BEYOND MODALITY-LEVEL COMPETITION: FREQUENCY-RESOLVED SPECTRAL REGULATION FOR MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modality competition is a persistent challenge in multimodal learning, where dominant modalities can suppress the effective learning of weaker ones. Existing methods typically assess and mitigate competition at the modality-level, overlooking heterogeneous competitive states within latent representations. By examining latent representations from a frequency-domain perspective, we find that modality preference varies substantially across spectral regions and can even reverse across frequency bands during optimization, exposing competitive heterogeneity that modality-level assessment can obscure and uniform regulation cannot effectively address. Motivated by this observation, we propose **SpecGuide**, a frequency-resolved method that mitigates modality competition at the spectral level. Specifically, SpecGuide derives band-wise preference from class-conditioned spectral discriminative margins, uses the resulting preference maps to guide adaptive spectral shaping, and encourages frequency-wise complementarity across modality-specific spectral responses. Extensive experiments on six benchmarks demonstrate that SpecGuide consistently achieves state-of-the-art performance, delivers gains across different backbones, and remains robust to missing and noise-corrupted modalities.

1 INTRODUCTION

Multimodal models integrate complementary information from heterogeneous data sources to overcome the ambiguity inherent in individual modalities, enabling substantial progress in applications such as autonomous driving (Liao et al. (2025); Hwang et al. (2025)) and medical diagnosis (Zhang et al. (2025); Li et al. (2024)). Unfortunately, multimodal models often suffer from the challenge of ***modality competition*** (Huang et al. (2022),Peng et al. (2022)) during joint optimization: the easier-to-learn modality tends to gain an early optimization advantage, which biases the shared gradient signal toward the stronger modality and progressively suppresses the optimization of the weaker one (Fan et al. (2023); Wei et al. (2025b); Zhang et al. (2024)).

Existing studies primarily mitigate modality competition through an assess-then-regulate paradigm, where modality preference is first estimated and subsequently used to guide multimodal optimization. Most of these approaches intervene at the optimization level by modulating modality-specific gradients (Lu et al. (2026); Wei et al. (2025c); Li et al. (2023)) or introducing additional learning objectives (Wei et al. (2025a); Kontras et al. (2026); Wei et al. (2025b)). Since such regulation acts through parameter updates, its influence on latent representations[1] is indirect and takes effect only in subsequent forward passes, limiting its ability to correct competitive states already present within the current representations. This limitation motivates direct assessment and regulation of modality competition within representations (Fan et al. (2023); Nguyen et al. (2024); Wei et al. (2025c)). Nevertheless, existing representation-level methods typically perform assessment or regulation at the granularity of entire representations, leaving fine-grained competitive heterogeneity within representations largely unexplored.

This naturally raises a fundamental question: **does the preference between modalities remain consistent across different regions of their representations?** To investigate this question, we

[1]Hereafter referred to as "representation".

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

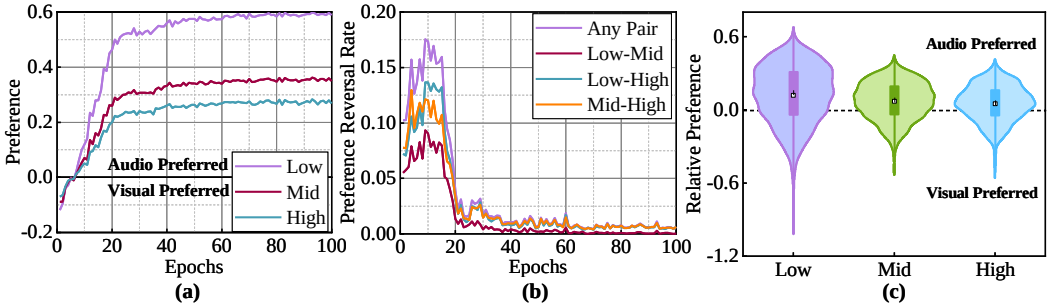


Figure 1: Frequency-resolved modality preference during joint training. (a) Evolution of modality preference across low-, middle-, and high-frequency regions. (b) Cross-band preference reversal rates over training. (c) Preference distributions across frequency bands; violin width indicates distribution density, boxes show the interquartile range, and white markers denote the mean.

examine representations in the frequency domain, where discriminative information can be resolved across distinct spectral regions. Previous studies (Cai et al. (2017); Ma et al. (2019)) have shown that different modalities rely on different spectral components for discrimination. This suggests that their discriminative advantages may also vary across frequency regions, leading to frequency-dependent modality preference.

Fig. 1 presents the frequency-resolved preference dynamics throughout joint training. We observe distinct behaviors across spectral regions: although all three bands eventually favor the audio modality, their preference strengths remain markedly different. More importantly, Fig. 1(b) reveals persistent cross-band preference reversals, which occur more frequently during early training but remain observable as optimization proceeds, indicating that the preference between modalities can reverse across different spectral regions of their representations. The preference distributions in Fig. 1(c) provide complementary evidence, showing marked variations in preference across frequency bands.

These observations show that a modality-level preference may miss both cross-band preference reversals and substantial differences in preference strength. This motivates finer-grained assessment and regulation of modality competition across frequency regions. Accordingly, we propose **SpecGuide**, a frequency-resolved method that diagnoses and mitigates modality competition across spectral regions within latent representations. Specifically, Frequency-wise Modality Preference (FMP) estimator measures cross-modal preference separately across frequency bands and constructs the corresponding frequency-wise guidance maps. Conditioned on these maps and the current spectral content, Preference-Guided Spectral Shaping (PGSS) adaptively shapes modality-specific spectral responses. Frequency-wise Complementarity Guidance (FCG) further promotes complementary spectral responses across modalities within individual frequency regions. Extensive experimental results demonstrate that SpecGuide achieves SOTA performance across six benchmarks and different backbone architectures, while maintaining strong robustness to missing and noise-corrupted inputs.

In summary, the main contributions of this paper are as follows: (1) We reveal frequency-wise heterogeneity in modality preference within representations, showing that different frequency bands can exhibit distinct and even reversed preferences. (2) We propose SpecGuide, a frequency-resolved method that assesses and mitigates modality competition across frequency regions within representations. (3) We conduct extensive experiments across diverse benchmarks and backbone architectures, demonstrating SOTA performance and strong robustness to missing and noise-corrupted modalities.

2 Frequency-Domain Analysis of Modality Competition

Problem Formulation. Following previous studies (Wei et al. (2025c;b)), we consider a multimodal classification problem with M modalities and C classes. Given a dataset $\mathcal{T} = \{(x_i^1, \ldots, x_i^M, y_i)\}_{i=1}^N$ with $y_i \in \{1, \ldots, C\}$, a modality-specific encoder E^m maps each input x_i^m to its representation $z_i^m = E^m(x_i^m)$, where $m = 1, \ldots, M$.

As shown in Fig. 1, the preference between modalities varies substantially across frequency bands, suggesting that cross-modal competitive states are not uniform across spectral regions. Prior studies (Rahaman et al. (2019); John Xu et al. (2020); Wang et al. (2020; 2023)) have shown that neural

networks exhibit frequency-dependent learning behaviors and exploit discriminative cues differently across spectral regions. In this section, we analyze how discriminative information is distributed across frequency regions and how the resulting discriminative advantages vary across modalities.

Frequency-wise Discriminative Heterogeneity. Let the spectrum of z_i^m be partitioned into S non-overlapping frequency regions $\{\Omega_s\}_{s=1}^S$, and let $h_{i,m}^s \in \mathbb{R}^D$ denote the spectral descriptor extracted from region Ω_s. For class c, let $\mu_{m,c}^s = \mathbb{E}[h_{i,m}^s \mid y_i = c]$ denote its mean descriptor and Σ_m^s the pooled within-class covariance in the spectral representation space.

Following the Fisher criterion (Fisher (1936)), we consider a non-zero projection direction $w \in \mathbb{R}^D$ and characterize the pairwise discriminability between classes c and c' within region s as the maximum ratio of projected mean separation to pooled within-class variation:

$$\mathcal{D}_{m,c,c'}^s = \max_{w \neq 0} \frac{\left[w^\top(\mu_{m,c}^s - \mu_{m,c'}^s)\right]^2}{w^\top(\Sigma_m^s + \epsilon I)w} = (\mu_{m,c}^s - \mu_{m,c'}^s)^\top(\Sigma_m^s + \epsilon I)^{-1}(\mu_{m,c}^s - \mu_{m,c'}^s), \quad (1)$$

where $\epsilon > 0$ ensures positive definiteness and stabilizes covariance inversion. More detailed derivations can be found in Appendix 6.2.1. Furthermore, we summarize the discriminative state of modality m within region s by averaging over all class pairs:

$$\mathcal{D}_m^s = \frac{2}{C(C-1)} \sum\nolimits_{c<c'} \mathcal{D}_{m,c,c'}^s. \quad (2)$$

where $\mathcal{D}_m^s$ characterizes class separation relative to within-class variation within a spectral region.

To connect band-wise and overall discriminability, let $h_{i,m} = [(h_{i,m}^1)^\top, \ldots, (h_{i,m}^S)^\top]^\top$ denote the concatenated spectral descriptor and define $\mathcal{D}_m$ analogously on $h_{i,m}$. When cross-band within-class correlations are relatively weak, we approximate $\Sigma_m \approx \mathrm{diag}(\Sigma_m^1, \ldots, \Sigma_m^S)$. Under this approximation, the pairwise Fisher discriminability decomposes across spectral regions, and averaging over all class pairs gives:

$$\mathcal{D}_m \approx \sum\nolimits_{s=1}^S \mathcal{D}_m^s. \quad (3)$$

The complete derivation is provided in Appendix 6.2.2. Different modalities can rely on distinct spectral regions for discrimination (Wang et al. (2020; 2023); Wu et al. (2018); Aryal & Lee (2023)). Such spectral heterogeneity is reflected in $\mathcal{D}_m^s$, which depends on modality- and region-specific class separation and within-class variation, allowing the relative discriminative advantage between modalities to vary across frequency regions.

Granularity Limitation of Modality-Level Assessment and Regulation. Existing methods assess modality competition using diverse criteria, including prediction confidence (Wei et al. (2025c)), mutual information (Yang et al. (2025a)), cluster purity (Wei et al. (2024b)), prototype-based discriminability (Fan et al. (2023)), and spectrum energy (Lu et al. (2026)), but ultimately operate at the modality level. Taking the frequency domain as a concrete instance, the analysis above shows that relative discriminative advantages can vary substantially across spectral regions, revealing heterogeneity obscured by global assessment.

For any two modalities m_1 and m_2, we define their discriminative preference within frequency region s as:

$$P_{m_1,m_2}^s = \mathcal{D}_{m_1}^s - \mathcal{D}_{m_2}^s, \quad (4)$$

where $P_{m_1,m_2}^s > 0$ indicates stronger discriminability of modality m_1, while $P_{m_1,m_2}^s < 0$ indicates stronger discriminability of modality m_2. Accordingly, $P_{m_1,m_2}^{s_1} P_{m_1,m_2}^{s_2} < 0$ represents a preference reversal between frequency regions s_1 and s_2.

Following Eq. 3, the global relative discriminative advantage between modalities can be expressed as:

$$P_{m_1,m_2}^{\text{global}} \triangleq \mathcal{D}_{m_1} - \mathcal{D}_{m_2} \approx \sum\nolimits_{s=1}^S P_{m_1,m_2}^s. \quad (5)$$

Aggregating frequency-wise preferences into a single modality-level scalar can obscure their heterogeneity. Preferences with opposite signs may partially cancel, such that a small global preference can coexist with strong region-specific preferences, while differences in preference strength remain unresolved even when all regions favor the same modality. Consequently, modality-level assessment may overlook frequency-wise variations in both preference direction and magnitude.

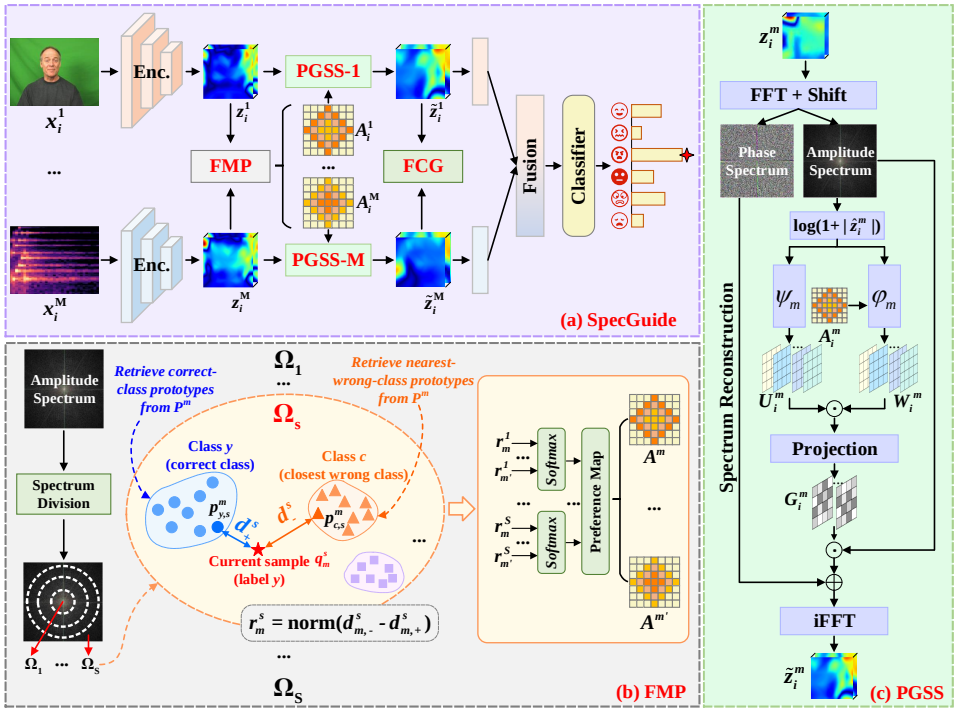


Figure 2: **Overview of SpecGuide.** (a) Overall framework. (b) Schematic illustration of FMP. (c) Schematic illustration of PGSS. More details of FMP and PGSS are provided in Appendix 6.4.

This granularity limitation also extends to regulation. Let $r_m^s \in \mathbb{R}$ denote the desired regulation intensity for modality m in frequency region s, while r_m denotes a single modality-level intensity shared across all regions. Their mismatch is quantified as:

$$\mathcal{E}(r_m) = \sum_{s=1}^{S} \eta_s (r_m - r_m^s)^2, \tag{6}$$

where $\eta_s \geq 0$ and $\sum_{s=1}^{S} \eta_s = 1$. Minimizing Eq. 6 gives:

$$r_m^* = \sum_{s=1}^{S} \eta_s r_m^s, \qquad \mathcal{E}(r_m^*) = \sum_{s=1}^{S} \eta_s (r_m^s - \sum_{s=1}^{S} \eta_s r_m^s)^2 = \text{Var}_\eta(r_m^s). \tag{7}$$

When the required regulation intensities differ across positively weighted frequency regions, the minimum approximation error is strictly positive, indicating that a shared modality-level correction cannot simultaneously match all region-specific targets. The complete derivation is provided in Appendix 6.2.3, motivating frequency-wise assessment and regulation.

Empirical Evidence for Frequency-wise Heterogeneity. Fig. 1(a) shows that the relative preference between modalities follows distinct trajectories across low-, middle-, and high-frequency regions. Fig. 1(b) further reveals persistent cross-region preference reversals, peaking in the early training stage. Notably, this stage overlaps with the **prime learning window** identified by InfoReg (Huang et al. (2025)), where modality imbalance strongly influences subsequent learning, underscoring the importance of resolving frequency-wise preference heterogeneity. Fig. 1(c) further shows markedly different preference distributions across frequency regions, demonstrating heterogeneity in both preference direction and magnitude. These results provide empirical evidence that cross-modal preference is frequency-dependent and motivate frequency-resolved assessment and regulation.

3 Proposed Method

3.1 Overview

Motivated by the analysis in Sec. 2, we propose SpecGuide, a frequency-resolved method that directly diagnoses and mitigates modality competition within representations. As illustrated in Fig. 2,

Algorithm 1 Training and Inference of SpecGuide

Require: Dataset $\mathcal{T}$; multimodal model parameters Θ.
Training: Initialize class-conditioned spectral prototypes $\{p^s_{m,c}\}$.
for each training epoch **do**
  for each mini-batch $\{(x^m_i)^M_{m=1}, y_i\}$ **do**
    Extract modality-specific representations $z^m_i = E^m(x^m_i)$.
    Estimate frequency-wise preferences $\{\alpha^s_{i,m}\}$ and construct preference maps $\{A^m_i\}$.
    Obtain spectrally modulated representations $\tilde{z}^m_i = \text{PGSS}(z^m_i; A^m_i)$.
    Update the class-conditioned spectral prototypes.
    Optimize Θ with $\mathcal{L}$ in Eq. 16.
  end for
  Store the epoch-wise mean frequency preferences $\{\bar{\alpha}^s_m\}$ as the inference prior.
end for
Inference: Construct preference maps $\{\bar{A}^m\}$ from the stored inference prior.
for each test sample $\{x^m_i\}^M_{m=1}$ **do**
  Extract $z^m_i = E^m(x^m_i)$ and obtain $\tilde{z}^m_i = \text{PGSS}(z^m_i; \bar{A}^m)$.
  Fuse $\{\tilde{z}^m_i\}^M_{m=1}$ and predict $\hat{y}_i$.
end for

SpecGuide augments a standard multimodal model with three components: FMP, PGSS, and FCG. Specifically, given the modality-specific representation z^m_i, FMP estimates the cross-modal preference over S spectral regions $\{\Omega_s\}^S_{s=1}$ and constructs the corresponding frequency-wise preference map A^m_i. Conditioned on the current spectral content and A^m_i, PGSS adaptively shapes the spectral response of each modality to obtain the modulated representation $\tilde{z}^m_i = \text{PGSS}(z^m_i; A^m_i)$, which is subsequently fused for multimodal prediction. FCG further promotes complementary spectral responses across modalities within individual frequency regions. The complete training and inference procedures are summarized in Algorithm 1.

3.2 Frequency-wise Modality Preference Estimation

To address the granularity limitation identified in Sec. 2, FMP estimates frequency-wise modality preference by measuring the class-conditioned spectral separability of each modality.

As shown in Fig. 2, given a representation z^m_i, we first obtain its centered spectrum $\hat{z}^m_i = \text{fftshift}(\text{FFT}(z^m_i))$ and partition it into S non-overlapping frequency regions $\{\Omega_s\}^S_{s=1}$. For each region, FMP constructs a channel-wise spectral descriptor $q^s_{i,m} \in \mathbb{R}^D$ from the normalized spectral-energy distribution and compares it with class-conditioned spectral prototypes. Let $d^s_{i,m,+}$ denote the distance from $q^s_{i,m}$ to the ground-truth class prototype and $d^s_{i,m,-}$ the distance to the nearest competing-class prototype. We define the normalized discriminative margin as:

$$r^s_{i,m} = \frac{d^s_{i,m,-} - d^s_{i,m,+}}{\gamma^s_{i,m} + \epsilon}, \tag{8}$$

where $\gamma^s_{i,m}$ denotes the prototype-distance scale in region s. A larger $r^s_{i,m}$ indicates stronger class separability of modality m within that frequency region.

We then compare modalities independently within each frequency region:

$$\alpha^s_{i,m} = \frac{\exp(r^s_{i,m}/\tau)}{\sum^M_{m'=1} \exp(r^s_{i,m'}/\tau)}, \tag{9}$$

where τ controls the sharpness of the preference distribution. The resulting $\alpha^s_{i,m}$ denotes the preference weight of modality m in region s. These band-wise preferences are mapped to their corresponding spectral locations to construct the frequency-wise preference map:

$$A^m_i(\omega) = \sum\nolimits_{s=1}^{S} \alpha^s_{i,m} \mathbf{1}_{\Omega_s}(\omega), \tag{10}$$

where $\mathbf{1}_{\Omega_s}(\omega)$ denotes the indicator function of region Ω_s. The resulting A^m_i preserves the preference associated with each frequency region and serves as the guidance signal for PGSS. Further calculation and implementation details are provided in Appendix 6.4.1, and the pseudocode of FMP is presented in Algorithm 2.

3.3 PREFERENCE-GUIDED SPECTRAL SHAPING

Guided by the frequency-wise preference map A_i^m produced by FMP, PGSS adaptively shapes the spectral response of each representation.

Given the centered spectrum $\hat{z}_i^m$, we use the log-compressed magnitude $X_i^m = \log(1 + |\hat{z}_i^m|)$ as the spectral input. PGSS first projects X_i^m into J learned spectral response channels through a lightweight projection ψ_m, while a preference-conditioning network φ_m jointly takes X_i^m and A_i^m to generate location-dependent response weights:

$$U_i^m = \psi_m(X_i^m), \qquad W_i^m = \text{Softmax}_J(\varphi_m([X_i^m; A_i^m])), \tag{11}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation. The weighted responses are then mapped through ϕ_m to produce the spectral modulation map:

$$G_i^m = \sigma(\phi_m(U_i^m \odot W_i^m)), \tag{12}$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ denotes the sigmoid function.

PGSS applies G_i^m to the spectral magnitude while retaining the original phase, allowing preference-guided modulation of spectral response strength without perturbing the phase structure. The choice of modulation components is further examined in Table 6. As shown in Appendix 6.2.4, the modulated spectrum preserves conjugate symmetry and thus admits a real-valued inverse Fourier reconstruction. Accordingly, the modulated representation is reconstructed as:

$$\tilde{z}_i^m = \Re\Big[\text{IFFT}\Big(\text{ifftshift}\Big((|\hat{z}_i^m| \odot G_i^m) e^{j\angle \hat{z}_i^m}\Big)\Big)\Big], \tag{13}$$

where $\Re$ denotes the real part. Further details of PGSS are provided in Appendix 6.4.2.

3.4 FREQUENCY-WISE COMPLEMENTARITY GUIDANCE

To promote complementary learning across modalities, FCG regularizes cross-modal spectral responses separately within each frequency region. Given the PGSS-modulated representation $\tilde{z}_i^m$, we partition its centered spectrum using the same regions $\{\Omega_s\}_{s=1}^S$ as FMP and construct a channel-wise spectral descriptor for each region:

$$c_{i,m}^s = \sqrt{\frac{1}{|\Omega_s|} \sum\nolimits_{\omega \in \Omega_s} \left|\hat{\tilde{z}}_i^m(\omega)\right|^2 + \epsilon}, \tag{14}$$

where $\hat{\tilde{z}}_i^m = \text{fftshift}(\text{FFT}(\tilde{z}_i^m))$ and $c_{i,m}^s \in \mathbb{R}^D$ denotes the channel-wise spectral descriptor of modality m in region s.

For two modalities m_1 and m_2, FCG penalizes their squared cosine similarity independently within each frequency region:

$$\mathcal{L}_{\text{FCG}} = \frac{1}{BS} \sum_{i=1}^{B} \sum_{s=1}^{S} \left(\frac{(c_{i,m_1}^s)^\top c_{i,m_2}^s}{\|c_{i,m_1}^s\|_2 \|c_{i,m_2}^s\|_2 + \epsilon} \right)^2, \tag{15}$$

where B denotes the mini-batch size.

By reducing cross-modal spectral similarity within individual frequency regions, FCG promotes frequency-wise complementary responses. We further examine whether such frequency-wise decorrelation may impair task-relevant cross-modal alignment. As shown in Appendix 6.3.4, FCG substantially reduces band-wise spectral similarity while maintaining comparable semantic alignment between modalities.

Training Objective. The overall training objective combines task supervision with frequency-wise complementarity:

$$\mathcal{L} = \lambda_{\text{CLS}} \mathcal{L}_{\text{CLS}} + \lambda_{\text{FCG}} \mathcal{L}_{\text{FCG}}, \tag{16}$$

where λ_{CLS} and λ_{FCG} control the contributions of the two objectives. The classification objective is defined as:

$$\mathcal{L}_{\text{CLS}} = \frac{1}{B} \sum_{i=1}^{B} \left[\text{CE}(o_i, y_i) + \sum\nolimits_{m=1}^{M} \text{CE}(o_i^m, y_i) \right], \tag{17}$$

where o_i and o_i^m denote the fused and modality-specific prediction logits, respectively.

Table 1: Comparison with different methods. The best and second-best results are highlighted in **red** and ***blue***, respectively.

Methods	CREMA-D	AVE	KS	UCF-101	Food-101	MOSI
Joint-Training	62.90	67.16	67.48	79.86	92.08	78.36
OGM-GE (CPVR, 2022)	69.35	67.91	67.67	80.48	92.72	78.86
MMPareto (ICML, 2024)	69.88	70.40	70.79	80.99	92.96	***79.45***
D&R (ECCV, 2024)	75.54	68.81	69.10	80.10	92.83	79.15
ReconBoost (ICML, 2024)	74.43	72.14	74.13	82.06	93.03	79.30
InfoReg (CPVR, 2025)	70.97	69.65	73.97	***82.27***	93.04	79.04
DGL (ICCV, 2025)	***76.08***	***72.39***	***74.83***	81.60	***93.14***	78.57
SpecGuide (Ours)	**77.96**	**73.88**	**75.70**	**83.18**	**93.59**	**82.07**

3.5 Differences from Previous Studies

SpecGuide differs from existing methods primarily in the granularity of preference assessment and regulation. Previous approaches typically characterize modality learning states using modality-level signals and regulate each modality as a whole (Peng et al. (2022); Li et al. (2023); Wei et al. (2025b); Huang et al. (2025); Hua et al. (2024); Wei et al. (2025a;c); Nguyen et al. (2024)). SpecGuide instead resolves preference across spectral regions and uses the resulting region-specific guidance to directly shape latent spectral responses. PMR (Fan et al. (2023)) and FRM (Lu et al. (2026)) are closely related to SpecGuide through prototype-based modeling and frequency-domain analysis, respectively, but operate at a coarser granularity. PMR constructs class prototypes from entire unimodal representations, whereas FMP maintains class-conditioned prototypes within individual spectral regions. FRM analyzes frequency information in the input space but aggregates spectral cues into a modality-level preference signal, whereas SpecGuide estimates region-specific preference within latent representations and preserves this structure for subsequent spectral shaping. FCG further promotes cross-modal complementarity within individual frequency regions.

4 Experiments

4.1 Datasets and Experimental Settings

Datasets. We evaluate SpecGuide on six multimodal benchmarks covering diverse modality combinations: CREMA-D (Cao et al. (2014)), Kinetics-Sounds (KS) (Arandjelovic & Zisserman (2017)), AVE (Tian et al. (2018)), UCF-101 (Soomro et al. (2012)), UPMC Food-101 (Wang et al. (2015)), and CMU-MOSI (Zadeh et al. (2018)). These benchmarks enable us to evaluate the generality of SpecGuide across heterogeneous multimodal settings. Further details are provided in Appendix 6.6.

Experimental Settings. Following previous studies (Wei et al. (2025c)), we adopt ResNet-18 as the default encoder unless otherwise specified. CREMA-D, KS, and AVE are trained from scratch, while the remaining benchmarks use task-appropriate pretrained encoders. All models are optimized using SGD with momentum 0.9 and a batch size of 64. To ensure a fair comparison, we train all methods using the same configuration. Implementation details are reported in Appendix 6.7.

4.2 Comparison with SOTA Methods

We compare SpecGuide with six SOTA methods, including OGM-GE (Peng et al. (2022)), MMpareto (Wei & Hu (2024)), D&R (Wei et al. (2024b)), ReconBoost (Hua et al. (2024)), InfoReg (Huang et al. (2025)), and DGL (Wei et al. (2025b)). As shown in Table 1, SpecGuide achieves the best performance on all benchmarks. The improvements span diverse modality combinations, backbones, and both two- and three-modality settings, demonstrating generality across architectures and multimodal configurations. These results support the effectiveness of our proposed SpecGuide.

4.3 Ablation Studies

Comparison of Different Fusion Strategies. Following previous work (Wei et al. (2025c)), we integrate SpecGuide with four representative fusion strategies: summation, concatenation, FiLM

Table 2: Comparison of different fusion strategies.

Fusion Strategies	CREMA-D	AVE
FiLM	64.11	67.16
FiLM + Ours	71.24	71.90
BiGated	65.32	67.41
BiGated + Ours	69.49	70.40
Summation	61.69	67.91
Summation + Ours	77.55	73.88
Concatenation	62.90	67.16
Concatenation + Ours	77.96	73.88

Table 3: Comparison of different representation modulation strategies. † and * denote spatial- and frequency-domain modulation, respectively.

Strategies	Audio	Visual	Multi
w/o Modulation	57.39	21.24	62.90
Random Dropout†	59.68	55.24	69.76
Learnable Weight†	***62.37***	***68.82***	73.25
OPM†	61.02	56.45	***77.28***
Random Dropout*	61.83	36.56	68.15
Learnable Weight*	59.68	68.55	74.87
PGSS (Ours)*	**63.04**	**69.22**	**77.96**

(Perez et al. (2017)), and BiGated (Kiela et al. (2018)). As shown in Table 2, SpecGuide consistently improves performance across all strategies and datasets. The consistent gains under both linear (summation and concatenation) and nonlinear (FiLM and BiGated) fusion mechanisms demonstrate the robustness of frequency-resolved representation regulation to different fusion designs, confirming the broad compatibility of SpecGuide with diverse multimodal fusion strategies.

Comparison of Different Modulation Strategies. We compare PGSS with representative spatial- and frequency-domain modulation strategies while keeping the remaining framework unchanged. Random Dropout masks features according to the estimated modality imbalance ratio, Learnable Weight directly learns modulation weights without frequency-wise preference guidance, and OPM (Wei et al. (2025c)) serves as a representative spatial-domain modulation method. For a controlled comparison, Learnable Weight adopts the same architecture and parameter budget as PGSS. As shown in Table 3, all modulation variants improve over the model without representation modulation. However, frequency-domain modulation alone does not consistently outperform its spatial counterpart, indicating that operating in the spectral domain is insufficient by itself. PGSS achieves the best multimodal performance, highlighting the importance of frequency-wise preference guidance for adaptive spectral shaping.

Comparison of Different Preference Estimators. To evaluate the effectiveness of FMP, we replace it with three representative modality-level preference estimators, including contribution-based (Wei et al. (2025c)), cluster-purity-based (Wei et al. (2024b)), and Mutual Information (MI)-based (Yang et al. (2025a)) metrics, while keeping the remaining framework unchanged. As shown in Table 4, all preference-guided variants substantially outperform Joint-Training, supporting the effectiveness of PGSS and FCG. Moreover, FMP achieves the best performance on both datasets, demonstrating that frequency-resolved preference provides more effective guidance than modality-level preference signals. Detailed implementations of the compared estimators are provided in Appendix 6.5.

Table 4: Comparison of different preference estimators

Estimators	CREMA-D	AVE
None (Joint-Training)	62.90	67.16
Contribution	76.08	***73.38***
Cluster Purity	***76.61***	72.89
Mutual Information	76.21	72.14
FMP (Ours)	**77.96**	**73.88**

Robustness under Imperfect Modalities. We further evaluate SpecGuide under two common forms of modality degradation: noise corruption and modality missing. For noise corruption, Gaussian white noise is added to the audio modality with varying signal-to-noise ratios, while Gaussian noise with increasing variance is injected into the visual modality. As shown in Figs. 3(a) and 3(b), SpecGuide consistently achieves the best performance across different corruption levels, indicating strong robustness to degraded modality quality. We further simulate modality missing by randomly dropping either the audio or visual modality at increasing ratios. As shown in Figs. 3(c) and 3(d), SpecGuide degrades substantially more gracefully than Joint-Training and OGM-GE and maintains strong performance under severe missing conditions. It achieves the best results across most settings, with particularly clear advantages under visual modality missing. These results suggest that

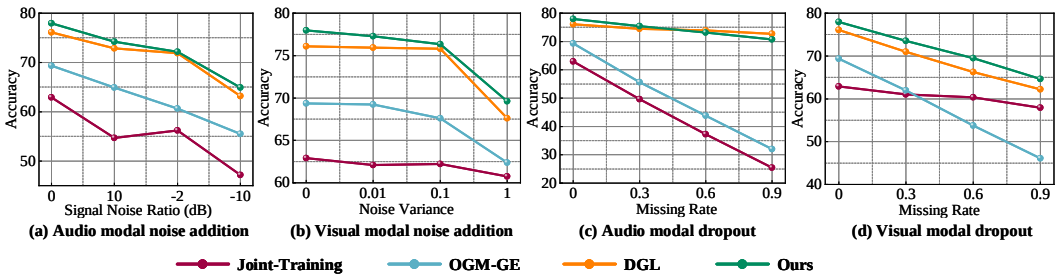


Figure 3: Robustness under imperfect modalities. (a)–(b) Noise-corrupted audio and visual modalities. (c)–(d) Missing audio and visual modalities.

frequency-resolved spectral shaping reduces excessive dependence on localized modality-specific cues and improves robustness to imperfect multimodal inputs.

Ablation of Key Components. Table 5 evaluates the contribution of each component in SpecGuide. With uniform preference maps, PGSS substantially improves the baseline, particularly on the visual branch, demonstrating the effectiveness of latent spectral shaping. Incorporating FMP further improves the visual and multimodal performance, showing that finer-grained preference estimation provides more informative guidance for spectral modulation. Adding FCG further improves the audio and multimodal performance while maintaining comparable visual accuracy, yielding the best multimodal result. These results demonstrate that PGSS, FMP, and FCG provide complementary contributions through spectral shaping, fine-grained guidance, and cross-modal spectral complementarity, respectively.

Table 5: Results of ablation of key components

Components	Audio	Visual	Multi
Baseline	57.39	21.24	62.90
+ PGSS	***62.90***	64.78	75.27
+ PGSS + FMP	61.42	**69.35**	***76.21***
Full SpecGuide	**63.04**	***69.22***	**77.96**

Ablation on Spectral Modulation Components. We investigate whether PGSS should modulate the amplitude, phase, or both components of the latent spectrum. As shown in Table 6, amplitude-only modulation achieves the best multimodal performance, while phase-only modulation results in a clear degradation. Jointly modulating amplitude and phase improves the visual branch slightly but remains inferior to amplitude-only modulation in multimodal performance. These results support our design of shaping spectral magnitude while retaining the original phase. Since the amplitude characterizes the response strength across frequencies, it provides a suitable space for preference-guided spectral shaping, whereas perturbing the phase can alter the spectral organization of the representation and weaken the resulting multimodal representation.

Table 6: Comparison of different spectral modulation targets.

Spectral Components		Accuracy		
Amplitude	Phase	Audio	Visual	Multi
✗	✗	57.39	21.24	62.90
✓	✓	59.41	**70.03**	***75.54***
✗	✓	***61.29***	61.42	72.31
✓	✗	**63.04**	***69.22***	**77.96**

5 Conclusion

We present SpecGuide, a frequency-resolved method for mitigating modality competition in multimodal learning. Our analysis reveals that modality preference can vary substantially across spectral regions and even reverse across frequency bands during optimization, exposing competitive states obscured by modality-level assessment and regulation. Motivated by this finding, SpecGuide mitigates modality competition at the frequency-level: FMP estimates band-specific preference, PGSS uses the resulting guidance to adaptively shape latent spectral responses, and FCG promotes cross-modal spectral complementarity within individual frequency regions. Extensive experiments across six multimodal benchmarks demonstrate SOTA performance, consistent improvements across different backbone architectures, and strong robustness to missing and noise-corrupted modalities. The limitations of SpecGuide are discussed in Appendix 6.8.

LLM USAGE STATEMENT

In this work, we only employed a Large Language Model (LLM) as an auxiliary tool to **polish the manuscript**. Specifically, its use was focused on enhancing the clarity and fluency of the writing to better articulate our core insights and findings. We do not use LLMs for tasks such as reference retrieval or figure generation. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the hyperparameters and environmental configurations for our main experiments are detailed in the manuscript, specifically within Sections 4.1, 6.6 and 6.7. Furthermore, the source code will be made publicly available in the next version.

IMPACT STATEMENT

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617, 2017. doi: 10.1109/ICCV.2017.73.

Nisan Aryal and Sang-Woong Lee. Frequency-based cnn and attention module for acoustic scene classification. *Applied Acoustics*, 210:109411, 2023.

Jiajun Cai, Qimin Cheng, Mingjun Peng, and Yang Song. Fusion of infrared and visible images based on nonsubsampled contourlet transform and sparse k-svd dictionary learning. *Infrared Physics & Technology*, 82:85–95, 2017.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7555–7577. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chaudhuri25a.html.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. ICML'23. JMLR.org, 2023.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038, 2023. doi: 10.1109/CVPR52729.2023.01918.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement, 2024. URL https://arxiv.org/abs/2405.09321.

Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. Adaptive unimodal regulation for balanced multimodal information acquisition, 2025. URL https://arxiv.org/abs/2503.18595.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably), 2022. URL https://arxiv.org/abs/2203.12221.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving, 2025. URL https://arxiv.org/abs/2410.23262.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020. ISSN 1991-7120. doi: 10.4208/cicp.oa-2020-0085. URL http://dx.doi.org/10.4208/cicp.OA-2020-0085.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Liang, Matthew Blaschko, and Maarten De Vos. Balancing multimodal training through game-theoretic regularization. *Advances in Neural Information Processing Systems*, 38:167953–167982, 2026.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22157–22167, 2023. doi: 10.1109/ICCV51070.2023.02030.

Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion, 2024. URL https://arxiv.org/abs/2412.08050.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047, 2025. doi: 10.1109/CVPR52734.2025.01124.

Siqi Lu, Wanying Xu, Yongbin Zheng, Wenting Luan, Peng Sun, and Jianhang Yao. Plug, play, and fortify: A low-cost module for robust multimodal image understanding models. *arXiv preprint arXiv:2602.22644*, 2026.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019.

Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. Ada2i: Enhancing modality balance for multimodal conversational emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9330–9339, 2024.

Haotian Ni, Yake Wei, Hang Liu, Gong Chen, Chong Peng, Hao Lin, and Di Hu. RollingQ: Reviving the cooperation dynamics in multimodal transformer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46288–46304. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ni25a.html.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation, 2022. URL https://arxiv.org/abs/2203.15332.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Chengxuan Qian, Kai Han, Jingchao Wang, Zhenlong Yuan, Chongwen Lyu, Jun Chen, and Zhe Liu. Dyncim: Dynamic curriculum for imbalanced multimodal learning, 2025. URL https://arxiv.org/abs/2503.06456.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos, 2018. URL https://arxiv.org/abs/1803.08842.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8681–8691. IEEE, 2020.

Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. What do neural networks learn in image classification? a frequency shortcut perspective. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1433–1442. IEEE, 2023.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, 2015. doi: 10.1109/ICMEW.2015.7169757.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning, 2025a. URL https://arxiv.org/abs/2507.10203.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning, 2025b. URL https://arxiv.org/abs/2507.10213.

Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance, 2024. URL https://arxiv.org/abs/2405.17730.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27328–27337, 2024a. doi: 10.1109/CVPR52733.2024.02581.

Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIV*, pp. 71–86, Berlin, Heidelberg, 2024b. Springer-Verlag. ISBN 978-3-031-73038-2. doi: 10.1007/978-3-031-73039-9_5. URL https://doi.org/10.1007/978-3-031-73039-9_5.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):469–485, 2025c. doi: 10.1109/TPAMI.2024.3468315.

Yu Wu, Hua Mao, and Zhang Yi. Audio classification using attention-augmented convolutional neural network. *Knowledge-Based Systems*, 161:90–100, 2018.

Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di 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)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096655.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4553–4566, 2025a. doi: 10.1109/TPAMI.2025.3547417.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025b. Curran Associates Inc. ISBN 9798331314385.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27446–27456, 2024. doi: 10.1109/CVPR52733.2024.02592.

Zheyu Zhang, Yayuan Lu, Feipeng Ma, Yueyi Zhang, Huanjing Yue, and Xiaoyan Sun. Incomplete multi-modal brain tumor segmentation via learnable sorting state space model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25982–25992, 2025. doi: 10.1109/CVPR52734.2025.02420.

6 APPENDIX

6.1 RELATED WORK

Modality competition is a major challenge in multimodal learning, where dominant modalities can hinder the effective optimization of weaker ones (Huang et al. (2022); Peng et al. (2022); Fan et al. (2023); Huang et al. (2025)). A major line of work mitigates competition through an assess-then-regulate paradigm, where modality preference is first estimated and subsequently used to guide multimodal optimization (Wei et al. (2024a); Qian et al. (2025); Yang et al. (2025a); Xu et al. (2023); Yang et al. (2025b); Wei et al. (2025a)). Representative studies characterize modality imbalance using metrics including unimodal prediction confidence (Peng et al. (2022); Wei et al. (2025c)), MI (Yang et al. (2025a)), Fisher information (Huang et al. (2025)), Shapley values (Wei et al. (2024a)), and prototype-based discriminability (Fan et al. (2023)). These estimators are then used to guide gradient modulation (Peng et al. (2022)), auxiliary supervision (Wei & Hu (2024)), knowledge distillation (Du et al. (2023)), and other optimization strategies (Wei et al. (2025c); Yang et al. (2025a); Nguyen et al. (2024)).

Recent studies have increasingly explored representation- and forward-level regulation to mitigate modality competition. OPM (Wei et al. (2025c)), PMR (Fan et al. (2023)), Ada2I (Nguyen et al. (2024)), and D&R (Wei et al. (2024b)) regulate modality learning through feature suppression, prototype-based stimulation, adaptive reweighting, and representation-space diagnosis, while MLA (Zhang et al. (2024)) and ReconBoost (Hua et al. (2024)) restructure multimodal optimization via alternating unimodal learning. Other approaches further intervene in internal representation structures, including representation-basis reallocation (Chaudhuri et al. (2025)) and cross-modal attention regulation (Ni et al. (2025)). Despite these advances, existing methods generally regulate modality representations without explicitly characterizing how competitive states vary within their internal structure.

Frequency-domain analysis provides a natural lens for resolving such internal heterogeneity in modality competition. FRM (Lu et al. (2026)) leverages spectral cues to estimate modality preference, but ultimately aggregates them into modality-level guidance, leaving frequency-resolved competitive states unmodeled. We find that modality preference varies substantially across spectral regions and can even reverse across frequency bands during optimization. Motivated by this observation, **SpecGuide** explicitly characterizes and regulates modality competition at the frequency level.

6.2 THEORETICAL PROOFS

6.2.1 DERIVATION OF PAIRWISE FISHER DISCRIMINABILITY

For modality m and frequency region s, consider two classes c and c'. For simplicity, define $\delta^s_{m,c,c'} = \mu^s_{m,c} - \mu^s_{m,c'}$, and $A^s_m = \Sigma^s_m + \epsilon I$, where $\epsilon > 0$. Since Σ^s_m is positive semidefinite, A^s_m is positive definite and therefore invertible. The Fisher criterion in Eq. 1 can then be written as:

$$\mathcal{D}^s_{m,c,c'} = \max_{w \neq 0} \frac{(w^\top \delta^s_{m,c,c'})^2}{w^\top A^s_m w}. \tag{18}$$

Let $u = (A^s_m)^{1/2} w$ and $w = (A^s_m)^{-1/2} u$. Substituting this change of variables into Eq. 18 gives:

$$\mathcal{D}^s_{m,c,c'} = \max_{u \neq 0} \frac{\left[u^\top (A^s_m)^{-1/2} \delta^s_{m,c,c'}\right]^2}{u^\top u}. \tag{19}$$

By the Cauchy–Schwarz inequality,

$$\left[u^\top (A^s_m)^{-1/2} \delta^s_{m,c,c'}\right]^2 \leq (u^\top u) \left[(\delta^s_{m,c,c'})^\top (A^s_m)^{-1} \delta^s_{m,c,c'}\right]. \tag{20}$$

Equality is attained when $u \propto (A^s_m)^{-1/2} \delta^s_{m,c,c'}$, or equivalently, $w \propto (A^s_m)^{-1} \delta^s_{m,c,c'}$. Therefore,

$$\mathcal{D}^s_{m,c,c'} = (\delta^s_{m,c,c'})^\top (A^s_m)^{-1} \delta^s_{m,c,c'}. \tag{21}$$

Substituting back the definitions of $\delta^s_{m,c,c'}$ and A^s_m yields:

$$\mathcal{D}^s_{m,c,c'} = (\mu^s_{m,c} - \mu^s_{m,c'})^\top (\Sigma^s_m + \epsilon I)^{-1} (\mu^s_{m,c} - \mu^s_{m,c'}), \tag{22}$$

which gives the closed-form expression in Eq. 1.

6.2.2 Decomposition of Discriminability across Frequency Regions

Let the complete spectral descriptor be the concatenation of the descriptors from all S frequency regions:

$$h_{i,m} = \left[(h^1_{i,m})^\top, \ldots, (h^S_{i,m})^\top \right]^\top. \tag{23}$$

Accordingly, the class mean can be written as $\mu_{m,c} = \left[(\mu^1_{m,c})^\top, \ldots, (\mu^S_{m,c})^\top \right]^\top$.

For a class pair (c, c'), define:

$$\delta_{m,c,c'} = \mu_{m,c} - \mu_{m,c'} = \left[(\delta^1_{m,c,c'})^\top, \ldots, (\delta^S_{m,c,c'})^\top \right]^\top. \tag{24}$$

The pooled within-class covariance of the complete descriptor has the block form:

$$\Sigma_m = \begin{bmatrix} \Sigma^1_m & \Sigma^{1,2}_m & \cdots & \Sigma^{1,S}_m \\ \Sigma^{2,1}_m & \Sigma^2_m & \cdots & \Sigma^{2,S}_m \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma^{S,1}_m & \Sigma^{S,2}_m & \cdots & \Sigma^S_m \end{bmatrix}, \tag{25}$$

where $\Sigma^{s,t}_m$ denotes the within-class cross-covariance between frequency regions s and t.

When these cross-region correlations are relatively weak, the off-diagonal blocks are small compared with the within-region covariance blocks, and we use the approximation:

$$\Sigma_m \approx \text{diag}\left(\Sigma^1_m, \ldots, \Sigma^S_m\right). \tag{26}$$

Consequently,

$$\Sigma_m + \epsilon I \approx \text{diag}\left(\Sigma^1_m + \epsilon I, \ldots, \Sigma^S_m + \epsilon I\right), \tag{27}$$

whose inverse is:

$$(\Sigma_m + \epsilon I)^{-1} \approx \text{diag}\left((\Sigma^1_m + \epsilon I)^{-1}, \ldots, (\Sigma^S_m + \epsilon I)^{-1}\right). \tag{28}$$

The pairwise discriminability of the complete descriptor therefore satisfies:

$$\mathcal{D}_{m,c,c'} = \delta^\top_{m,c,c'} (\Sigma_m + \epsilon I)^{-1} \delta_{m,c,c'} \approx \sum\nolimits_{s=1}^{S} (\delta^s_{m,c,c'})^\top (\Sigma^s_m + \epsilon I)^{-1} \delta^s_{m,c,c'} = \sum\nolimits_{s=1}^{S} \mathcal{D}^s_{m,c,c'}. \tag{29}$$

Averaging Eq. 29 over all unordered class pairs gives:

$$\mathcal{D}_m = \frac{2}{C(C-1)} \sum\nolimits_{c<c'} \mathcal{D}_{m,c,c'} \approx \frac{2}{C(C-1)} \sum\nolimits_{c<c'} \sum\nolimits_{s=1}^{S} \mathcal{D}^s_{m,c,c'} = \sum\nolimits_{s=1}^{S} \mathcal{D}^s_m. \tag{30}$$

6.2.3 Residual Error of Uniform Modality-level Regulation

Let r^s_m denote the desired regulation intensity for modality m in region s, while a modality-level method applies one shared intensity r_m to all regions. We measure their weighted squared mismatch as:

$$\mathcal{E}(r_m) = \sum\nolimits_{s=1}^{S} \eta_s (r_m - r^s_m)^2, \tag{31}$$

where $\eta_s \geq 0$ and $\sum_{s=1}^{S} \eta_s = 1$.

Taking the derivative with respect to r_m gives:

$$\frac{\partial \mathcal{E}(r_m)}{\partial r_m} = 2 \sum\nolimits_{s=1}^{S} \eta_s (r_m - r^s_m) = 2 \left(r_m - \sum\nolimits_{s=1}^{S} \eta_s r^s_m \right). \tag{32}$$

Therefore, the unique minimizer is:

$$r_m^* = \sum\nolimits_{s=1}^{S} \eta_s r_m^s. \tag{33}$$

Since:

$$\frac{\partial^2 \mathcal{E}(r_m)}{\partial r_m^2} = 2\sum\nolimits_{s=1}^{S} \eta_s = 2 > 0, \tag{34}$$

Eq. 33 is the global minimum.

Substituting r_m^* into the mismatch gives:

$$\mathcal{E}(r_m^*) = \sum\nolimits_{s=1}^{S} \eta_s \left(r_m^s - \sum\nolimits_{t=1}^{S} \eta_t r_m^t \right)^2 = \mathrm{Var}_\eta(r_m^s), \tag{35}$$

where

$$\mathrm{Var}_\eta(r_m^s) \triangleq \sum\nolimits_{s=1}^{S} \eta_s (r_m^s - \bar{r}_m)^2, \qquad \bar{r}_m = \sum\nolimits_{s=1}^{S} \eta_s r_m^s. \tag{36}$$

Hence, $\mathcal{E}(r_m^*) \geq 0$, with equality if and only if all frequency regions with non-zero weight η_s require the same regulation intensity. Whenever the desired intensities differ across positively weighted regions, $\mathrm{Var}_\eta(r_m^s) > 0$, and no single modality-level regulation intensity can exactly match all frequency-wise targets.

6.2.4 Conjugate Symmetry of the Modulated Spectrum

We show that PGSS preserves the conjugate symmetry of the centered latent spectrum, thereby ensuring a real-valued inverse Fourier reconstruction. For a real-valued representation $z_i^m \in \mathbb{R}^{D\times H\times W}$, its centered spectrum $\hat{z}_i^m = \mathrm{fftshift}(\mathrm{FFT}(z_i^m))$ satisfies:

$$\hat{z}_i^m(-\omega) = \hat{z}_i^m(\omega)^*, \tag{37}$$

where $-\omega$ denotes the conjugate frequency of ω. Consequently,

$$|\hat{z}_i^m(-\omega)| = |\hat{z}_i^m(\omega)|, \tag{38}$$

and the log-compressed magnitude used by PGSS satisfies:

$$X_i^m(-\omega) = X_i^m(\omega). \tag{39}$$

The frequency-wise preference map inherits the same symmetry. Since FMP partitions the spectrum into radial regions $\{\Omega_s\}_{s=1}^S$, each region satisfies $\omega \in \Omega_s \Leftrightarrow -\omega \in \Omega_s$. Therefore,

$$A_i^m(-\omega) = A_i^m(\omega), \tag{40}$$

where $A_i^m(\omega) = \sum_{s=1}^S \alpha_{i,m}^s \mathbf{1}_{\Omega_s}(\omega)$.

Recall that PGSS computes:

$$U_i^m = \psi_m(X_i^m), \qquad W_i^m = \mathrm{Softmax}_J\left(\varphi_m([X_i^m; A_i^m])\right), \tag{41}$$

followed by:

$$G_i^m = \sigma(\phi_m(U_i^m \odot W_i^m)). \tag{42}$$

The mappings ψ_m, φ_m, and ϕ_m consist of spatially shared 1×1 convolutions, normalization, and point-wise nonlinearities (Detailed model structure can be found in Appendix 6.4). Hence, identical inputs at conjugate frequency locations produce identical outputs. From Eqs. 39 and 40, it follows that:

$$U_i^m(-\omega) = U_i^m(\omega), \qquad W_i^m(-\omega) = W_i^m(\omega), \tag{43}$$

and therefore:

$$G_i^m(-\omega) = G_i^m(\omega). \tag{44}$$

Moreover, G_i^m is real-valued. PGSS modulates the spectral magnitude while retaining the original phase, which can equivalently be written as:

$$\tilde{\hat{z}}_i^m(\omega) = G_i^m(\omega)\hat{z}_i^m(\omega). \tag{45}$$

Combining Eqs. 37 and 44, we obtain:

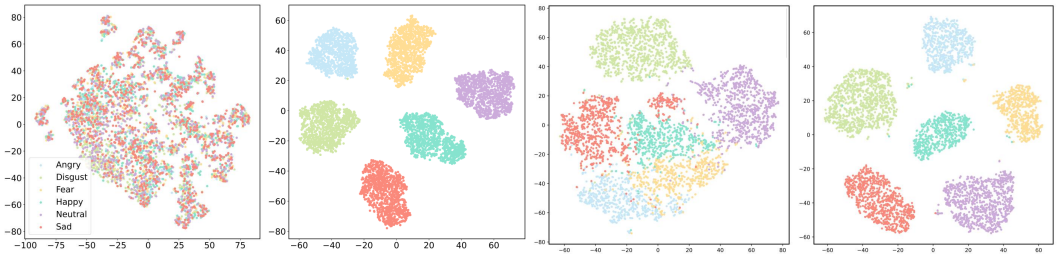
$$\tilde{\hat{z}}_i^m(-\omega) = G_i^m(-\omega)\hat{z}_i^m(-\omega) = G_i^m(\omega)\hat{z}_i^m(\omega)^* = (G_i^m(\omega)\hat{z}_i^m(\omega))^* = \tilde{\hat{z}}_i^m(\omega)^*. \tag{46}$$

Thus, the modulated spectrum remains conjugate symmetric, and its inverse Fourier transform is real-valued up to numerical precision.

6.3 Supplementary Experiments

6.3.1 Visualization of Modality-specific Representations

We further visualize the modality-specific representations learned by Joint-Training and SpecGuide using t-SNE. As shown in Fig. 4, Joint-Training produces well-separated clusters for the audio modality, whereas the visual representations exhibit severe class overlap, revealing a pronounced discrepancy in the discriminative structures learned by the two modality encoders. In contrast, SpecGuide substantially improves the class separability of the visual representations while preserving clear category structures in the audio modality. This qualitative comparison suggests that SpecGuide alleviates the under-optimization of the weaker modality without sacrificing the discriminative representation of the stronger one, leading to more effective utilization of both modalities for multimodal learning.



(a) Joint-Training Visual **(b) Joint-Training Audio** **(c) SpecGuide Visual** **(d) SpecGuide Audio**

Figure 4: t-SNE visualization of modality-specific representations on CREMA-D. We visualize the encoder representations learned by Joint-Training and SpecGuide for the visual and audio modalities. Joint-Training exhibits severe class overlap in the visual representation space despite clear separation in the audio modality, whereas SpecGuide substantially improves visual class separability while maintaining discriminative audio representations.

6.3.2 Parameter Sensitivity Analysis

We analyze the sensitivity of SpecGuide to four key hyperparameters on CREMA-D: the number of learned spectral response channels J in PGSS, the prototype warm-up count in FMP, the weight λ_{FCG} of the FCG and the temperature τ used for frequency-wise preference normalization. For each hyperparameter, we vary its value while keeping all other settings fixed. The results are shown in Fig. 5.

As shown in Fig. 5(a), performance first improves as J increases and reaches its best result at $J = 4$, while further increasing the number of response channels brings no additional benefit. This suggests that a small number of learned spectral response channels is sufficient to capture useful modulation patterns, whereas excessive decomposition may introduce unnecessary redundancy and increase optimization difficulty. Fig. 5(b) shows that SpecGuide is relatively insensitive to the prototype warm-up count over a broad range, with the best performance obtained around 16 samples. A very small warm-up count may activate prototype-based preference estimation before sufficiently reliable class statistics are accumulated, whereas an excessively long warm-up delays the use of frequency-wise preference guidance. For λ_{FCG}, Fig. 5(c) exhibits a clear optimum around $\lambda_{\text{FCG}} = 1.0$. A small value provides insufficient complementarity regularization, while an overly large value may excessively constrain cross-modal spectral descriptors and interfere with task-driven optimization. Finally, Fig. 5(d) evaluates the temperature τ in FMP. The best performance is achieved at a moderate temperature, while SpecGuide remains competitive across the tested range. Since τ controls the sharpness of the cross-modal preference distribution, an appropriate temperature provides sufficiently discriminative frequency-wise guidance without making the preference distribution overly sharp or excessively smooth.

Overall, SpecGuide maintains stable performance across a reasonably broad range of hyperparameter settings, indicating that its effectiveness does not rely on highly sensitive parameter tuning.

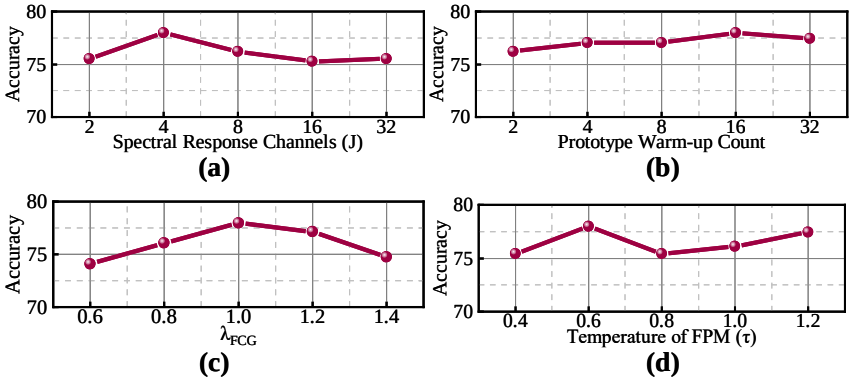


Figure 5: Parameter sensitivity analysis on CREMA-D. Sensitivity to (a) the number of spectral response channels J in PGSS, (b) the prototype warm-up count in FMP, (c) the FCG loss weight λ_{FCG}, and (d) the preference temperature τ in FMP.

6.3.3 Visualization of Preference-Guided Spectral Responses

To further examine how PGSS performs preference-guided spectral shaping, we visualize its intermediate spectral responses for a representative audio-visual sample in Fig. 6. The first row shows the $J = 4$ learned spectral response channels U_i^m, the second row shows the corresponding preference-conditioned weight maps W_i^m, and the last row presents the weighted responses $U_i^m \odot W_i^m$ that are subsequently fused to generate the final spectral modulation map.

As shown in Fig. 6, the learned response channels capture clearly different spectral patterns within each modality. More importantly, the corresponding weight maps exhibit non-uniform and location-dependent responses, indicating that PGSS assigns different importance to the learned spectral responses according to the current spectral content and frequency-wise preference guidance. Consequently, the weighted responses selectively retain different spectral structures across response channels. The audio and visual modalities also exhibit distinct weighting patterns, showing that PGSS performs modality-specific spectral shaping instead of applying a shared modulation pattern. These visualizations provide qualitative evidence that the frequency-wise preference estimated by FMP effectively conditions the spectral response of PGSS.

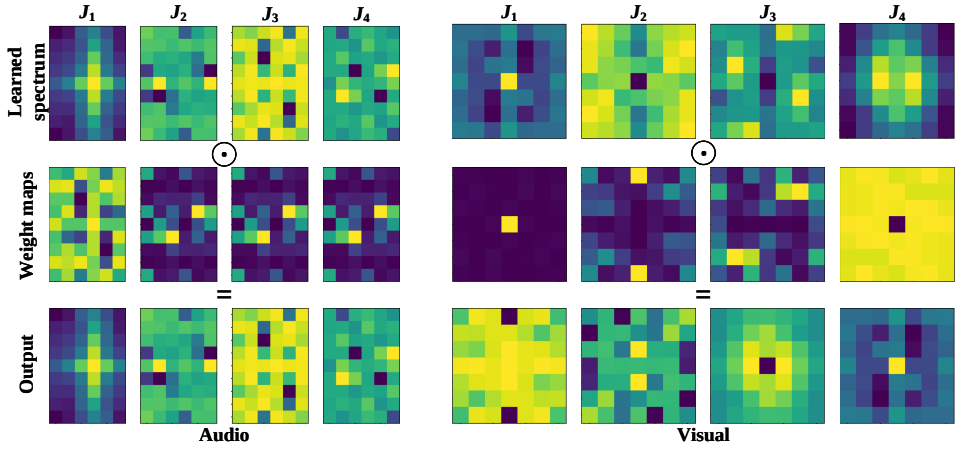


Figure 6: Visualization of preference-guided spectral responses in PGSS. For each modality, the first row shows the four learned spectral response channels U_i^m, the second row shows the corresponding preference-conditioned weight maps W_i^m, and the third row shows the weighted responses $U_i^m \odot W_i^m$.

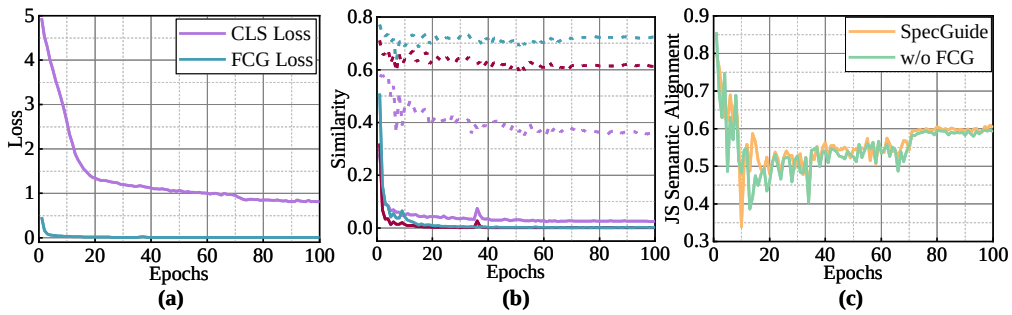


Figure 7: Effect of FCG on spectral complementarity and cross-modal alignment. (a) Classification and FCG losses during training. (b) Frequency-wise cross-modal spectral similarity with and without FCG. Purple, red, and blue denote low-, middle-, and high-frequency regions, respectively; solid and dashed lines correspond to SpecGuide and w/o FCG. (c) Jensen–Shannon semantic alignment between the two unimodal predictive distributions for SpecGuide and w/o FCG.

6.3.4 Effect of FCG on Cross-Modal Alignment.

Introducing FCG naturally raises a concern: whether the orthogonality regularization that reduces band-wise cosine similarity may inadvertently weaken useful cross-modal alignment. We first note that the ablation results in Table 5 show a further improvement in classification accuracy after introducing FCG, indicating that the imposed complementarity is beneficial to the downstream task. We then examine the optimization dynamics in greater detail. As shown in Fig. 7(a), both the classification loss and the FCG loss decrease smoothly during training, showing that the two objectives can be jointly optimized and that the behavior observed in Fig. 7(c) is not simply caused by an ineffective FCG constraint.

Fig. 7(b) further shows that substantial cross-modal spectral similarity persists without FCG, particularly in the middle- and high-frequency regions, whereas introducing FCG markedly reduces the similarity across all three frequency regions. Most importantly, this reduction in spectral similarity does not lead to degraded semantic alignment. As shown in Fig. 7(c), the Jensen–Shannon semantic alignment between the two unimodal predictive distributions remains comparable to that of the variant without FCG and converges to a similar level in the later stage of training. This comparison indicates that reducing frequency-wise spectral correlation does not prevent the two modalities from maintaining consistent task-semantic predictions. Together, these results show that FCG promotes frequency-wise complementarity while preserving task-relevant cross-modal alignment.

6.4 Implementation Details of SpecGuide

6.4.1 Frequency-wise Modality Preference Estimation

For FMP, we partition the centered spectrum into three predefined radial regions corresponding to low-, middle-, and high-frequency components. Taking a 2D feature map as an example, for a spectrum of spatial size $H \times W$, the normalized radial coordinate is defined as:

$$\rho(u, v) = \frac{\sqrt{f_u^2 + f_v^2}}{\max_{u,v} \sqrt{f_u^2 + f_v^2}}, \tag{47}$$

where f_u and f_v denote the centered discrete frequency coordinates. The three regions are:

$$\Omega_{\mathrm{L}} = \{\omega : \rho(\omega) < 0.25\}, \quad \Omega_{\mathrm{M}} = \{\omega : 0.25 \le \rho(\omega) < 0.55\}, \quad \Omega_{\mathrm{H}} = \{\omega : \rho(\omega) \ge 0.55\}. \tag{48}$$

For each modality, FMP computes the power spectrum of the representation. The spectral energy of channel d within region s is:

$$e_{i,m}^{s,d} = \frac{1}{|\Omega_s|} \sum\nolimits_{\omega \in \Omega_s} \left| \hat{z}_{i,m}^{d}(\omega) \right|^2. \tag{49}$$

After logarithmic compression, the energy is normalized across frequency regions within each channel:

$$q_{i,m}^{s,d} = \frac{\log(1 + e_{i,m}^{s,d})}{\sum_{t=1}^{S} \log(1 + e_{i,m}^{t,d}) + \epsilon}. \tag{50}$$

The resulting $q_{i,m}^{s} \in \mathbb{R}^D$ is used as the spectral descriptor for preference estimation.

For each modality m, class c, and frequency region s, FMP maintains a class-conditioned prototype $p_{m,c}^{s} \in \mathbb{R}^D$. Let $\bar{q}_{m,c}^{s}$ denote the mean descriptor of class c in the current mini-batch. The prototype is updated by:

$$p_{m,c}^{s} \leftarrow \beta p_{m,c}^{s} + (1 - \beta)\bar{q}_{m,c}^{s}, \tag{51}$$

where $\beta = 0.9$ in our experiments. The prototype distance is computed using the mean ℓ_1 distance:

$$d_{i,m,c}^{s} = \frac{1}{D} \left\| q_{i,m}^{s} - p_{m,c}^{s} \right\|_1. \tag{52}$$

Accordingly, $d_{i,m,+}^{s}$ denotes the distance to the ground-truth class prototype, and $d_{i,m,-}^{s}$ denotes the minimum distance to the remaining class prototypes. The scale term $\gamma_{i,m}^{s}$ in Eq. 8 is computed as the mean distance from $q_{i,m}^{s}$ to all valid class-conditioned prototypes within the same spectral region:

$$\gamma_{i,m}^{s} = \frac{1}{|\mathcal{C}_m^{\text{val}}|} \sum_{c \in \mathcal{C}_m^{\text{val}}} d_{i,m,c}^{s}, \tag{53}$$

where $\mathcal{C}_m^{\text{val}}$ denotes the set of classes whose prototypes have accumulated sufficient samples.

The complete pseudocode for FMP is shown in Algorithm 2.

Algorithm 2 Frequency-wise Modality Preference (FMP) Estimation

Require: Mini-batch representations $\{z_i^m\}_{i=1,m=1}^{B,M}$, labels $\{y_i\}_{i=1}^{B}$, spectral regions $\{\Omega_s\}_{s=1}^{S}$, class-conditioned spectral prototypes $\{p_{m,c}^{s}\}$.

for each modality $m = 1, \ldots, M$ **do**
  Compute the centered spectra $\hat{z}_i^m = \text{fftshift}(\text{FFT}(z_i^m))$.
  Construct spectral descriptors $\{q_{i,m}^{s}\}_{i=1,s=1}^{B,S}$ from the normalized spectral-energy distributions.
  Determine the valid class set $\mathcal{C}_m^{\text{val}}$ according to the prototype sample counts.
  for each spectral region $s = 1, \ldots, S$ **do**
    Compute the prototype distances $\{d_{i,m,c}^{s}\}_{c \in \mathcal{C}_m^{\text{val}}}$ using Eq. 52.
    Obtain the ground-truth distance $d_{i,m,+}^{s}$ and the nearest competing-class distance $d_{i,m,-}^{s}$.
    Compute the prototype-distance scale $\gamma_{i,m}^{s}$ using Eq. 53.
    Compute the normalized discriminative margin $r_{i,m}^{s}$ using Eq. 8.
  end for
end for
for each spectral region $s = 1, \ldots, S$ **do**
  Compute the cross-modal preference weights $\{\alpha_{i,m}^{s}\}_{m=1}^{M}$ using Eq. 9.
end for
for each modality $m = 1, \ldots, M$ **do**
  Construct the frequency-wise preference maps $\{A_i^m\}_{i=1}^{B}$ using Eq. 10.
  Update the class-conditioned spectral prototypes $\{p_{m,c}^{s}\}$ using the current mini-batch.
end for
return Frequency-wise preference maps $\{A_i^m\}_{i=1,m=1}^{B,M}$.

6.4.2 Preference-Guided Spectral Shaping

PGSS operates on the centered Fourier spectrum $\hat{z}_i^m = \text{fftshift}(\text{FFT}(z_i^m))$ and uses $X_i^m = \log(1 + |\hat{z}_i^m|)$ as input. The projection ψ_m is implemented using a 1×1 convolution:

$$U_i^m = \text{Conv}_{1\times1}^{D\to J}(X_i^m), \tag{54}$$

where J denotes the number of learned spectral response channels.

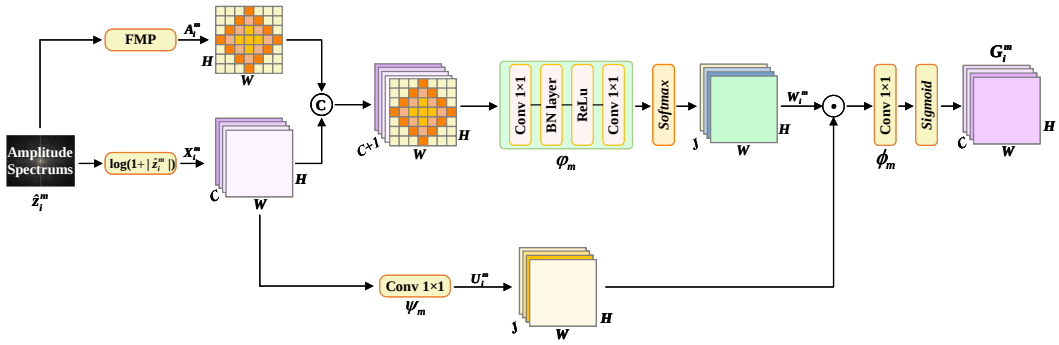


Figure 8: Detailed architecture of PGSS. The frequency-wise preference map A_i^m conditions the response weights W_i^m, which modulate the learned spectral responses U_i^m to generate the spectral modulation map G_i^m.

The preference-conditioning network φ_m takes $[X_i^m; A_i^m]$ as input and consists of two 1×1 convolutional layers with normalization and nonlinear activation, as illustrated in Fig. 8. The response weights are obtained as:

$$W_i^m = \text{Softmax}_J\left(\varphi_m([X_i^m; A_i^m])\right). \tag{55}$$

The weighted responses are projected back to the original channel dimension and passed through a sigmoid:

$$G_i^m = \sigma\left(\text{Conv}_{1\times1}^{J\to D}(U_i^m \odot W_i^m)\right). \tag{56}$$

PGSS then modulates the spectral magnitude as:

$$|\tilde{\hat{z}}_i^m| = |\hat{z}_i^m| \odot G_i^m, \tag{57}$$

while retaining the original phase. The modulated representation is reconstructed according to Eq. 13. During training, PGSS uses the sample-wise preference map estimated by FMP, while inference uses the training-derived frequency-wise preference prior. Detailed configurations for each network in PGSS are shown in Table 7.

Table 7: Detailed network configuration of PGSS.

Network	Structure	Configuration
ψ_m	Conv	$D \to J$, kernel $= 1 \times 1$, stride $= 1$, padding $= 0$
φ_m	Conv	$(D+1) \to D/4$, kernel $= 1 \times 1$, stride $= 1$, padding $= 0$
	BN	$D/4$
	ReLU	–
	Conv	$D/4 \to J$, kernel $= 1 \times 1$, stride $= 1$, padding $= 0$
ϕ_m	Conv	$J \to D$, kernel $= 1 \times 1$, stride $= 1$, padding $= 0$

6.5 Details of Imbalanced Ratio Estimators

We detail the three representative families of estimators used in our experiments: a contribution-based estimator, a mutual-information-based estimator, and a cluster-purity-based estimator. For clarity, we describe the two modalities case with an audio modality a and a visual modality v. The extension to more modalities is straightforward. We denote by $f^m(\cdot)$ the modality specific logits, by y_i the ground-truth label of sample i, and by B the batch size.

Contribution-based Estimator. Following the discriminative-contribution view of (Wei et al. (2025c)), the contribution of a modality is measured by how much confidence it assigns to the correct class. For each modality $m \in \{a, v\}$, we accumulate the softmax probability at the ground-truth class over the batch:

$$s^m = \sum_{i=1}^{B} \text{Softmax}\big(f^m(x_i)\big)_{y_i}. \tag{58}$$

The imbalanced ratio is then the normalized contribution of each modality:

$$\rho^v \;=\; \frac{s^v}{s^a + s^v}, \qquad \rho^a \;=\; 1 - \rho^v, \tag{59}$$

where a larger s^m yields a larger ρ^m. This estimator is cheap to compute and depends only on the current predictions.

MI-based Estimator. The second estimator characterizes a modality by the amount of label-relevant information it carries (Yang et al. (2025a)). Intuitively we want to approximate the normalized MI $I(X^m; Y)/H(X^m)$, i.e. how much of the modality's own uncertainty is explained by the label.

Let $\hat{p}(y)$ be the empirical label distribution in the batch and let $H(Y) = -\sum_y \hat{p}(y) \log \hat{p}(y)$ be the label entropy. We estimate the label-relevant term of modality m by the negative cross-entropy between its logits and the ground-truth labels,

$$s^m \;=\; -\,\mathrm{CE}\big(f^m,\, y\big), \tag{60}$$

which acts as a proxy for $I(X^m; Y)$: the better modality m predicts the label, the larger s^m. The intrinsic uncertainty of the modality is estimated by:

$$H_m \;=\; \mathrm{CE}\big(f^m,\, \mathrm{Softmax}(f^m)\big), \tag{61}$$

i.e. the entropy of its own predictive distribution, serving as a proxy for $H(X^m)$.

The unnormalized informativeness ratio of each modality is:

$$r^m \;=\; \frac{s^m + H(Y) + \epsilon_0}{H_m}, \tag{62}$$

where the constant ϵ_0 shifts the numerator to a stable positive regime. To turn the two informativeness scores into a competitive ratio we apply a temperature-scaled softmax with temperature T:

$$\rho^a \;=\; \frac{\exp(r^v/T)}{\exp(r^a/T) + \exp(r^v/T)}, \qquad \rho^v \;=\; 1 - \rho^a. \tag{63}$$

Note the cross assignment in Eq. 63: the coefficient of modality a is driven by the informativeness of the other modality v, so that a more informative v enlarges ρ^a and encourages the weaker modality to be strengthened.

Cluster-purity-based Estimator. The third estimator inspects the geometry of the learned representations rather than the predictions, following (Wei et al. (2024b)). The key idea is that a modality whose feature space is well organized with respect to the labels (high purity) and generalizes consistently between training and validation (small train–validation purity gap) is comparatively well optimized.

For each modality m and each split $s \in \{\text{train}, \text{val}\}$, we first rescale the features to $[-1, 1]$ and run K-means clustering (K equal to the number of classes). Given the resulting clusters $\{c_k\}$ and the ground-truth partition $\{t_j\}$, cluster purity is:

$$\mathrm{Pur}^m_s \;=\; \frac{1}{N} \sum\nolimits_k \max_j \big|c_k \cap t_j\big|, \tag{64}$$

where N is the number of samples in the split. We then take the absolute train–validation purity gap of each modality as its imbalance signal,

$$g^m \;=\; \big|\,\mathrm{Pur}^m_{\text{train}} - \mathrm{Pur}^m_{\text{val}}\,\big|, \tag{65}$$

and normalize across modalities to obtain the ratio:

$$\rho^a \;=\; \frac{g^a}{g^a + g^v + \epsilon}, \qquad \rho^v \;=\; 1 - \rho^a, \tag{66}$$

with a small ϵ for numerical stability. A larger gap g^m indicates a modality whose representation transfers less reliably, and thus receives a larger ρ^m.

6.6 Dataset Details and Preprocessing

We conduct experiments on six benchmark datasets covering diverse modality combinations and learning tasks. Five datasets involve two modalities, including audio-visual recognition (CREMA-D (Cao et al. (2014)), Kinetics-Sounds (Arandjelovic & Zisserman (2017)), and AVE (Tian et al. (2018))), RGB-optical-flow action recognition (UCF-101 (Soomro et al. (2012))), and vision-language recognition (UPMC Food-101 (Wang et al. (2015))). We further evaluate SpecGuide on CMU-MOSI (Zadeh et al. (2018)), which contains three modalities (audio, visual, and text), to examine its applicability beyond the two-modality setting.

- **CREMA-D** (Cao et al. (2014)). CREMA-D is a benchmark for multimodal emotion recognition containing 7,442 video clips of six emotion categories: angry, happy, sad, neutral, disgust, and fear. Following the standard split, 6,698 samples are used for training and validation and 744 for testing. For the visual modality, frames are sampled at 1 FPS, and one frame is randomly selected from all sampled frames except the first, followed by random 224×224 cropping and horizontal flipping. For the audio modality, following (Peng et al. (2022); Wei et al. (2025c)), raw waveforms are converted into 257×299 spectrograms using a window length of 512 and an overlap of 353.
- **Kinetics-Sounds** (Arandjelovic & Zisserman (2017)). Kinetics-Sounds is an audio-visual action recognition dataset containing 34 human action categories, where each sample is a 10-second video clip centered on the corresponding action. The dataset contains approximately 15k training samples and 1.9k samples each for validation and testing. For the visual modality, three frames are randomly sampled from each video and augmented with random 224×224 cropping and horizontal flipping. The audio modality is processed following the official preprocessing protocol.
- **AVE** (Tian et al. (2018)). AVE is a benchmark for audio-visual event recognition containing 4,143 10-second video clips from 28 event categories. The dataset contains approximately 3,339 training samples and 402 samples each for validation and testing. For the visual modality, one frame is randomly sampled from each video and augmented with random 224×224 cropping and horizontal flipping. The audio modality is processed following the official preprocessing protocol.
- **UCF-101** (Soomro et al. (2012)). UCF-101 is an action recognition dataset containing 13,320 video clips from 101 action categories, with 9,586 samples for training and validating and 3,734 for testing. We use MMAction2 to sample videos at 25 FPS and extract RGB frames together with the corresponding optical-flow fields in the x and y directions. The model input consists of three uniformly sampled RGB frames and two corresponding optical-flow maps, with the latter concatenated along the channel dimension. Both RGB and optical-flow inputs are augmented using random 224×224 cropping and horizontal flipping.
- **UPMC Food-101** (Wang et al. (2015)). UPMC Food-101 is an vision-language food recognition dataset containing 90k image-text pairs from 101 food categories. We use 67k samples for training and validation, and 22k for testing. For the visual modality, images are resized to 224×224 and augmented using random rotations and horizontal flipping.
- **CMU-MOSI** (Zadeh et al. (2018)). CMU-MOSI is a multimodal sentiment analysis benchmark containing 2,199 video segments with audio, visual, and textual modalities. We follow the standard preprocessing and data split adopted in previous studies, without applying additional data augmentation.

6.7 Supplementary Implementation Details

For fair comparison, all methods are trained for 100 epochs using SGD with a momentum of 0.9 and a batch size of 64. The initial learning rate is set to 1×10^{-3} for Food-101 and CMU-MOSI, which contain a textual modality, and to 2×10^{-3} for the remaining datasets. Following previous works, we adopt a StepLR scheduler with a step size of 70 epochs and a decay factor of 0.1. Unless otherwise specified, the remaining training settings follow previous works (Peng et al. (2022); Wei et al. (2025b;a)), and consistent configurations are used across all compared methods.

For SpecGuide, each class-conditioned spectral prototype in FMP is used for preference estimation after accumulating at least 16 samples. In PGSS, the number of learned spectral response channels in Eq. 11 is fixed to $J = 4$ across all datasets. For the overall objective in Eq. 16, we set $\lambda_{\text{CLS}} = 2$. The temperature τ in Eq. 9 and the FCG weight λ_{FCG} in Eq. 16 are set to 1 by default. We use $\tau = 0.6$ for CREMA-D, while λ_{FCG} is set to 0.5 for AVE and CMU-MOSI and to 0.9 for Kinetics-Sounds and UCF-101.

6.8 Limitations and Discussion

Although SpecGuide demonstrates consistent effectiveness across diverse multimodal settings, several limitations remain. First, FMP currently partitions the spectrum into predefined low-, middle-, and high-frequency regions using fixed boundaries. While this design performs consistently across the evaluated datasets and architectures, adaptive spectral partitioning may further improve its flexibility across different modalities and tasks. Second, FMP uses class-conditioned spectral prototypes and ground-truth labels to estimate sample-wise frequency preferences during training, while PGSS employs the learned frequency-wise preference statistics as a prior during inference. Extending FMP to label-free, sample-adaptive preference estimation at inference time could further enhance the flexibility of SpecGuide. We will focus on these topics in our following works.