Learning dynamics of multimodal deep learning: An academic survey

Multimodal machine learning faces a fundamental optimization paradox: networks that receive more information often underperform their unimodal counterparts (arXiv) due to **modality competition**, gradient imbalance, and convergence rate disparities between modalities. This survey synthesizes theoretical foundations and recent advances (2018-2025) in understanding why multimodal training is difficult and how gradient-based interventions, loss design, and architectural choices can restore balanced learning. The core insight across the literature is that standard joint optimization fails because **dominant modalities converge faster, suppressing gradient updates for weaker modalities** (arXiv) (NeurIPS) through what researchers term the "greedy learner hypothesis" or "modality laziness." (arXiv) (arXiv) Solutions have evolved from heuristic loss reweighting to principled gradient modulation methods with theoretical convergence guarantees.

---

1. Loss functions shape cross-modal representations and training stability

Contrastive losses and the InfoNCE foundation

The dominant paradigm for multimodal representation learning emerged from contrastive objectives that maximize agreement between paired samples while pushing apart unpaired ones. **Van den Oord, Li, and Vinyals (2018)** introduced InfoNCE (Information Noise-Contrastive Estimation) in "Representation Learning with Contrastive Predictive Coding," establishing the mathematical foundation now underlying most multimodal systems. (arXiv) The InfoNCE loss optimizes a lower bound on mutual information between representations: (arXiv)

$$\mathcal{L}_{NCE} = -\mathbb{E}\left[\log \frac{\exp(f(x_{t+k}, c_t))}{\sum_j \exp(f(x_j, c_t))}\right]$$

This formulation found its most influential application in **CLIP (Radford et al., ICML 2021)**, "Learning Transferable Visual Models From Natural Language Supervision," which trained on 400 million image-text pairs (arXiv) using symmetric cross-entropy over cosine similarities. CLIP computes an N×N similarity matrix between batch elements and applies cross-entropy in both directions (image-to-text and text-to-image), with a **learned temperature parameter τ** controlling the sharpness of the distribution. (Machinecurve) This symmetric formulation ensures neither modality dominates the optimization objective.

Lin et al. (WACV 2023) proposed Relaxed Contrastive (ReCo) loss in "Relaxing Contrastiveness in Multimodal Representation Learning," addressing InfoNCE's limitation of penalizing all negative pairs equally. ReCo uses the formulation $\mathcal{L}_{RC} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \max(0, C_{ij})^2$, which does not penalize orthogonal negative pairs, improving performance particularly in medical imaging tasks where semantic similarity varies continuously.

Uncertainty weighting for automatic loss balancing

Manual tuning of loss weights across modalities remains impractical at scale. **Kendall, Gal, and Cipolla (CVPR 2018)** introduced homoscedastic uncertainty weighting in "Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," deriving task weights from learned uncertainty parameters:

$$\mathcal{L}_{total} = \sum_i \left[\frac{1}{2\sigma_i^2} \mathcal{L}_i(W) + \log \sigma_i \right]$$

Here σ represents task-dependent uncertainty that the network learns jointly with task parameters. The $\log \sigma$ regularization term prevents the trivial solution of infinite uncertainty. This approach eliminates hyperparameter tuning and outperforms grid search on joint semantic segmentation, instance segmentation, and depth regression tasks.

Gradient-based loss balancing through GradNorm and PCGrad

Chen et al. (ICML 2018) proposed GradNorm in "Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks," which dynamically normalizes gradient magnitudes across tasks. (arXiv) GradNorm introduces a single hyperparameter α controlling asymmetry in task balancing, computing modulation based on relative inverse training rates to ensure all tasks converge at similar speeds. (Patrick-llgc)

Yu et al. (NeurIPS 2020) addressed gradient conflicts directly in "Gradient Surgery for Multi-Task Learning" with PCGrad (Projecting Conflicting Gradients). (Substack) When gradients from different tasks point in opposing directions (negative cosine similarity), PCGrad projects each gradient onto the normal plane of the other: $g_i' = g_i - \frac{g_i \cdot g_j}{\|g_j\|^2} g_j$. (NeurIPS) This "gradient surgery" eliminates destructive interference while preserving beneficial gradient components.

Auxiliary losses preventing modality collapse

Modality collapse—where multimodal representations degenerate to rely on a single modality—requires explicit countermeasures. **Javaloy et al. (ICML 2022)** addressed this in "Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization," introducing an impartiality block that modifies gradient computation to prevent dominant modalities from suppressing others in heterogeneous likelihood scenarios.

Recent work by **Liu et al. (arXiv 2025)** in "Principled Multimodal Representation Learning" optimizes the dominant singular value of the representation matrix using a softmax-based loss over singular values, combined with instance-wise contrastive regularization on leading eigenvectors. This achieves simultaneous alignment of multiple modalities without anchor dependency and provides more stable training than volume-based methods.

---

2. Gradient dynamics reveal why multimodal networks fail

The modality competition theorem

Huang et al. (ICML 2022) provided the first rigorous proof of modality competition in "Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably)." (arXiv) They proved that for late-fusion networks with smoothed ReLU activation trained by gradient descent, **different modalities compete and only a subset of encoder networks learn useful feature representations**. (Proceedings of Machine Learni…) "Losing" modalities fail to be discovered during optimization, causing systematic underperformance compared to well-trained unimodal baselines. (arXiv)

The theoretical analysis reveals that networks visit **saddle manifolds corresponding to unimodal solutions during early training**—a "prime learning window" phenomenon where the network commits to one modality before adequately exploring others. This explains the empirical observation that multimodal networks with more parameters often underperform simpler unimodal networks.

Greedy learning and conditional utilization

Wu et al. (ICML 2022) formalized the "greedy learner hypothesis" in "Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks," (thecvf) introducing the **conditional utilization rate**: the accuracy gain when accessing one modality in addition to another. (Proceedings of Machine Learni…) Their experiments across Colored MNIST, ModelNet40, and gesture recognition datasets revealed consistent imbalance—models extract information from the fastest-learning modality while underutilizing others. (NeurIPS) They proposed **conditional learning speed** as a training-time proxy enabling real-time rebalancing. (NYU Scholars)

Gradient magnitude imbalance across modalities

Peng et al. (CVPR 2022, Oral) demonstrated in "Balanced Multimodal Learning via On-the-fly Gradient Modulation" that dominant modalities produce larger gradients and converge faster, suppressing updates to weaker modality encoders. (arXiv) (arXiv) They introduced the **discriminative discrepancy ratio ρ** measuring imbalance between modalities' contributions to the classification objective. When ρ deviates from zero, one modality dominates learning.

Guo et al. (NeurIPS 2024) extended this analysis in "Classifier-guided Gradient Modulation for Enhanced Multimodal Learning" (CGGM), showing that **gradient direction misalignment** (negative cosine similarity between modality-specific gradients) compounds magnitude imbalance. Previous methods addressing only magnitude fail when gradients point in opposing directions. (GitHub)

On-the-fly Gradient Modulation with Generalization Enhancement (OGM-GE)

OGM-GE, introduced by **Peng et al. (CVPR 2022)**, dynamically monitors contribution discrepancy and modulates gradients accordingly. (GitHub +2) The modulation coefficient $k_m = 1 - \tanh(\alpha \cdot \rho_m)$ weakens gradients of the dominant modality, allowing under-optimized modalities to catch up. Crucially, the

Generalization Enhancement (GE) component adds dynamic Gaussian noise to compensate for reduced SGD noise intensity from gradient modulation— arXiv addressing the finding that modulation inadvertently harms generalization by reducing stochastic gradient noise strength.

The extended TPAMI version (**Wei et al., 2024**) added **On-the-fly Prediction Modulation (OPM)**, which drops dominant modality features with dynamic probability during the feed-forward pass, addressing imbalance from both forward and backward perspectives. arXiv

MMPareto resolves gradient conflicts through Pareto optimization

Wei and Hu (ICML 2024) discovered in "MMPareto: Boosting Multimodal Learning with Innocent Unimodal Assistance" that conventional Pareto optimization methods fail in multimodal settings—performing worse than uniform baselines. Proceedings of Machine Learni… The cause: multimodal gradients have smaller magnitude and lower batch sampling covariance than unimodal gradients, meaning Pareto solutions reduce SGD noise strength and harm generalization. arXiv

MMPareto resolves this through a two-case algorithm: when gradients do not conflict (cosine similarity ≥ 0), it equally weights gradients to enhance noise strength; when gradients conflict, it first finds a Pareto-optimal direction, then scales magnitude to maintain generalization. arXiv This achieves **flatter minima** in loss landscape analysis, linked to better generalization. Results on CREMA-D improved from 66.13% baseline to 75.13%. Liner

CGGM modulates both gradient magnitude and direction

Guo et al. (NeurIPS 2024) introduced Classifier-guided Gradient Modulation (CGGM), the first method addressing both magnitude and direction of multimodal gradients. CGGM employs modality-specific classifiers to evaluate utilization rates and uses their gradients to guide optimization direction, ResearchGate maximizing cosine similarity between the fusion gradient and weighted unimodal classifier gradients. Liner Unlike OGM-GE, CGGM places no limitations on loss functions, optimizers, or the number of modalities, arXiv achieving state-of-the-art results on UPMC-Food 101 (92.94%) and CMU-MOSI sentiment analysis.

---

3. Theoretical foundations explain when multimodal learning succeeds

Provable multimodal superiority under latent space assumptions

Huang et al. (NeurIPS 2021) provided the first rigorous proof that multimodal learning achieves smaller population risk than unimodal subsets in "What Makes Multi-Modal Learning Better than Single (Provably)." Under an encode-to-latent-space-then-map-to-task-space framework, multiple modalities provide more accurate estimates of shared latent representations. OpenReview The key condition: **modalities must contribute complementary (not purely redundant) information about a common latent space**.

This theoretical framework explains both successes and failures: when modalities share a latent representation and encoding functions are sufficiently expressive, multimodal learning provably outperforms. When these

conditions fail—or when training dynamics prevent discovery of complementary information—multimodal systems underperform.

Saddle point dynamics in high-dimensional multimodal landscapes

Dauphin et al. (NeurIPS 2014) established in "Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization" that saddle points dominate high-dimensional loss landscapes. Using arguments from statistical physics and random matrix theory, they showed critical points are exponentially more likely to be saddles than local minima in high dimensions. (arXiv) Multimodal networks, with their high-dimensional parameter spaces from multiple encoders, face proliferating saddle points. (Pronod's Blog)

The strict saddle property, characterized by **Ge et al. (COLT 2015)**, provides escape guarantees: if all saddle points have at least one negative curvature direction, noisy gradient descent finds local minima in polynomial time. (Medium) However, **Kawaguchi et al. (2017)** hypothesized that deep networks may converge to "highly degenerate" saddle points with many zero eigenvalues, complicating the theoretical picture. (arXiv)

Modality gap as a persistent phenomenon

Liang et al. (NeurIPS 2022) explained the modality gap in "Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Learning" through three mechanisms: (1) deep architectures create a "cone effect" restricting embeddings to narrow regions, (2) different random initializations place modalities in different cones, and (3) contrastive objectives preserve rather than eliminate this gap. Their theorem shows each neural network layer shrinks angles between embedding pairs with high probability, explaining why the gap persists even with extensive training. (NeurIPS)

Convergence through multi-objective optimization formulations

Recent work reformulates multimodal learning as multi-objective optimization (MOO). **Fernando et al. (2024)** in "Mitigating Modality Imbalance via Multi-Objective Optimization" (MIMO) optimizes the worst-performing unimodal objective while constraining multimodal loss, achieving ~20× reduction in computation versus existing MOO methods while providing convergence guarantees to stationary points under Lipschitz smoothness assumptions. (arXiv)

Learning rate scheduling across modalities

Yao and Mihalcea (ACL Findings 2022) demonstrated in "Modality-specific Learning Rates for Effective Multimodal Additive Late-fusion" that global learning rates cause vanishing gradients for some modalities due to differences in modality nature and computational flow. They proposed MSLR strategies: fixed ratios (MSLR-K), validation-scaled (MSLR-S), and dynamic adjustment (MSLR-D). (ACL Anthology)

Guerra-Manzanares and Shamout (2025) introduced MILES (Modality-Informed Learning Rate Scheduler), which uses conditional utilization rates to dynamically adjust per-modality learning rates, outperforming seven state-of-the-art baselines across four multimodal tasks without requiring offline unimodal pretraining. (arXiv)

---

4. Modality balancing techniques restore learning equilibrium

Gradient-Blending based on overfitting-to-generalization ratios

Wang, Tran, and Feiszli (CVPR 2020) identified the fundamental problem in "What Makes Training Multi-Modal Classification Networks Hard?": different modalities overfit and generalize at different rates, and standard regularizers prove ineffective. They introduced the **Overfitting-to-Generalization Ratio (OGR)** (thecvf) and proposed G-Blending, which computes optimal blending weights (IJCAI) $w_k^* = \frac{1}{Z}\frac{G_k}{O_k^2}$ where G represents generalization ability and O represents overfitting tendency.

Online G-Blend, which recomputes weights periodically, achieved 75.8% on Kinetics versus 71.4% for naive fusion and 72.6% for the best unimodal (RGB-only). (TheCVF) The method reached state-of-the-art on Kinetics (83.3% with IG-65M pretraining), EPIC-Kitchen, and AudioSet. (thecvf)

Prototypical Modal Rebalance targets slow-learning modalities directly

Fan et al. (CVPR 2023) observed in "PMR: Prototypical Modal Rebalance for Multimodal Learning" that existing methods modulate based on fused representations dominated by the better modality. (arXiv) PMR instead stimulates **specific slow-learning modalities without interference** using prototypes representing general class features. (Wenchao Xu) (arXiv) Prototype Cross-Entropy (PCE) loss accelerates clustering of slow modalities while Prototype Entropy Regularization (PER) penalizes dominant modalities during early training.

Adaptive Gradient Modulation through Shapley values

Li et al. (ICCV 2023) proposed AGM in "Boosting Multi-modal Model Performance with Adaptive Gradient Modulation," using **Shapley value-based attribution** to isolate individual modality contributions. AGM identifies "mono-modal" states as competition-free references and measures competition strength to determine modulation intensity. A key finding: the model's "preferred modality" (with lowest competition strength) need not be the one that appears dominant in accuracy metrics. (ResearchGate)

Diagnosing and re-learning for intrinsic modality limitations

Wei et al. (ECCV 2024) raised an important caveat in "Diagnosing and Re-learning for Balanced Multimodal Learning": not all modalities deserve equal emphasis. **Scarcely informative modalities pushed by balancing methods may counterproductively lose efficacy.** Their approach diagnoses modality separability and implements targeted re-initialization of uni-modal encoders based on learning state, respecting intrinsic capacity limitations.

Modality dropout as implicit regularization

Modality dropout—stochastically excluding entire modalities during training—provides multiple benefits: preventing over-reliance on dominant modalities, acting as implicit ensembling, and enabling robustness to missing modalities at inference. Recent advances include **Masked Modality Projection (MMP)**, which projects available modality tokens to missing modality representations, and **Dynamic Modality Scheduling (DMS)**, which weights contributions based on predictive confidence and uncertainty. (arXiv)

ReconBoost applies boosting principles to modality reconciliation

Hua et al. (ICML 2024) introduced ReconBoost in "Boosting Can Achieve Modality Reconcilement," using alternating updates for each modality with KL divergence-based coordination between modality learners. Memory consolidation and global aggregation mechanisms prevent catastrophic forgetting while achieving balanced learning across modalities. (arXiv)

---

5. Architecture determines gradient flow and modality interaction

Early fusion provides noise robustness at the cost of modality specialization

Barnum, Talukder, and Yue (NeurIPS 2020 Workshop) demonstrated in "On the Benefits of Early Fusion in Multimodal Representation Learning" that initial-layer fusion outperforms late-layer fusion and provides robustness across signal-to-noise ratios. Early fusion encourages models to utilize both modalities more effectively, (openreview) though it sacrifices the specialized encoders possible with late fusion. (OpenReview)

Wang et al. (NeurIPS 2020) proposed parameter-free fusion through "Deep Multimodal Fusion by Channel Exchanging," using batch normalization scaling factors γ to measure channel importance and dynamically exchange channels between modality-specific sub-networks. (NeurIPS) Their theoretical analysis proves linear combinations of cross-modal channels can decrease or maintain loss.

Multimodal transformers achieve efficiency through architectural innovation

ViLT (Kim, Son, and Kim, ICML 2021) demonstrated that heavy visual feature extraction is unnecessary by processing images as patch embeddings identical to text tokens, (Papertalk) achieving **10× speedup** over previous vision-language pretraining models. (Semantic Scholar) (Hugging Face) This finding reshaped the field toward lightweight fusion.

BLIP-2 (Li et al., ICML 2023) introduced the Querying Transformer (Q-Former) bridging frozen image encoders and LLMs with **54× fewer trainable parameters** than Flamingo80B. The two-stage pretraining—vision-language representation learning followed by vision-to-language generative learning—exemplifies efficient multimodal training through frozen backbone strategies. (arXiv)

FLAVA (Singh et al., CVPR 2022) unified vision, language, and cross-modal tasks through Masked Multimodal Modeling (MMM), jointly masking image patches and text tokens. (TheCVF) The architecture jointly optimizes image-text contrastive (ITC), image-text matching (ITM), and MMM objectives.

Transformer training exhibits distinct phases

Yang et al. (NeurIPS 2024) analyzed transformer gradient dynamics in "Training Dynamics of Transformers to Recognize Word Co-occurrence," discovering an **"automatic balancing of gradients"** property and two-phase training: Phase 1 where MLPs align with target signals while attention remains unchanged, and Phase 2 where

attention and MLPs evolve jointly. (arXiv) This phase structure has implications for multimodal training schedules.

Song et al. (NeurIPS 2024) in "Unraveling the Gradient Descent Dynamics of Transformers" proved that with appropriate initialization, gradient descent achieves global optimality with either softmax or Gaussian attention kernels. (NeurIPS) Notably, **Gaussian kernels enable convergence in cases where softmax fails**, showing faster convergence rates experimentally. (Amazon)

Mixture of Experts enables modality-specialized routing

Mixture-of-Experts with Expert Choice Routing (NeurIPS 2022) from Google Research inverted traditional routing by having experts select top-k tokens rather than tokens selecting experts, achieving **2×+ training efficiency** over GShard and Switch Transformer while addressing load imbalance. (Google Research)

MoME (Shen et al., NeurIPS 2024) introduced Mixture of Vision Experts (MoVE) with adaptive deformable transformation and Mixture of Language Experts (MoLE) with sparsely-activated adapters, using instance-level soft routing. Statistical analysis confirms specialization emerges in both vision and language branches, addressing task interference inherent to multimodal generalist models. (NeurIPS)

MoE-Fusion (Cao et al., ICCV 2023) combined Mixture of Local Experts (MoLE) for local features with Mixture of Global Experts (MoGE) guided by multimodal gates, enabling dynamic fusion that adapts to changing modality importance across samples. (TheCVF)

Architectural principles for stable multimodal training

The literature suggests several design principles for training stability:

- **Frozen backbones** (BLIP-2): Keep pretrained encoders frozen; train only lightweight connectors
- **Gated cross-attention** (Flamingo): Gradually inject cross-modal information via learned gates initialized near zero
- **Contrastive pre-alignment** (ALBEF, Li et al., NeurIPS 2021): Align representations through contrastive loss before fusion (arXiv)
- **Alternating training** (MLA, Zhang et al., CVPR 2024): Process modalities in separate iterations to eliminate interference (TheCVF)

Conclusion: Toward principled multimodal optimization

The field has progressed from observing multimodal failures to understanding their theoretical foundations and developing principled solutions. **Modality competition is now provable**, gradient imbalance is quantifiable through discriminative discrepancy ratios, and interventions like OGM-GE, MMPareto, and CGGM provide theoretically grounded remedies.

Key advances include the recognition that **gradient magnitude alone is insufficient**—direction matters equally (CGGM); (Liner) that **Pareto optimization requires adaptation** to preserve SGD noise strength (MMPareto); and that **not all modalities deserve equal emphasis** when intrinsic informativeness varies (D&R). Architecturally, the field has moved toward frozen backbones with lightweight fusion modules, MoE-based specialization, and attention mechanisms with explicit gating for stable gradient flow.

Open challenges remain: scaling gradient modulation to more than two modalities efficiently, theoretical guarantees under realistic neural network conditions, and understanding the interaction between architectural choices and optimization dynamics. The convergence of optimization theory, gradient analysis, and architectural innovation points toward increasingly principled multimodal learning systems that reliably outperform their unimodal constituents.