains strong fused performance and avoids brittle training dynamics observed in several baselines. This controlled setting supports the role of margin-thresholded gating and budget-aware stabilization in reducing the risk that prolonged dominance suppresses the weaker modality. For CAT-GS on CG-MNIST, the 95% confidence interval is [96.85, 97.99].

Evidence for Reduced Gradient Conflicts. To quantify cross-modal interference, we follow the implementation and define a modality-specific fusion gradient surrogate $\mathbf{g}_f^m$ on the fusion output layer (e.g., `fc_out`) using the logit-sum gradients from Eq. (14). We measure gradient alignment using the cosine similarity $\cos(\mathbf{g}_f^a, \mathbf{g}_f^v)$. As shown in Figure 4 (top), CAT-GS maintains a positive average cosine similarity throughout training, indicating a cooperative optimization regime at the fusion layer. Correspondingly, Figure 4 (bottom) shows that CAT-GS substantially reduces the fraction of training batches with negative cosine similarity, directly evidencing suppression of destructive gradient interference.

Gate Stability and Gradient Starvation Analysis. We quantify gating stability using the Total Variation (TV) of the gating coefficients. CAT-GS demonstrates smooth gating transitions with low total variation, confirming that calibration and EMA smoothing effectively filter out high-frequency noise ("gate thrashing"). To assess starvation risk under gating, we track the gradient norms of the unimodal encoders. CAT-GS's ε-floored soft regimes reduce accidental gradient collapse, and its capped budget reallocation stabilizes update magnitudes during hard-gating events.

Observed failure-mode behavior. Figure 5 summarizes optimization dynamics from CREMA-D training logs to illustrate gate thrashing and dominant-modality trapping. In panel (a), the audio gate α_{audio} is plotted over 600 steps: OGM-GE and PMR exhibit pronounced oscillations under batch-wise modulation, whereas CAT-GS produces smooth trajectories using EMA-smoothed reliability ($\beta = 0.9$) and margin-based regime switching. In panel (b), we report the cumulative gradient budget ratio $\Sigma\|\nabla_v\|/\Sigma\|\nabla_a\|$. Joint training accumulates a 3.2× imbalance; gradient-modulation baselines remain around ∼ 2.4×; G2D reaches ∼ 2.0×; and CAT-GS maintains 1.7× through budget-preserving reallocation (Eq. (13)). Together, these diagnostics indicate that prior approaches can suffer from gate instability and persistent gradient imbalance, which CAT-GS is designed to address.

For clearer interpretation of panel (a), Figure 6 shows decomposed gating trajectories (raw and smoothed) for Joint, OGM-GE, PMR, G^2D, and CAT-GS. In this decomposed view, each method is shown with a raw per-step trajectory and a smoothed trend (moving average), which separates high-frequency jitter from underlying control behavior. Joint remains largely static, indicating weak reliability-driven adaptation. OGM-GE and PMR show strong high-frequency fluctuation and jagged

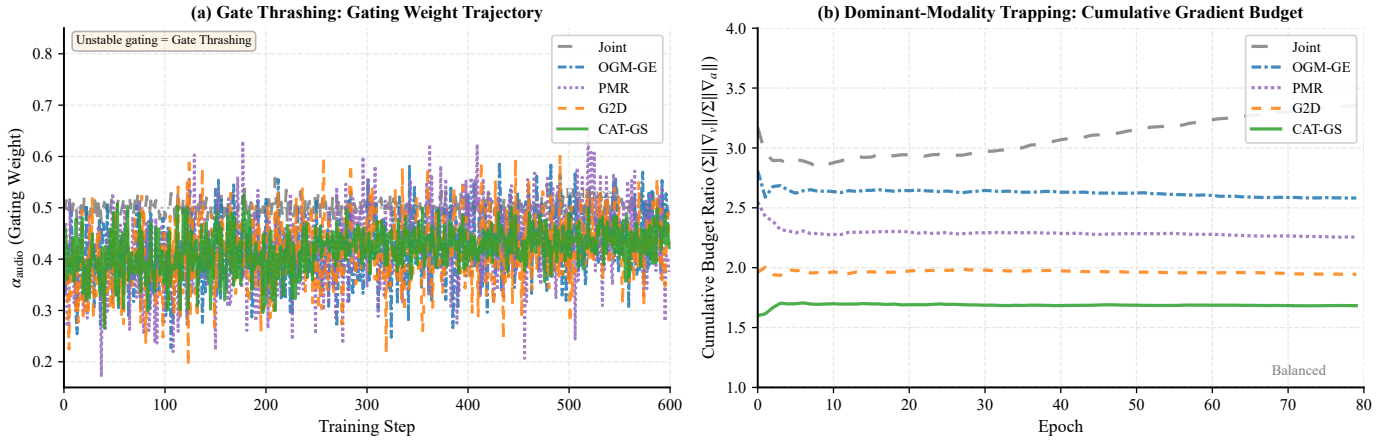

Figure 5: Optimization diagnostics on CREMA-D. (a) Audio-gating trajectory over 600 steps: OGM-GE and PMR show high-frequency oscillations (gate thrashing), whereas CAT-GS remains smooth under EMA-stabilized control. (b) Cumulative gradient-budget ratio $\Sigma\|\nabla_v\|/\Sigma\|\nabla_a\|$: Joint-Train reaches 3.2× imbalance, G^{2}D remains near 2.0×, and CAT-GS reduces this to 1.7× through budget-preserving reallocation.

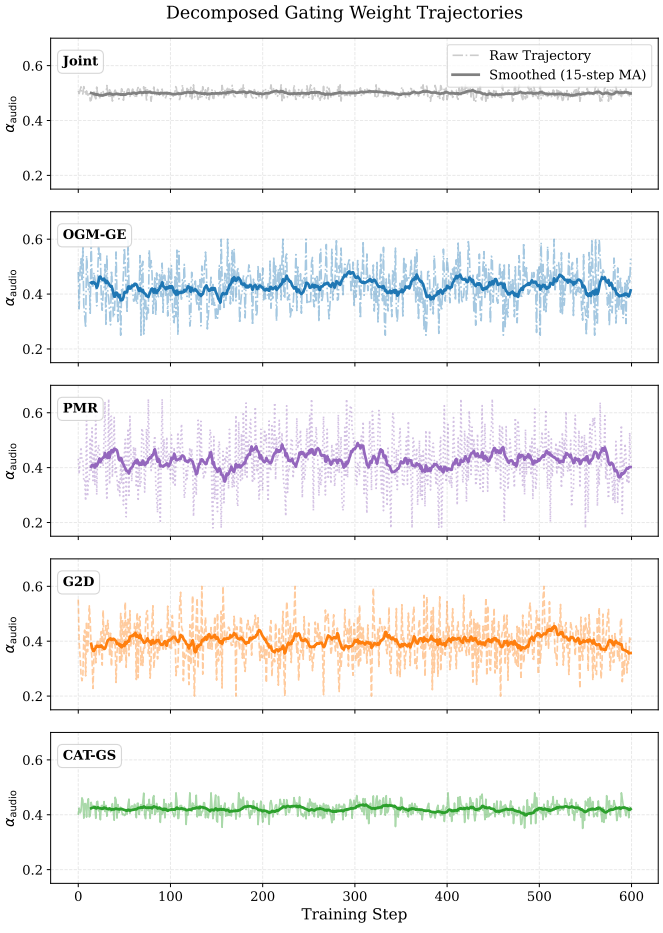

Figure 6: Decomposed gating weight trajectories (α_{audio}) across training steps for Joint, OGM-GE, PMR, G^{2}D, and CAT-GS.

trends, consistent with unstable batch-reactive routing. G^{2}D reduces fluctuation amplitude relative to OGM-GE/PMR but still exhibits frequent short-window switching. CAT-GS is the smoothest, reducing oscillation while preserving gradual trend adaptation, which is consistent with its EMA-stabilized and margin-thresholded controller design.

To complement this optimization-dynamics view, we further provide a confusion-matrix-style qualitative analysis on the three core audio–visual benchmarks (CREMA-D, AV-MNIST, and VGGSound), in which each test sample is assigned to a single cell according to which unimodal branches predicted it correctly and whether the fused prediction was correct (Figure 7). The top-left cell holds the samples that neither branch predicted correctly and that fusion did not recover, which are limited by the unimodal encoders; the wrong row of the remaining columns holds those on which at least one branch was correct but the fused prediction was not, which are limited by the fusion stage. Together the two account for all errors. On CREMA-D, CAT-GS reduces both relative to Joint-Train, from 19.5% to 7.4% and from 13.4% to 6.3%, consistent with the reduced negative fusion-gradient cosine events in Figure 4. The fusion-stage loss is lower than that of G^{2}D on all three benchmarks (6.3% vs. 7.5%, 10.7% vs. 11.7%, 11.8% vs. 12.4%), but on VGGSound the encoder-limited error is larger (35.1% vs. 33.4%) and outweighs that advantage. This places the VGGSound shortfall in the unimodal branches rather than at the fusion layer, consistent with the weakly informative reliability margin Δ and the large label space at this scale.

4.3. Ablation Studies

We perform two complementary ablations: a unified component matrix with both single-component activation and leave-one-out removal (Table 5), and architectural ablation over fusion modules (Table 6).

Table 5 merges both ablation views using a tick-mark configuration: (i) starting from Joint-Train and enabling one component at a time, and (ii) starting from full CAT-GS and removing one component at a time. This gives a compact view of standalone contribution and component necessity in one place.

Unified Component Ablation (Table 5). Gating acts as the controller's hub: the calibrated, EMA-smoothed reliability mar-

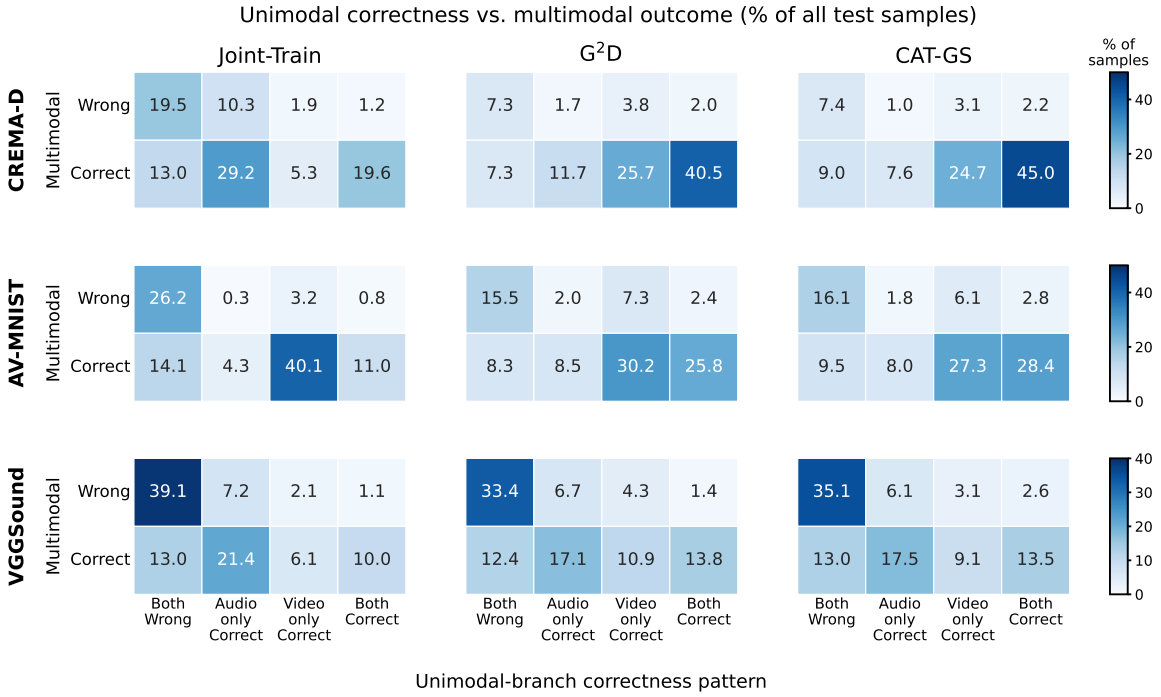

Figure 7: Unimodal correctness versus multimodal outcome for Joint-Train, G^{2}D, and CAT-GS (columns) on CREMA-D, AV-MNIST, and VGGSound (rows). Within each panel, columns group test samples by which unimodal branches are correct and rows give the multimodal outcome.

gin drives only the gating regime selector, the weak-bias rule is itself a gating regime, and gradient-budget reallocation is invoked only at hard-gate (α=0) events. Calibrated gating is therefore active in every variant that improves over Joint-Train, so we omit calibration and gating as columns and vary the four remaining components; removing gating reduces CAT-GS to fusion-only PCGrad, the "+ Fusion-only PCGrad only" row (77.80, +10.33). In the leave-one-out block, "w/o Weak-Bias" disables only the weak-bias factor by setting λ_{bias}=0, while keeping the adaptive gating controller (warm-up, dominance, and soft regime selection via $\tau_{\text{low}}, \tau_{\text{high}}$) unchanged. Turning off *gradient-budget reallocation* produces the largest drop in the removal block, showing that gating alone is insufficient to prevent long-term update collapse. Discarding *EMA smoothing* also causes a substantial drop, indicating strong dependence on stabilized reliability signals. Omitting *fusion-layer PCGrad surgery* hurts performance to a lesser extent, indicating conflict resolution is important but secondary to reliability and budget stabilization. The activation block complements this view by showing each component helps over Joint-Train, while the full combination remains strongest.

Fusion Strategy Analysis (Table 6). Table 6 compares different fusion modules when used inside CAT-GS. We explore simple additive fusion (Sum), concatenation (Concat), FiLM-style conditioning [36], bidirectional gating (BiGated) [37], cross-attention [38], and explicit Early/Late Fusion variants [39, 40]. We observe three trends. First, naive Sum fusion is consistently weaker than more expressive alternatives, especially on CREMA-D and CG-MNIST, suggesting that richer cross-modal interactions remain beneficial even when gradients are well-balanced.

Table 5: Component-wise contribution and necessity analysis on CREMA-D. Checkmarks indicate enabled components. Calibrated reliability and gating are the always-on base controller (not shown as columns); we vary the four components listed.

Variant	EMA	Weak-Bias	Budget	PCGrad	MM	Δ_{Joint}
Single-component activation (from Joint-Train)						
Joint-Train (baseline)	–	–	–	–	67.47	0.00
+ Calibrated+EMA reliability	✓	–	–	–	78.90	+11.43
+ Adaptive gating (incl. weak-bias)	–	✓	–	–	77.20	+9.73
+ Gradient-budget only	–	–	✓	–	79.40	+11.93
+ Fusion-only PCGrad only	–	–	–	✓	77.80	+10.33
Leave-one-out removal (from full CAT-GS)						
CAT-GS (full)	✓	✓	✓	✓	**86.29**	**+18.82**
w/o EMA smoothing	–	✓	✓	✓	84.60	+17.13
w/o Weak-Bias	✓	–	✓	✓	85.50	+18.03
w/o Grad. Budget	✓	✓	–	✓	83.20	+15.73
w/o PCGrad	✓	✓	✓	–	85.08	+17.61

Table 6: Performance comparison of **CAT-GS** under different fusion strategies across multiple datasets in terms of accuracy (%).

Fusion	CREMA-D	AV-MNIST	CG-MNIST	UR-FUNNY
Sum	82.10	71.85	94.26	64.02
Concat	84.72	72.93	95.88	65.01
FiLM [36]	83.95	72.12	92.43	63.77
BiGated [37]	81.88	72.44	91.36	63.25
Cross-Attention [38]	85.04	72.98	96.73	66.41
Transformer Fusion [16]	85.73	72.95	97.06	66.56
Early Fusion [39]	82.62	72.18	93.45	64.02
Late Fusion [39]	**86.29**	**73.21**	**97.42**	**67.55**

Second, cross-attention and concatenation form a strong pair of contenders, often approaching the best performance.

We additionally tested a lightweight transformer-style token-fusion variant [16] under the same training protocol; CAT-GS remained stable and trainable, while absolute performance differ-

ences remained architecture-dependent. In our experiments, Late Fusion is the primary setting (best absolute accuracy), and we report Early Fusion and Transformer Fusion as transfer checks under the same training protocol.

The fusion ablation isolates the role of the architecture used at the multimodal head while keeping the CAT-GS controller fixed. FiLM and BiGated fusion can already capitalize on CAT-GS, but their performance is inconsistent across datasets, likely due to the stronger inductive assumptions they impose (e.g., one modality modulating the other). Cross-attention provides a strong trade-off and often matches or slightly trails the strongest-performing variant across datasets. To assess transfer across fusion heads, we evaluate Transformer Fusion, Early Fusion, and Late Fusion under the same training protocol. CAT-GS remains stable in all cases, with expected architecture-dependent absolute differences. Taken together, Tables 5 and 6 show that (i) each component of CAT-GS contributes complementary gains, and (ii) the method is not overly sensitive to the specific fusion module.

Overall, CAT-GS delivers consistent gains across classification, tri-modal understanding, and controlled synthetic settings. The ablations confirm that its advantages arise from the combination of stabilized reliability signals, adaptive regime switching, budget-aware gradient stabilization, and conflict-aware fusion.

Teacher Miscalibration Stress Test (Table 7). CAT-GS relies on unimodal teachers to provide modality reliability signals. To assess robustness to errors in these signals, we introduce a controlled teacher miscalibration stress test by applying a temperature mismatch at teacher inference time. For modality m, teacher logits are perturbed as $z_m \leftarrow z_m / T^m_{\text{mis}}$ before computing the confidence used by the CAT-GS controller.

We evaluate four regimes: (R0) a calibrated baseline, (R1) an overconfident audio teacher, (R2) an underconfident audio teacher, and (R3) both modalities miscalibrated. To isolate the role of reliability stabilization, we further evaluate the worst-performing mismatch (R2) with EMA smoothing disabled (R4) and temperature calibration removed (R5). All other settings are kept identical. For R5, we bypass the calibration step when computing reliability; to avoid confusion, we omit the mismatch-temperature entries in Table 7. Table 7 reports multimodal (MM) accuracy under each condition. CAT-GS degrades gracefully under moderate miscalibration (R1–R2), remaining close to the calibrated baseline. Performance drops further under the stressed setting (R3) but remains stable. In contrast, removing EMA smoothing (R4) or temperature calibration (R5) leads to a larger degradation, highlighting the importance of reliability stabilization. Overall, CAT-GS is robust to moderate teacher miscalibration, with stabilization mechanisms playing a critical role. R4–R5 are best interpreted as internal stress controls (ablation-style reliability degradation), not as a standalone weak-teacher/domain-shift benchmark. Under these stressed settings, CAT-GS remains trainable and stable, but final accuracy decreases relative to the strong-teacher setting. This supports a balanced conclusion: robustness to moderate unreliability, with clear dependence on teacher quality under stronger reliability degradation.

Table 7: Teacher miscalibration stress test on CREMA-D dataset. Multimodal (MM) accuracy of the fused model under temperature-mismatched teachers.

Run	T^a_{mis}	T^v_{mis}	MM Acc. (%) ↑	Δ (%) ↓
R0 (Baseline)	1.0	1.0	86.3	0.0
R1 (Audio overconf.)	0.5	1.0	85.2	−1.1
R2 (Audio underconf.)	2.0	1.0	85.1	−1.2
R3 (Both mismatched)	2.0	0.5	84.4	−1.9
R4 (R2 w/o EMA)	2.0	1.0	83.6	−2.7
R5 (R2 w/o calibration)	–	–	84.0	−2.3

Table 8: Modality-imbalance validation on CREMA-D (clean vs degraded modality). Values are fused multimodal (MM) test accuracy (%). Drop is measured in percentage points as $\text{MM}_{\text{clean}} - \text{MM}_{\text{condition}}$.

Condition	Joint (MM %)	CAT-GS (MM %)	Joint Drop (pp)	CAT-GS Drop (pp)
Clean (none)	67.47	86.29	0.00	0.00
Audio-degraded	46.23	71.79	21.24	14.50
Visual-degraded	55.40	78.29	12.07	8.00

Modality-Imbalance Validation (clean vs degraded modality). To directly validate robustness to modality imbalance, we run a controlled clean-vs-degraded protocol on CREMA-D under identical training settings for Joint/Normal and CAT-GS. We evaluate three conditions: clean (none), audio-degraded, and visual-degraded. All reported values are fused multimodal (MM) test accuracy (%). Results are summarized in Table 8. Under audio degradation, Joint/Normal drops from 67.47% to 46.23% (drop 21.24), whereas CAT-GS drops from 86.29% to 71.79% (drop 14.50). Under visual degradation, Joint/Normal drops from 67.47% to 55.40% (drop 12.07), whereas CAT-GS drops from 86.29% to 78.29% (drop 8.00). Therefore, CAT-GS yields smaller degradation in both stressed conditions, supporting the claim that it improves robustness under modality imbalance.

Threshold Sensitivity Analysis. We evaluate the sensitivity of CAT-GS to τ_{low}, τ_{high}, and λ_{bias} with a grid search over the three thresholds. Across all tested combinations, the maximum absolute change in fused accuracy relative to the default setting stays below 0.5 percentage points (variance $< 0.5\%$), with stable performance for $\tau_{\text{low}} \in [0.03, 0.08]$, $\tau_{\text{high}} \in [0.10, 0.20]$, and $\lambda_{\text{bias}} \in [0.1, 0.3]$, so CAT-GS is not highly sensitive within these ranges. These ranges translate into a simple setting procedure: (1) run a short pilot (e.g., first 3–5 epochs) and log the reliability margin Δ per mini-batch, (2) inspect the empirical histogram of Δ, (3) set τ_{low} near the lower-middle mass of the histogram (about the 40th percentile) and τ_{high} near the clear-dominance region (about the 80th percentile), while keeping a gap of at least 0.05, and (4) keep $\lambda_{\text{bias}} \in [0.1, 0.3]$ unless validation indicates otherwise.

Runtime Analysis. CAT-GS introduces minimal computational overhead compared to standard joint training. The additional operations—teacher inference (which can be precomputed or run in parallel), reliability calibration, and gradient rescaling—are lightweight vector operations. The fusion-layer surgery involves only a single-sided projection step on the fusion output layer (e.g., `fc_out`) using the logit-sum gradients from Eq. (14), avoiding the high cost of full-network gradient projection. Em-

pirically, we observe a per-epoch training time increase of approximately 2–5% across our benchmarks, which is negligible given the performance gains and improved convergence stability.

4.4. Discussion

CAT-GS is most beneficial when unimodal reliability varies across training and the fused head is a meaningful shared bottleneck where cross-modal gradients can interfere. On large-scale in-the-wild datasets such as VGGSound, CAT-GS yields comparable rather than dominant improvements, and in fact does not surpass the strongest baselines (G^{2}D, UMT) there, suggesting that representation capacity and data scale may outweigh optimization control when the teacher-reliability signal is weak and the label space is large. Moreover, CAT-GS inherits any systematic bias in the unimodal teachers: if a teacher is consistently miscalibrated, the controller may over- or under-prioritize a modality. Our temperature-mismatch stress test (Table 7) indicates the method degrades gracefully under moderate miscalibration, and that EMA smoothing and calibration materially improve robustness. Under stronger reliability degradation stress settings, the method remains stable but shows lower final accuracy, reinforcing that teacher quality is an important practical factor. At the same time, teacher guidance alone does not explain the gains: under matched training settings on CREMA-D, Joint-Train (CAT-GS disabled) reaches 67.47%, the no-teacher standard-training reference reaches 64.78%, and CAT-GS reaches 86.29%±0.15 (Table 1). In practice, we recommend CAT-GS when unimodal teachers are reasonably strong and when training dynamics exhibit either gate thrashing, prolonged starvation, or frequent negative fusion-gradient cosine events. CAT-GS is implemented at the optimization stage, so extension across fusion paradigms is operationally straightforward. In our additional architecture checks, CAT-GS remained stable for Early Fusion, Late Fusion, and a transformer-style token-fusion head (Table 6). Together, these results indicate transferability across fusion paradigms, with expected architecture-dependent accuracy differences. For token-level transformer architectures, reliability signals can be computed from modality-specific token branches, with gating/budget stabilization on modality-specific adapters or blocks and conflict handling on shared/cross-attention parameters.

Limitations. CAT-GS relies on pre-trained unimodal teachers to provide reliability signals. If teachers are weak or systematically biased, the controller may inherit these failure modes. The teacher-miscalibration stress test (Table 7) is the closest evidence available: temperature mismatch perturbs the teacher signal (R1–R3) with only gradual accuracy loss, and the R4–R5 controls show EMA smoothing and calibration buffer it. Since this perturbs teacher *calibration* rather than *accuracy*, a genuinely weak or domain-shifted teacher remains untested and is left to future work. The main remaining limitation is dependence on teacher reliability; fully teacher-free CAT-GS control remains an important direction for future work. Future work should investigate teacher-free reliability estimation and stronger domain-shift robustness so that control quality is less dependent on teacher calibration. Extending to many-modality or fully entangled token-level settings mainly requires careful definition of modality-linked parameter groups so that gating/budget operations remain well-posed while shared-layer conflict handling remains localized. From a practical deployment perspective, CAT-GS introduces several hyperparameters (e.g., τ_{low}, τ_{high}, λ_{bias}, β, β_g, γ_{cap}, ε). However, most can be fixed to stable defaults across datasets (we use β=0.9, β_g=0.9, γ_{cap}=1.5, and ε=0.1 throughout), while the primary tuning knobs are the regime thresholds and the weak-bias factor. Our sensitivity analysis indicates broad plateaus where performance is stable (Section 4.3), suggesting limited tuning effort in practice. A simple guideline is to set τ_{low} and τ_{high} to separate "close" vs. "clear-dominance" regions of the observed reliability-margin distribution, and then choose $\lambda_{\text{bias}} \in [0.1, 0.3]$ to softly favor the weaker modality when reliabilities are comparable.

5. Conclusion

We introduced CAT-GS, an optimization-stage learning-dynamics controller for balanced and robust multimodal learning. Instead of modifying architectures or designing bespoke losses, CAT-GS treats multimodal training as a regime-switching control problem and enforces stability constraints on gradient flow, update magnitude, and fusion-gradient geometry. Across audio–visual benchmarks (CREMA-D, AV-MNIST), additional benchmarks (AVE and CMU-MOSI), tri-modal humor understanding (UR-FUNNY), and a controlled synthetic dataset (CG-MNIST), CAT-GS improves multimodal accuracy and training stability over strong imbalance-aware baselines while maintaining competitive unimodal performance. The additional AVE/CMU-MOSI comparisons further indicate that CAT-GS transfers beyond the original dataset family to event-centric and sentiment-centric settings. On the large-scale in-the-wild VGGSound benchmark, however, CAT-GS remains competitive but does not surpass the strongest baselines (G^{2}D, UMT), a limitation we attribute to the weak, noisy teacher-reliability signal and the large label space at this scale. Scaling optimization-stage control to such data, for instance through stronger or teacher-free reliability estimation, is an explicit direction for future work. Ablations show that calibration, gating, budget reallocation, and fusion surgery contribute complementary gains, and CAT-GS remains effective across fusion architectures under a consistent late-fusion-centered evaluation with transfer checks to alternative fusion heads. Future work includes scaling to more modalities, integrating token-level/temporal gating in transformer-based models, and studying interactions with large-scale pretraining.

6. Code Availability

The implementation of CAT-GS is publicly available at https://github.com/mahirshahriar1/CAT-GS.

References

[1] P. Xu, X. Zhu, D. A. Clifton, Multimodal learning with transformers: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (10) (2023) 12113–12132.

[2] X. Peng, Y. Wei, A. Deng, D. Wang, D. Hu, Balanced multimodal learning via on-the-fly gradient modulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[3] Y. Yao, R. Mihalcea, Modality-specific learning rates for effective multimodal additive late-fusion, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1824–1834.

[4] Z. Han, F. Yang, J. Huang, C. Zhang, J. Yao, Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20675–20685.

[5] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 1321–1330.

[6] M. Rakib, A. Bagavathi, G^{2}D: Boosting multimodal learning with gradient-guided distillation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 4059–4068.

[7] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, C. Finn, Gradient surgery for multi-task learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 5824–5836.

[8] K. Kontras, C. Chatzichristos, M. B. Blaschko, M. D. Vos, Improving multimodal learning with multi-loss gradient modulation, in: 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024, BMVA, 2024.

[9] R. Xu, R. Feng, S.-X. Zhang, D. Hu, MMCosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[10] Y. Wei, S. Li, R. Feng, D. Hu, Diagnosing and re-learning for balanced multimodal learning, Springer-Verlag, 2024, p. 71–86.

[11] K. Kontras, T. Strypsteen, C. Chatzichristos, P. P. Liang, M. B. Blaschko, M. D. Vos, Balancing multimodal training through game-theoretic regularization, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

[12] M. Pezeshki, S.-O. Kaba, Y. Bengio, A. Courville, D. Precup, G. Lajoie, Gradient starvation: a learning proclivity in neural networks, in: Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21, Curran Associates Inc., Red Hook, NY, USA, 2021.

[13] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[14] C. Wu, Y. Wei, X. Chu, W. Sun, F. Su, L. Wang, Hierarchical attention-based multimodal fusion for video captioning, Neurocomputing 315 (2018) 362–370.

[15] J. Arevalo, T. Solorio, M. Montes-y Gomez, F. A. González, Gated multimodal networks, Neural Computing and Applications 32 (14) (2020) 10209–10228.

[16] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 6558–6569.

[17] C. Du, J. Teng, T. Li, Y. Liu, T. Yuan, Y. Wang, Y. Yuan, H. Zhao, On uni-modal feature learning in supervised multi-modal learning, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[18] S. Wei, C. Luo, Y. Luo, Boosting multimodal learning via disentangled gradient learning, in: 2025 IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 1–10.

[19] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network (03 2015).

[20] N. Fujimori, R. Endo, Y. Kawai, T. Mochizuki, Modality-specific learning rate control for multimodal classification, in: S. Palaiahnakote, G. Sanniti di Baja, L. Wang, W. Q. Yan (Eds.), Pattern Recognition, Springer International Publishing, Cham, 2020, pp. 412–422.

[21] H. Li, X. Li, P. Hu, Y. Lei, C. Li, Y. Zhou, Boosting multimodal model performance with adaptive gradient modulation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22214–22224.

[22] Y. Fan, W. Xu, H. Wang, J. Wang, S. Guo, PMR: Prototypical modal rebalance for multimodal learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20029–20038.

[23] X. Zhang, J. Yoon, M. Bansal, H. Yao, Multimodal representation learning by alternating unimodal adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[24] C. Hua, Q. Xu, S. Bao, Z. Yang, Q. Huang, Reconboost: boosting can achieve modality reconcilement, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24, JMLR.org, 2024.

[25] Y. Wei, D. Hu, Mmpareto: boosting multimodal learning with innocent unimodal assistance, ICML'24, JMLR.org, 2024.

[26] Y. Yang, F. Wan, Q.-Y. Jiang, Y. Xu, Facilitating multimodal classification via dynamically learning modality gap, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24, Curran Associates Inc., Red Hook, NY, USA, 2024.

[27] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, CREMA-D: Crowd-sourced emotional multimodal actors dataset, IEEE transactions on affective computing 5 (4) (2014) 377–390.

[28] V. Vielzeuf, A. Lechervy, S. Pateux, F. Jurie, CentralNet: A multilayer approach for multimodal fusion, in: Computer Vision – ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part VI, Springer-Verlag, Berlin, Heidelberg, 2019, pp. 575–589.

[29] H. Chen, W. Xie, A. Vedaldi, A. Zisserman, VGGSound: A large-scale audio-visual dataset, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 721–725.

[30] M. K. Hasan, W. Rahman, A. Bagher Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency, M. E. Hoque, UR-FUNNY: A multimodal language dataset for understanding humor, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2046–2056.

[31] B. Kim, H. Kim, K. Kim, S. Kim, J. Kim, Learning Not to Learn: Training Deep Neural Networks With Biased Data , in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 9004–9012.

[32] Y. Tian, J. Shi, B. Li, Z. Duan, C. Xu, Audio-visual event localization in unconstrained videos, in: ECCV, 2018.

[33] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, arXiv preprint arXiv:1606.06259 (2016).

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition , in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2016, pp. 770–778.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.

[36] E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. Courville, Film: visual reasoning with a general conditioning layer, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[37] D. Kiela, E. Grave, A. Joulin, T. Mikolov, Efficient large-scale multi-modal classification, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[38] C.-F. R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 347–356.

[39] H. Gunes, M. Piccardi, Affect recognition from face and body: early fusion vs. late fusion, in: 2005 IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, 2005, pp. 3437–3443 Vol. 4.

[40] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia systems 16 (6) (2010) 345–379.