From Token Imbalance to Balanced Routing: An ELBO-Regularized Probabilistic Framework for Contrastive Multimodal Learning

Habibeh Naderi
Dalhousie University
habibeh.naderi@dal.ca

Behrouz Haji Soleimani
Kinaxis
bhajisoleimani@kinaxis.com

Stan Matwin
Dalhousie University
stan@cs.dal.ca

Abstract

We introduce CoPRIME (Contrastive Probabilistic Routing for IMbalanced tokens with ELBO-regularized mixture of experts), a probabilistic routing framework for multimodal representation learning that generalizes multimodal representation learning beyond vision-text by tackling the fundamental challenge of extreme token imbalance across modalities. This imbalanced-ness is particularly pronounced between spectrogram-tokenized audio and text. CoPRIME augments contrastive pretraining with an ELBO-regularized routing objective that jointly promotes 1) expert specialization, requiring experts to explain the tokens they receive, and 2) diverse utilization via KL regularization to a uniform prior. To stabilize routing, we further replace standard CoV-based regularizers with entropy-based importance and load losses, yielding smoother gradients and flexible, modality-aware routing without rigid uniformity constraints. On MOSEI and IEMOCAP datasets, CoPRIME achieves state-of-the-art zero- and few-shot emotion and sentiment results, outperforming dense Transformers and prior multimodal MoE variants while retaining the efficiency of sparse conditional computation. Ablations isolate the role of each loss and show that ELBO is the primary driver of stable specialization under modality imbalance, with entropy-based regularizers further improving convergence and utilization.

1 INTRODUCTION

The success of multimodal contrastive learning, as exemplified by CLIP (Radford et al., 2021) and SimCLR (Chen et al., 2020a,b), has demonstrated the ability to learn high-quality representations that generalize well to a wide range of downstream tasks. Pre-training on large-scale datasets enables such models to achieve zero-shot transferability and robustness to distribution shifts (Chen et al., 2023; Xue et al., 2023; Ming and Li, 2024; Nakada et al., 2023; Ren and Li, 2023). In parallel, conditional computation (Bengio, 2013) has emerged as a strategy to increase model capacity while maintaining a roughly constant training and inference cost by selectively activating subsets of model parameters for different inputs. In NLP, sparsely activated Mixture of Experts (MoEs) have gained traction (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022), enabling efficient training and inference while scaling up to trillion-parameter models. MoEs have also been effectively applied to vision (Riquelme et al., 2021; Lou et al., 2021), text (Lepikhin et al., 2020; Zoph et al., 2022), and most recently, multimodal models (Mustafa et al., 2022). MoEs are a natural fit for a multimodal backbone, since expert layers can learn an appropriate partitioning of modalities. LIMoE (Mustafa et al., 2022), as the first proposed multimodal MoEs, identified unique failure modes that arise when integrating multiple modalities in a single MoE framework. To address those, they adopted customized entropy-based regularization techniques to stabilize the multimodal MoEs training.

Intuitively, the sparse MoEs models may better handle distinct downstream tasks with low-data settings as their primary motivation is to scale model parameters while keeping computational costs in check by activating only a fraction of model parameters in both pre-training and inference time. Additionally, MoEs' sparse nature offers benefits such as mitigating catastrophic forgetting in continual learning (Collier et al., 2020) and enhancing multitask learning (Ma et al., 2018). Sparse MoEs introduce an inductive bias that aligns well with specialized tasks, as they preserve task-specific learned knowledge. This knowledge is effectively transferred to new downstream tasks through the dynamic and selective activation of parameter subsets during inference.

Building on these advancements, we further investigate the application of sparse MoEs to multimodal setting. In particular, we propose a single multimodal MoEs architecture

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

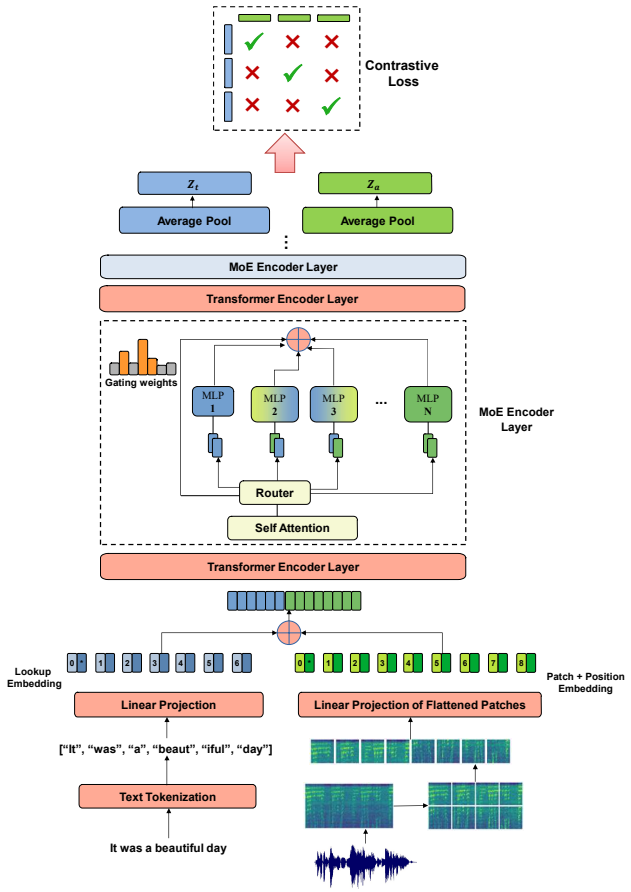

Figure 1: CoPRIME, a sparsely activated contrastive ELBO-regularized multimodal model.

that aligns audio and text representations using contrastive learning (Radford et al., 2021). We refer to our resulting model as CoPRIME (Contrastive Probabilistic Routing for IMbalanced tokens with ELBO-regularized mixture of experts) a probabilistic routing framework that generalizes multimodal representation learning beyond vision-text to the especially challenging audio-text case. The majority of existing multimodal literature is on image-text which is fueled by the availability of large public datasets. Extension to audio is possible by transforming raw audio waveforms into sequences of discrete tokens analogous to text.

In this work, we leverage a Vision Transformer (ViT)-style (Dosovitskiy et al., 2021) tokenization on the spectrogram of audio and integrating it into a multimodal contrastive learning framework. We couple ViT-style spectrogram tokenization with a shared Transformer and sparse expert layers, enabling efficient contrastive alignment without modality-specific encoders or decoders. While recent vision-text models such as CLIP (Radford et al., 2021) have demonstrated remarkable success despite token disparities, audio-text modeling suffers from a far more severe imbalance, as spectrogram tokenization produces disproportionately many audio tokens relative to text. CoPRIME specifically targets this distinct challenge not fully explored in prior multimodal MoE work: the extreme token imbalance. To address this, we design specialized loss functions including an Evidence Lower Bound (ELBO) loss combined with entropy-based auxiliary losses that are rooted in principles of probabilistic regularization and information theory. In particular, we propose an ELBO-regularized routing objective that encourages expert specialization while preventing expert collapse. The ELBO loss achieves this by maximizing the marginal likelihood of modality-specific experts, ensuring that the selected experts effectively model the assigned tokens. Additionally, it regularizes the expert marginal probability distribution toward a uniform prior by minimizing the Kullback-Leibler (KL) divergence between the prior and the expert likelihood conditioned on a modality. To further stabilize training, we reformulate standard MoE regularizers into entropy-based importance and load losses (instead of Coefficient of Variation (CoV)-based), which promote balanced yet flexible routing by maximizing entropy rather than imposing hard constraints, complemented by Z-loss and Mutual Information (MI) for additional stability and diversity. Finally, we distinguish ELBO from other auxiliary objectives through a dedicated weighting factor λ_{ELBO}, reflecting its central role in balancing specialization and diversity under extreme modality imbalance. These objectives promote diverse yet balanced expert usage and in turn prevent a subset of experts from dominating due to modality bias. In summary, our contributions are as follows:

- Unified contrastive MoE: We present CoPRIME, a unified sparse MoE framework for multimodal contrastive learning that explicitly addresses the severe token imbalance across modalities, enabling stable and effective cross-modal alignment.

- ELBO-regularized routing for modality imbalance: We introduce an ELBO objective that 1) rewards experts for explaining the tokens they receive, and 2) applies a KL prior to prevent expert collapse, directly addressing the extreme audio-text token imbalance that destabilizes multimodal MoEs.

- Entropy-based routing stabilizers: We replace CoV-style importance and load loss terms in standard MoE with entropy-based counterparts (complemented by Z-loss and MI) to encourage balanced yet flexible expert utilization, yielding smoother gradients and more stable training under sparse, top-k routing.

- Comprehensive validation: On MOSEI and IEMOCAP datasets (zero and few-shot, cross and intra-domain), CoPRIME consistently outperforms dense and LIMoE-style baselines, ablations show ELBO is the primary driver of gains, with entropy-based losses improving utilization and convergence.

2 PROPOSED METHOD

2.1 CoPRIME Model Architecture

We propose a unified Transformer-based architecture (Vaswani, 2017) for processing both audio and text modalities, leveraging a shared encoder design that facilitates multimodal learning. Figure 1 illustrates our proposed model architecture.

Audio Tokenization using Spectrogram For audio, raw waveform data is first transformed into a spectrogram representation in time-frequency domain which is well-suited for capturing temporal and spectral features. This representation is subsequently tokenized using a Vision Transformer (ViT)-style approach (Dosovitskiy et al., 2021), where the spectrogram is divided into non-overlapping patches of size $P \times P$. Each patch is flattened into a vector, and linearly projected into a fixed-dimensional token embedding, $emb_{audio} = W_{patch} \cdot \text{flatten}(S_{patch}) + b_{patch}$, where S_{patch} is a spectrogram patch, W_{patch} and b_{patch} are learnable parameters for the linear embedding layer.

Text Tokenization For text, we utilize a standard sentencepiece tokenizer (Kudo and Richardson, 2018) to encode input sequences into discrete tokens, which are then mapped to the desired embedding dimension using a learned vocabulary and a linear projection layer, $emb_{text} = W_t \cdot t_{id}$, where t_{id} is the one-hot token ID, and W_t is the text embedding matrix. The linear embedding layers for text and audio have the same dimensionality. These modality-specific embedding layers ensure that audio and text data are projected into a shared feature space with consistent dimensionality.

After the linear projection layers, the embeddings are augmented with an extra flag token to specify which modality the token belongs to.

Shared Dense Transformer Encoder Layers Once tokenized, the audio and text tokens are concatenated and passed through a shared Transformer encoder. This shared encoder design is intentionally modality-agnostic, meaning that it processes audio and text tokens without explicit conditioning on modality type. The Transformer encoder, consisting of self-attention and feedforward layers, processes all tokens jointly, allowing for implicit cross-modal interaction, knowledge transfer, and alignment during representation learning. The output token representations from the final layer of the Transformer are average-pooled along the sequence dimension to generate a single modality-specific representation vector z_m for each input, where $m \in \{\text{audio}, \text{text}\}$. To train the model, we employ a contrastive learning objective (Radford et al., 2021; Yuan et al., 2021; Yu et al., 2022) that encourages paired audio and text inputs to have similar representations in the shared latent space while ensuring that unpaired inputs are dissimilar. Specifically, the representations z_{audio} and z_{text} are linearly projected using modality-specific weight matrices W_{audio} and W_{text}, respectively, to produce the final representations, $\{(W_{\text{audio}} z_{\text{a}_k}, W_{\text{text}} z_{\text{t}_k})\}_{k=1}^{n}$. The contrastive loss $\mathcal{L}_{\text{contrastive}}$ is then computed over paired and unpaired examples. These dense Transformer encoder layers are trained without explicit modality separation, demonstrating the ability to capture shared patterns across both audio and text inputs. However, to enhance efficiency and scalability, we introduce sparse Mixture of Experts (MoE) layers (Zhou et al., 2022; Zoph et al., 2022) into the architecture, following the principles established in prior work (Riquelme et al., 2021; Mustafa et al., 2022; Lepikhin et al., 2020). These layers selectively activate subsets of parameters based on input characteristics, significantly reducing computational overhead while maintaining representational capacity.

Sparse Mixture of Experts (MoE) Encoder Layers Sparse MoE encoder layers consist of multiple experts, which are Multi-Layer Perceptrons (MLPs), and are activated selectively based on input characteristics. In this framework, each token $x \in \mathbb{R}^D$ is processed by only K out of E available experts, allowing for sparse computation. To determine the K experts to activate, a lightweight router computes gating weights for each token, $g(x) = \text{softmax}(W_g x) \in \mathbb{R}^E$, where $W_g \in \mathbb{R}^{E \times D}$ is a learnable parameter matrix. Then, the top K experts with the highest gating weight are selected. The final output of the MoE encoder layer is obtained by linearly combining the outputs of the K selected experts, weighted by their corresponding gating values, $\text{MoE}(x) = \sum_{e=1}^{K} g_e(x) \cdot \text{MLP}_e(x)$. This mechanism enables the model to efficiently allocate computational resources by dynamically selecting the most relevant experts for each token, thereby enhancing scalability and performance.

2.2 Contrastive Loss

Given n pairs of audio and text transcriptions $\{(a_i, t_i)\}_{i=1}^{n}$, the multimodal contrastive model learns representations $Z_n = \{(z_{a_i}, z_{t_i})\}_{i=1}^{n}$ such that corresponding audio-text pairs are closer in the latent space than the unpaired inputs. Therefore, the contrastive loss of our model is defined as follows:

$$\mathcal{L}_{\text{contrastive}}(Z_n) = \sum_{i=1}^{n} \Bigg(\overbrace{-\frac{1}{2} \log \frac{\exp(\frac{1}{\tau} \frac{z_{a_i}^T z_{t_i}}{\|z_{a_i}\| \|z_{t_i}\|})}{\sum_{j=1}^{n} \exp(\frac{1}{\tau} \frac{z_{a_i}^T z_{t_j}}{\|z_{a_i}\| \|z_{t_j}\|})}}^{\text{audio-to-text loss}} \underbrace{-\frac{1}{2} \log \frac{\exp(\frac{1}{\tau} \frac{z_{a_i}^T z_{t_i}}{\|z_{a_i}\| \|z_{t_i}\|})}{\sum_{j=1}^{n} \exp(\frac{1}{\tau} \frac{z_{a_j}^T z_{t_i}}{\|z_{a_j}\| \|z_{t_i}\|})}}_{\text{text-to-audio loss}} \Bigg) \quad (1)$$

where τ is the temperature.

2.3 Multimodal Mixture of Experts Losses

Here we described the auxiliary losses that are used for the MoE encoders to stabilize the multimodal training and to avoid collapse to a few experts. These losses are designed to balance routing and expert utilization. Inspired by the (Riquelme et al., 2021; Mustafa et al., 2022), our auxiliary loss includes an importance loss to enforce a balanced gating weights schema however we define a different entropy based importance loss $\omega_{(imp)}$ for any expert. For a given token $x \in \mathbb{R}^D$, we define the gating weights across the E experts as $g(x) = \text{softmax}(Wx) \in \mathbb{R}^E$ where $W \in \mathbb{R}^{E \times D}$ represents the routing parameters. When processing a batch of n tokens, denoted as $\{x_i\}_{i=1}^n$, we represent the batch collectively as $X \in \mathbb{R}^{n \times D}$.

Importance Loss The importance loss ensures a balanced distribution of gating weights across experts. For each expert e among E experts, the importance is defined as: $\text{imp}_e(X) = \sum_{x \in X} g(x)_e$, where $g(x) = \text{softmax}(Wx)$ represents the gating weights, and W are the routing parameters. The importance loss in MoE is traditionally defined as the squared Coefficient of Variation (CoV) (Riquelme et al., 2021; Mustafa et al., 2022), i.e. $\mathcal{L}_{\text{imp}}(X) = \left(\frac{\text{std(imp}(X))}{\text{mean(imp}(X))}\right)^2$ where $\text{imp}(X) = \{\text{imp}_e(X)\}_{e=1}^E$. However, in this work we propose using an entropy based loss function for routing probabilities:

$$\mathcal{L}_{\text{imp}}(X) = -\mathcal{H}(\text{imp}(X)) = \sum_{e=1}^{E} \frac{\text{imp}_e(X)}{\sum_{i=1}^{E} \text{imp}_i(X)} \log\left(\frac{\text{imp}_e(X)}{\sum_{i=1}^{E} \text{imp}_i(X)}\right) \tag{2}$$

CoV loss forces low variance in expert assignments which can be too strict whereas the entropy loss encourages high entropy leading to smoother and more probabilistic routing instead of hard constraints. Essentially, CoV loss harshly penalizes any deviations from uniformity. While entropy maximization only encourages balance but allows some flexibility. This means the proposed entropy based loss allows for dynamic expert allocation instead of enforcing perfect balance. This is beneficial when some experts should specialize in certain data patterns. Moreover, entropy based importance loss is more gradient-friendly for the optimization as it is simply differentiable. This is specially important for large-scale MoE models with thousands of experts and make the training more stable.

Load Loss As we discussed above, the importance loss encourages all experts to receive roughly equal numbers of training examples. However, overall balancing of the weights does not guarantee that still, some small subset of experts do not get all the assignments. Particularly in realistic multimodal settings where we cannot assume there exist balanced data across different modalities, modality-specific experts tend to emerge naturally. In the modality imbalanced setting, all the tokens from the minority modality tend to be routed to a single expert. Despite this local congestion, on a global level, the overall routing still appears balanced, as tokens from the majority modality are evenly distributed among experts, ensuring that modality-agnostic auxiliary losses remain satisfied. The router can optimize the importance loss by perfectly balancing the majority modality's tokens but dropping all the minority modality's tokens. This however leads to unstable training and under performing models.

Therefore, following the initial proposal of (Shazeer et al., 2017), our loss objective also includes a load loss for load-balancing purposes. Similar to entropy based importance loss, we define an entropy based load loss. Since expert assignments are discrete and non-differentiable, a smooth approximation is used to facilitate gradient-based optimization. Given an input token x, the gating function computes routing probabilities via a noisy softmax: $g_{\text{noisy}}(x) = \text{softmax}(Wx + \epsilon)$, where W is the routing parameter matrix, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ represents noise added for smoothness, with $\sigma = \frac{1}{E}$. Since routing is determined by a Top-K selection, we define the K-th largest gating score as the threshold: $\tau_K(x) = \max_{k\text{-th}}(Wx + \epsilon)$.

The probability of expert e being selected for token x is: $p_e(x) = P\left((Wx)_e + \epsilon_{\text{new}} > \tau_K(x)\right)$, where $\epsilon_{\text{new}} \sim \mathcal{N}(0, \sigma^2)$. Using the Gaussian cumulative distribution function Φ, we can express this as: $p_e(x) = 1 - \Phi\left(\frac{\tau_K(x) - (Wx)_e}{\sigma}\right)$.

The load of expert e over a batch X is computed as: $\text{load}_e(X) = \sum_{x \in X} p_e(x)$. To encourage balanced expert utilization, we propose to maximize the entropy of the load distribution. This is similar to importance loss, where we use an entropy based loss, instead of minimizing the coefficient of variation (CoV) that is $\mathcal{L}_{\text{load}}(X) = \left(\frac{\text{std(load}(X))}{\text{mean(load}(X))}\right)^2$ where $\text{load}(X) = \{\text{load}_e(X)\}_{e=1}^E$. Our entropy based load loss is defined as:

$$\mathcal{L}_{\text{load}}(X) = -\mathcal{H}(\text{load}(X)) = \sum_{e=1}^{E} \frac{\text{load}_e(X)}{\sum_{i=1}^{E} \text{load}_i(X)} \log\left(\frac{\text{load}_e(X)}{\sum_{i=1}^{E} \text{load}_i(X)}\right) \tag{3}$$

The entropy based loss provide several benefits as explained in the importance loss section. Unlike CoV loss, which strictly enforces uniform expert assignments, entropy-based loss promotes balance while allowing flexibility in expert specialization. This softer approach leads to smoother, more probabilistic routing and improves gradient flow, making it particularly beneficial for large-scale MoE models with many experts.

Z-Loss for Router Stability Z-Loss is an auxiliary loss function designed to improve the stability of Mixture of Experts (MoE) models. It controls the magnitude of router logits, preventing them from becoming excessively large and causing numerical instabilities. In MoE architectures, router logits determine token assignments to experts. Large values in these logits can make training unstable by overconfidently assigning tokens to specific experts. Given activa-

tions $A = \{Wx_i\}_{i=1}^n$, where each entry is $a_{i,e} = (Wx_i)_e$, the Z-Loss regulates the activation magnitude of the router by penalizing large values, as follows:

$$\mathcal{L}_{\text{z-loss}}(X) = \frac{1}{n}\sum_{i=1}^{n}\Big(\log\sum_{e=1}^{E}\exp(a_{i,e})\Big)^2 \tag{4}$$

Z-Loss offers several advantages: it improves numerical stability by constraining large router logits, thereby reducing roundoff errors. Additionally, it prevents overconfident routing, ensuring that token assignments remain probabilistic rather than overly deterministic, which enhances flexibility in expert selection. Lastly, its squared log-sum-exp formulation results in smoother gradients, leading to more stable optimization, particularly in large-scale models.

Mutual Information (MI) Loss Mutual information loss is an auxiliary loss function designed to encourage diversity in expert selection by considering the entropy of expert assignments conditioned on modalities (Mustafa et al., 2022). It helps prevent expert collapse and ensures varied expert utilization by maximizing the dependency between token assignments and expert selection.

In each MoE layer, the router computes a gating matrix $G_m \in \mathbb{R}^{n_m \times E}$ for each modality m, where each row of G_m represents the probability distribution over E experts for one of the n_m tokens in the batch. The routing probabilities for token x_i are given by $p_m(\text{experts}|x_i) \in \mathbb{R}^E$. To ensure a balanced and effective routing strategy, two entropy-based loss functions are defined (Mustafa et al., 2022):

Local entropy loss: Encourages confident expert assignments by maximizing the entropy of individual token assignments:

$$\Omega_{\text{local}}(G_m) = \frac{1}{n_m}\sum_{i=1}^{n_m}\mathcal{H}(p_m(\text{experts}|x_i)) \tag{5}$$

where $\mathcal{H}(p) = -\sum_{e=1}^{E} p_e \log p_e$ denotes the entropy.

Global entropy loss: Encourages diversity across expert assignments by maximizing the entropy of the marginal expert distribution: $\Omega_{\text{global}}(G_m) = -\mathcal{H}(\tilde{p}_m(\text{experts}))$, where the marginal expert probability is: $\tilde{p}_m(\text{experts}) = \frac{1}{n_m}\sum_{i=1}^{n_m} p_m(\text{experts}|x_i)$. These two loss terms work together, the local entropy loss ensures that individual tokens have strong expert assignments, while the global entropy loss promotes expert diversity.

$$\mathcal{L}_{MI}(X) = -\frac{1}{M}\sum_{m=1}^{M}\mathcal{H}(\tilde{p}_m(\text{experts})) + \mathcal{H}\Big(\frac{1}{M}\sum_{m=1}^{M}\tilde{p}_m(\text{experts})\Big) \tag{6}$$

Evidence Lower Bound (ELBO) Loss Unlike the other auxiliary losses, which primarily focus on balancing routing and expert utilization (e.g., importance, load, Z-loss, and mutual information), the ELBO loss serves a dual purpose: 1) it encourages expert specialization by ensuring assigned experts explain modality-specific tokens well, 2) it prevents expert collapse via KL regularization that encourages diverse expert utilization. Since ELBO loss directly affects how well experts specialize in different data patterns, its contribution to the final loss function may need to be weighted differently.

Let X_m be the set of tokens from modality m, and let E be the set of experts. The approximate marginal probability of an expert over the tokens in modality m is defined as: $p_m(e) = \frac{1}{|X_m|}\sum_{x\in X_m} p(e|x)$, where $p(e|x)$ is the probability of selecting expert e for token x and $|X_m|$ is the number of tokens in modality m. This represents the likelihood of an expert being used across all tokens from a given modality.

Since the true posterior distribution $p(e|X_m)$ is intractable, we approximate it using a variational distribution $q(e|X_m)$. The ELBO formulation is obtained by minimizing the KL divergence:

$$D_{\text{KL}}(q(e|X_m)||p(e|X_m)) = \mathbb{E}_{q(e|X_m)}\left[\log q(e|X_m) - \log p(e|X_m)\right] \tag{7}$$

which leads to the Evidence Lower Bound (ELBO) (Belghazi et al., 2018):

$$\mathcal{L}_{\text{ELBO}}(X_m) = \mathbb{E}_{q(e|X_m)}\left[\log p(X_m|e)\right] - D_{\text{KL}}(q(e|X_m)||p(e)) \tag{8}$$

where the first term ensures that the assigned experts explain the tokens well, and the second term regularizes the expert distribution towards a prior $p(e)$. Since we approximate $p(e|X_m)$ using routing probabilities, we set: $q(e|X_m) \approx p_m(e) = \frac{1}{|X_m|}\sum_{x\in X_m} p(e|x)$. Thus, the ELBO loss for modality m in our framework becomes:

$$\mathcal{L}_{\text{ELBO}}(X_m) = \sum_{x\in X_m}\mathbb{E}_{p(e|x)}\left[\log p(X_m|e)\right] - D_{\text{KL}}(p_m(e)||p(e)) \tag{9}$$

where $p(e)$ is a prior distribution over experts, which we set to be uniform. To compute the expert likelihood $p(X_m \mid e)$ in the embedding space we use a prototype-Gaussian per expert, $\mathcal{N}(X_m; \mu_e^{(m)}, \sigma^2 I)$, with $\mu_e^{(m)}$ updated by EMA over tokens softly routed to e. Essentially, each expert e defines a simple density over token embeddings. The sequence-level likelihood is:

$$\log p(X_m \mid e) = -\frac{1}{2\sigma^2}\sum_{t=1}^{|X_m|}\|x_t - \mu_e^{(m)}\|_2^2 - \frac{|X_m|d}{2}\log(2\pi\sigma^2) \tag{10}$$

We aggregate token-level routing into a set-level posterior $q(e \mid X_m) = p_m(e) = \frac{1}{|X_m|}\sum_{x\in X_m} p(e \mid x)$. This formulation is decoder-free and adds negligible parameters. And the total ELBO loss becomes:

$$\mathcal{L}_{\text{ELBO}}(X) = -\frac{1}{M}\sum_{m=1}^{M}\mathcal{L}_{\text{ELBO}}(X_m) \tag{11}$$

The proposed ELBO loss encourages expert specialization. The expectation term ensures experts focus on modality-specific tokens. It simultaneously prevents expert collapse as the KL regularization ensures experts are utilized fairly. Additionally it balances routing decisions by integrating routing probabilities.

Final Objective Function Our final loss function is a combination of the contrastive loss and all the auxiliary losses. Auxiliary losses regularize the contrastive loss and ensure effective expert utilization and modality utilization. In image-text multimodal typically the image modality contains richer information than the text modality. However, in audio-text multimodal usually the text modality captures richer information and often times audio does not provide additional information other than the spoken words. For this reason it is trickier to balance the modalities and expert utilization. Moreover, the less important modality in CoPRIME (i.e. audio) contains more tokens than the more important counterpart (i.e. text) which is the opposite case for image-text multimodal mixture of expert models. We use a different λ for the ELBO versus other auxiliary losses to address this. The other auxiliary losses (importance, load, Z-loss, and MI) focus on ensuring a balanced and stable routing mechanism, preventing expert underuse or overuse. ELBO, however, promotes specialization and separation of experts for different modalities while maintaining fair utilization. Using a different λ_{ELBO} allows for finer control over this trade-off where too high a value may force excessive specialization, while too low a value may lead to underutilized experts.

$$\mathcal{L}(X) = \mathcal{L}_{\text{contrastive}}(X) + \lambda_{\text{ELBO}}\mathcal{L}_{\text{ELBO}}(X) + \lambda_{\text{aux}}\big(\mathcal{L}_{\text{imp}}(X) + \mathcal{L}_{\text{load}}(X) + \mathcal{L}_{\text{z-loss}}(X) + \mathcal{L}_{MI}(X)\big) \tag{12}$$

3 EXPERIMENTS

In this section, we present the experimental evaluation of CoPRIME model for emotion recognition tasks. We utilize the LibriSpeech960 dataset (Panayotov et al., 2015) for pre-training and evaluate the model's performance on the MOSEI (Zadeh et al., 2018) and IEMOCAP (Busso et al., 2008) datasets under various settings. **LibriSpeech960:** The LibriSpeech960 dataset (Panayotov et al., 2015) comprises approximately 960 hours of English speech data, accompanied by transcriptions. It serves as a comprehensive resource for training models on large-scale speech recognition tasks. **MOSEI:** The Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) dataset (Zadeh et al., 2018) is a large-scale multimodal dataset containing over 23,000 video clips annotated with sentiment and emotion labels. For our experiments, we focus on the emotion recognition aspect, which includes six emotion classes: Happy, Sad, Angry, Fear, Surprise, and Disgust. **IEMOCAP:** The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset (Busso et al., 2008) consists of approximately 12 hours of audiovisual data from dyadic sessions, annotated with emotion labels. We utilize the emotion recognition annotations, which include emotions such as Happy, Sad, Angry, Neutral, and others.

3.1 Evaluation Scheme and Experimental Setup

Table 1: Model information for base and large variants of CoPRIME and LIMoE.

	CoPRIME-B/32	LIMoE-B/32	CoPRIME-L/32	LIMoE-L/32
Transformer Blocks	3	3	6	6
Attention Heads	8	8	8	8
MoE Blocks	6	6	12	12
Experts per MoE Block	8	8	16	16
Routing Mechanism	Top-2	Top-2	Top-4	Top-4
Hidden Size	512	512	768	768
MLP Size	2048	2048	3072	3072
Patch Size	32x32	32x32	32x32	32x32

We evaluate the performance of CoPRIME under two primary settings: 1) Cross-domain transfer learning: The model is pre-trained on the LibriSpeech960 dataset and directly evaluated on the MOSEI and IEMOCAP datasets using 0-shot and 10-shot learning approaches. 2) Domain-specific transfer learning: The model is pre-trained on the LibriSpeech960 dataset and then fine-tuned on the MOSEI dataset. Subsequently, it is evaluated on the IEMOCAP dataset using 0-shot and 10-shot learning approaches. MOSEI dataset is much larger than IEMOCAP, hence it is more suitable for fine-tuning large architectures. Therefore, we fine tune on the larger dataset and evaluate the transfer learning quality on the smaller dataset. Table 1 shows the model information for the two variants of CoPRIME-B/32 (i.e. base model) and CoPRIME-L/32 (i.e. large variant). Additionally, we use a dense baseline in the experiments which cosists of only transformer encoder blocks and no MoE blocks. The dense baseline architecture is similar to that of CoPRIME-B/32 in terms of the number of blocks, attention heads, etc. We use AdamW optimizer with learning rate of 1e-4, and batch size of 256 throughout all experiments.

We perform a log-spaced grid search over λ_{ELBO} and The λ_{aux}. As we can see from the sensitivity analysis heatmap in Figure 3 across a wide plateau $\lambda_{\text{ELBO}} \in [0.02, 0.08]$ and $\lambda_{\text{aux}} \in [0.005, 0.04]$, CoPRIME's task performance remains within 1.3% of the best setting, while convergence speed and stability (variance of validation loss and gradient norms) degrade outside this interval. Our chosen setting sits on a Pareto front between final accuracy and convergence speed. The λ_{aux} and λ_{ELBO} were tuned and values of 0.02 and 0.04 are used, respectively in all experiments. Source code is available at: https://github.com/hanadk/coprime.

3.2 Results

Cross-domain transfer learning: In this case we pre-trained CoPRIME on the LibriSpeech960 dataset and evaluated on both MOSEI and IEMOCAP datasets using 0-shot and 10-shot learning approaches. LibriSpeech is mostly used for automatic speech recognition while MOSEI and IEMOCAP are specifically used for emotion recognition. Therefore, the domain of pre-training is a little different than

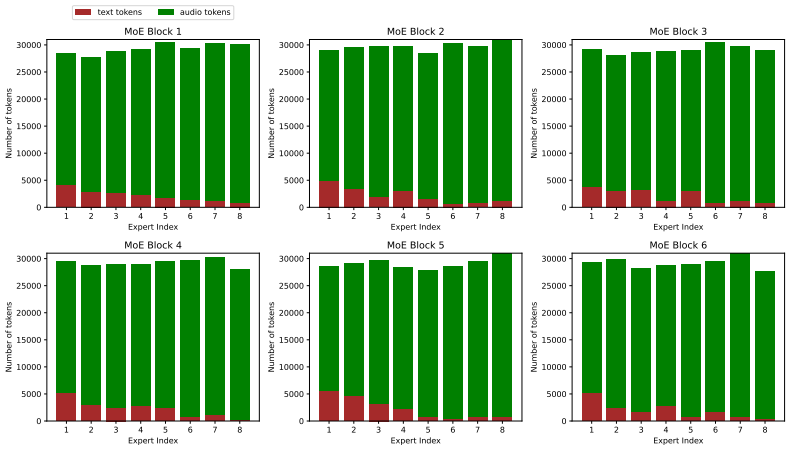

Figure 2: Token distribution on a batch of size 256 for CoPRIME-B/32 model with 6 MoE blocks each having 8 experts. We can see the extreme token imbalance even with audio patches of 32x32.

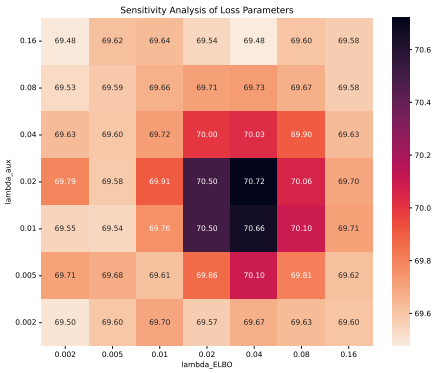

Figure 3: Sensitivity analysis with respect to regularization parameters (MOSEI 10-shot accuracy).

the evaluation target domain. Table 2 presents the results of this cross-domain transfer learning. As we can see from the table, CoPRIME outperforms the dense baseline which shows the effectiveness of the MoE blocks and architecture. Additionally, we observe that CoPRIME-L/32 outperforms CoPRIME-B/32 which is expected considering the increase in size and capacity of the model. Moreover, it is noteworthy to mention that 10-shot transfer learning yields higher accuracy than zero-shot learning specially since the target domain is quite different than the source (i.e. pre-training) domain. As for the comparison with the state-of-the-art, Hubert-large (Hsu et al., 2021) achieves 67.62% accuracy on IEMOCAP emotion recognition task. Our multimodal model achieves competitive accuracy in 0-shot while being significantly more efficient. In 10-shot CoPRIME accuracy is significantly higher.

Domain-specific transfer learning: In this case, we first pre-train CoPRIME on LibriSpeech dataset. We then fully fine-tune the model on the MOSEI dataset, and finally evaluate the model on the IEMOCAP dataset under 0-shot and 10-shot learning scenarios. Since MOSEI and IEMOCAP are both emotion recognition tasks, the source and target domains are similar and this is considered intra-domain transfer learning. Moreover, different emotion categories of these datasets do not affect the results as we use linear few-shot evaluation for 10-shot and class prototype similarity for 0-shot. Table 3 shows the results of domain-specific transfer learning. As we can see, fine-tuning on MOSEI significantly improved the accuracy of emotion recognition on IEMOCAP dataset, by an average of 2.6%. It is noteworthy to mention that the 10-shot transfer learning accuracy of CoPRIME is competitive to the SOTA models that are directly trained on the dataset. CORECT (Nguyen et al., 2023) achieves 84.7% accuracy on IEMOCAP with directly training a supervised multimodal classification model on the dataset. CoPRIME achieved 82.17% using 10-shot transfer learning which shows the representations learned from our multimodal contrastive framework capture rich information and can generalize across tasks.

Routing distribution: Figure 2 shows the routing distribution in different MoE blocks. We observe the emergence of both modality-specific experts, and multimodal experts which process both images and texts. This validates the effectiveness of our losses in encouraging diversity as well as modality specialization to a certain degree.

Loss ablation studies: Our final objective function consists of a contrastive loss, ELBO loss, and a combination of 4 entropy-based auxiliary losses (importance, load, z-loss, and mutual information). Therefore, there are five auxiliary losses and the total number of their possible combinations is 27 ($\sum_{r=0}^{N} \binom{N}{r} - N = 2^N - N$ where $N = 5$ and $r = 2$). We performed a brute force usage of auxiliary loss combinations to analyze the impact of each individual loss. We provide the ablation study of the CoPRIME loss in Table 4. Each row in the table shows the highest performing model with and without that specific loss function. As we can see from the

Table 2: Cross-domain transfer learning results of different methods on MOSEI and IEMOCAP datasets under 0-shot and 10-shot settings. The accuracy & F1 values are shown in percentages (%).

Model	MOSEI 0shot			MOSEI 10shot			IEMOCAP 0shot		IEMOCAP 10shot	
	Emotion	Sentiment		Emotion	Sentiment		Emotion		Emotion	
	Acc	Acc	F1	Acc	Acc	F1	Acc	F1	Acc	F1
Dense-B/64	44.53	52.01	52.49	56.61	63.44	64.09	43.98	44.54	57.91	57.39
LIMoE-B/64	58.13	67.80	68.29	70.56	78.56	79.18	57.88	58.54	71.31	70.73
CoPRIME-B/64	59.68	67.99	68.47	71.10	77.22	77.87	60.75	61.32	72.52	71.95
Dense-B/32	47.87	55.43	55.83	58.97	66.00	66.65	46.91	47.48	61.01	60.46
LIMoE-B/32	61.44	69.13	69.61	72.50	78.73	79.41	61.28	61.90	74.73	74.14
CoPRIME-B/32	63.75	70.90	71.33	74.95	81.00	81.68	63.26	63.91	75.22	74.71
Dense-L/64	49.02	56.62	57.04	60.99	66.78	67.38	48.43	49.04	61.59	61.06
LIMoE-L/64	62.73	70.57	70.99	74.24	80.61	81.23	61.99	62.61	76.23	75.65
CoPRIME-L/64	63.86	71.29	71.76	74.89	81.80	82.49	63.58	64.24	76.45	75.85
Dense-L/32	51.35	59.24	59.69	64.37	70.04	70.64	50.75	51.31	65.31	64.73
LIMoE-L/32	64.57	72.43	72.87	77.93	83.06	83.68	65.28	65.84	78.33	77.78
CoPRIME-L/32	**66.43**	**74.38**	**74.82**	**78.21**	**84.8**	**85.44**	**66.17**	**66.78**	**79.7**	**79.15**

Table 3: Domain-specific transfer learning results of different methods on IEMOCAP datasets (after being fine-tuned on MOSEI) under 0-shot and 10-shot settings. The error bars are the 2-sigma variations calculated using 10-fold cross validation. The accuracy & F1 values are shown in percentages.

Model	IEMOCAP 0shot		IEMOCAP 10shot	
	Emotion		Emotion	
	Acc	F1	Acc	F1
Dense-B/32	49.94 ± 0.74	50.21 ± 0.75	64.72 ± 0.71	64.36 ± 0.69
LIMoE-B/32	63.82 ± 0.52	63.48 ± 0.57	76.92 ± 0.42	76.83 ± 0.39
CoPRIME-B/32	65.47 ± 0.45	65.3 ± 0.49	78.36 ± 0.38	78.2 ± 0.41
Dense-L/32	54.78 ± 0.68	55.07 ± 0.65	67.32 ± 0.61	67.85 ± 0.59
LIMoE-L/32	67.8 ± 0.54	67.51 ± 0.53	79.92 ± 0.48	80.19 ± 0.44
CoPRIME-L/32	$\mathbf{68.9 \pm 0.56}$	$\mathbf{69.28 \pm 0.51}$	$\mathbf{82.17 \pm 0.48}$	$\mathbf{82.5 \pm 0.46}$

Table 4: Across 27 combinations, each row shows the best accuracy (%) of all combinations that included the auxiliary loss ✓vs. those that did not ✗. Validation accuracy is the average contrastive accuracy in a minibatch of size 256 on MOSEI dataset.

	Validation		MOSEI			
			0shot		10shot	
Auxiliary Loss	✗	✓	✗	✓	✗	✓
Importance	64.4	64.6	59.2	59.3	70.6	70.9
Load	64.5	64.6	59.1	59.1	69.8	69.7
Z-loss	64.3	64.4	59.6	59.8	70.1	70.3
MI	64.1	64.8	59.7	60.4	69.7	70.4
ELBO	62.5	64.8	58.9	60.3	69.4	70.9

table, all losses are beneficial with the exception of load loss. Load loss significantly improves efficiency while penalizing the accuracy in a negligible amount. We also observe that ELBO loss has a significant impact on the training, and hence the reason we used a different λ for it.

4 CONCLUSION

In this work, we introduced CoPRIME, a novel sparse MoE model designed to advance multimodal representation learning. By leveraging contrastive learning, adaptive expert routing, and spectrogram-based audio tokenization, CoPRIME effectively addresses key challenges in multimodal learning, particularly the imbalance in token density between audio and text, while maintaining computational efficiency. CoPRIME uses a ViT-style spectrogram tokenization approach in audio-text multimodal settings, transforming raw waveforms into patch-based embeddings. CoPRIME incorporates a sparse MoE framework with an entropy-regularized routing mechanism. Traditional MoE models often suffer from expert overloading or under-utilization, leading to unstable training and poor generalization. To mitigate these issues, we introduced a set of entropy-based auxiliary losses that ensure balanced and dynamic activation of experts across modalities. Additionally, we proposed an ELBO-based loss function that optimizes expert specialization while preventing collapse and encouraging diversity.

Experimental results on MOSEI and IEMOCAP demonstrate that CoPRIME significantly outperforms dense transformers and prior SOTA multimodal MoE models in both zero-shot and few-shot learning settings. On IEMOCAP, CoPRIME-L/32 achieved 79.7% accuracy in 10-shot learning, surpassing LIMoE accuracy significantly. Similarly, on MOSEI, CoPRIME-L/32 reached 84.8% sentiment accuracy in 10-shot learning, demonstrating substantial improvements over dense transformers and LIMoE while maintaining high computational efficiency. Our loss ablation studies further validate the effectiveness of our approach, showing that ELBO loss contributes to an average 2.5% gain in contrastive accuracy, while entropy-based auxiliary losses lead to more stable training and balanced expert utilization.

References

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Yoshua Bengio. Deep learning of representations: Looking forward. In *International conference on statistical language and speech processing*, pages 1–37. Springer, 2013.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

Mark Collier, Efi Kokiopoulou, Andrea Gesmundo, and Jesse Berent. Routing networks with co-training for continual learning. *arXiv preprint arXiv:2009.04381*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012/.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. Cross-token modeling with conditional computation. *arXiv preprint arXiv:2109.02008*, 2021.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018.

Yifei Ming and Yixuan Li. Understanding retrieval-augmented task adaptation for vision-language models. *arXiv preprint arXiv:2405.01468*, 2024.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35: 9564–9576, 2022.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pages 4348–4380. PMLR, 2023.

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. *arXiv preprint arXiv:2311.04507*, 2023.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Yunwei Ren and Yuanzhi Li. On the importance of contrastive loss in multimodal learning. *arXiv preprint arXiv:2304.03717*, 2023.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Yihao Xue, Siddharth Joshi, Dang Nguyen, and Baharan Mirzasoleiman. Understanding the robustness of multimodal contrastive learning to distribution shift. *arXiv preprint arXiv:2310.04971*, 2023.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35: 7103–7114, 2022.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

From Token Imbalance to Balanced Routing: An ELBO-Regularized Probabilistic Framework for Contrastive Multimodal Learning Supplementary Materials

A MOTIVATION AND INTUITION

CoPRIME builds upon a series of prior research starting from the seminal paper of Sparsely Gated MoE which introduced the importance loss for balancing expert utilization. Then GShard, Switch Transformer, and V-MoE introduced the buffer capacity and added auxiliary load loss to tackle expert imbalance for NLP and vision. Subsequently, LIMoE added another auxiliary MI loss to handle token imbalance in multimodal tasks. We extend these ideas to the audio-text multimodal setting. Due to the high dimensionality of audio spectrograms which is substantially larger than typical image data, we introduce an additional ELBO loss for multimodal per-token routing to address token imbalance more effectively. CoPRIME targets a distinct challenge not fully explored in prior multimodal MoE work: the extreme token imbalance. Unlike image-text ($\sim$12:1 ratio), audio produces far more tokens than text due to spectrogram tokenization ($\sim$50:1 ratio) causing optimization instability and expert collapse. CoPRIME tackles this via its ELBO loss and entropy-based aux losses.

Our MoE architecture resembles LIMoE one, but with novel objective functions. We introduced an ELBO loss to address token imbalance more effectively in audio-text multimodal which is way more imbalanced than image-text multimodal. It encourages expert specialization while preventing expert collapse. Additionally we replace CoV-based load/importance losses with entropy-based variants that promote flexible expert utilization. This encourages balance while allowing controlled flexibility, leading to a more dynamic and probabilistic routing mechanism instead of imposing hard constraints.

The design of our ELBO and entropy-based auxiliary losses are rooted in principles of probabilistic regularization and information theory. The ELBO is derived from a variational lower bound on the marginal likelihood of modality-specific expert assignments. By maximizing this likelihood while minimizing KL divergence to a uniform prior, we promote diverse yet balanced expert usage. This prevents a subset of experts from dominating due to modality bias (overwhelming in volume but not in semantic content). The entropy-based importance and load losses further enhance probabilistic routing flexibility, providing gradient-friendly regularization without enforcing rigid uniformity. Empirical results (Table 4) validates these theoretical motivations: ELBO alone boosts contrastive accuracy by 2.3%, while entropy-based losses improve stability and convergence.

In conclusion, CoPRIME introduces theoretically grounded mechanisms for handling token imbalance, and empirically demonstrates their necessity for effective contrastive alignment.

B THEORETICAL RATIONALE FOR LOSS FUNCTIONS

The design of our ELBO and entropy-based auxiliary losses are rooted in principles of probabilistic regularization and information theory. The ELBO is derived from a variational lower bound on the marginal likelihood of modality-specific expert assignments. By maximizing this likelihood while minimizing KL divergence to a uniform prior, we promote diverse yet balanced expert usage. This prevents a subset of experts from dominating due to modality bias (overwhelming in volume but not in semantic content). The entropy-based importance and load losses further enhance probabilistic routing flexibility, providing gradient-friendly regularization without enforcing rigid uniformity. Empirical results in Table 4 validates these theoretical motivations. ELBO alone boosts contrastive accuracy by 2.3%, while entropy-based losses improve stability and convergence.

C ELBO CONTRIBUTION BEYOND MI/ENTROPY REGULARIZERS

The MI/entropy-based auxiliary losses (importance, load, and the MI objective) shape the routing distributions. They encourage balanced usage (global entropy/load) and confident per-token assignments (local entropy) but do not require experts to model the geometry of the data in feature space.

The ELBO introduces two complementary, data-aware pressures:

1. Expert specialization via likelihood (feature geometry): The term $\mathbb{E}_{q(e|X_m)}\big[\log p(X_m|e)\big]$ directly rewards assignments to experts that result in high feature likelihood (e.g., small within-expert squared distance under the Gaussian), pushing experts to model coherent, well-separated regions of representation space tied to semantic or acoustic phenomena. MI/entropy alone can balance traffic without ensuring any expert actually fits the data manifold.

2. Sample/modality-level posterior-prior alignment: The KL term $\mathrm{KL}\big(q(e|X_m)\|p(e)\big)$ with $q(e|X_m) = p_m(e)$ regularizes the sequence/set-level posterior, complementing batch-marginal balancing (global entropy). This mitigates unstable solutions where token-level posteriors are sharp yet each sample still collapses to the same expert.

In summary, our findings show:

- Lower within-expert scatter and higher between-expert separation of token embeddings (lower trace of covariance, higher centroid separation).
- More stable utilization (lower variance of per-expert load over training), especially under strong modality imbalance.
- Faster/steadier convergence of the contrastive objective (fewer epochs to a fixed retrieval/accuracy threshold).

The MI/entropy terms keep usage balanced and avoid starvation. The ELBO ensures that the balanced usage is meaningful, i.e., experts explain the embeddings they receive. In ablations, MI-only models distribute load but form blurrier clusters while adding ELBO sharpens clusters and typically improves retrieval with minimal extra cost.

D ATTRIBUTION OF PERFORMANCE GAINS TO ARCHITECTURAL OR LOSS DESIGN

We ran loss ablations (Table 4) to isolate the contribution of each component, confirming the utility of all losses. We also implemented LIMoE-style variants and found CoPRIME to outperform them on emotion recognition. These findings indicate that CoPRIME's performance gains originates from improved routing under modality imbalance and loss innovations, rather than increased parameter count. CoPRIME introduces theoretically grounded mechanisms for handling token imbalance, and empirically demonstrates their necessity for effective contrastive alignment.

E REGARDING LOAD LOSS AND ITS IMPACT ON EFFICIENCY

The load loss was introduced to improve routing balance and stability in sparse MoE architectures, especially under modality imbalance where audio tokens far outnumber text. As discussed in Section 2.3, we adopt an entropy-based formulation to softly encourage expert balance without imposing hard uniformity. This prevents expert collapse and token congestion, a known failure mode in multimodal MoEs. While prior works such as GShard and LIMoE validate the utility of load loss, our own ablation study (Table 4) further substantiates its role; the inclusion of load loss improves the contrastive accuracy from 64.5% to 64.6%, it also enhances efficiency which is defined as routing stability, expert utilization balance, and improved training dynamics in modality-imbalanced settings. We have included empirical efficiency indicators (e.g. token-to-expert assignment distributions and activation sparsity heatmaps) in our paper.

F REGARDING ENTROPY-BASED VS COV-BASED IMPORTANCE LOSS

Traditionally, importance loss is computed via the CoV to enforce uniform expert usage. However, CoV imposes rigid constraints and penalizes minor deviations. In contrast, we propose an entropy-based loss (Section 2.3) that encourages balanced yet flexible expert allocation. This is especially advantageous when expert specialization is desirable (e.g., modality-specific routing). Our intuition is that entropy loss offers smoother gradients and is more robust in large-scale setups with noisy or diverse data. Empirically, we have also run internal ablations with CoV-based variants, and results were slightly weaker (64.4% to 64.5%). Replacing CoV with entropy improves stability and generalization: validation accuracy improves from 64.4% to 64.6% (Table 4), and consistent gains are observed across few-shot setups. While these deltas may appear modest, they accumulate with other auxiliary losses (e.g., ELBO), leading to more stable convergence.

G LOSS FUNCTION COMPLEXITY AND TRADE-OFF COEFFICIENTS

CoPRIME indeed introduces six loss terms (contrastive, ELBO, and four entropy-based auxiliary losses) to regulate expert specialization, diversity, and balanced routing. This is necessitated by the unique challenge of audio-text multimodal alignment, especially with spectrogram-based audio tokenization, which introduces an extreme token imbalance. Tuning multiple loss weights is non-trivial; so to mitigate this, we design the framework with two trade-off coefficients: one dedicated to ELBO and one shared across the four auxiliary losses. This design was motivated by the complementary roles of the losses: ELBO regularizes global expert specialization, while the auxiliary losses promote balanced, diverse routing. The model showed robustness across a range of lambda values during tuning (Section 3.2).

H NECESSITY OF ALL LOSSES AND LOAD LOSS EFFECTIVENESS

As shown in Table 4, we conducted a comprehensive ablation over 27 combinations of auxiliary losses, confirming that the ELBO loss contributes the most (2.5% gain in contrastive accuracy), while Load loss primarily enhances stability and routing smoothness. While its impact on final accuracy is limited, it plays a complementary role in preventing expert saturation in modality-imbalanced settings, as discussed in Section 2.3.

I COMPARISON WITH OTHER MOE-BASED MULTIMODAL MODELS AND BEHAVIORAL ANALYSIS OF ROUTING

CoPRIME targets a distinct challenge not fully explored in prior multimodal MoE work: the extreme token imbalance. Unlike image-text ($\sim$12:1 ratio), audio produces far more tokens than text due to spectrogram tokenization ($\sim$50:1 ratio) causing optimization instability and expert collapse. CoPRIME tackles this via its ELBO loss and entropy-based auxiliary losses. A direct comparison with LIMoE is infeasible due to lack of official code and differing modalities/datasets. To ensure fairness, we have implemented LIMoE-style variants ourselves and observed that CoPRIME outperforms them in both zero-shot and few-shot emotion recognition tasks. We also benchmarked against strong baselines like dense transformers and SOTA models on the tasks (HuBERT & CORECT).

J HYPERPARAMETER TUNING AND MODEL SELECTION

We conducted a brute-force ablation study on loss combinations (Section 3.2 and Table 4), validating the effectiveness of all five losses. As for the weights, (0.02, 0.04) were determined through grid search and proved robust across experiments. We conduct an extensive hyperparameter tuning covering lambda coefficients, top-K routing thresholds, and analyze training stability under different expert counts. We will release code to facilitate reproducibility and ease of adoption.

K COMPUTE COSTS

The datasets used in this work are relatively small and the network architecture design (in terms of number of blocks, etc.) is also relatively small to match that. Sparse MoE architecture also makes the training more efficient. All the experiments were ran on a single NVIDIA GH200 GPU with 96GB VRAM on LambdaLabs cloud.

L SOCIETAL IMPACTS

In terms of environmental impact, training large models is costly. In this work we used relatively small datasets and small network architecture, so that all the experiments were run on a high-end PC with Titan RTX GPU. However, if someone wants to train a larger version of CoPRIME on larger datasets it could become costly. In either case, sparse MoE architecture as used in CoPRIME helps in reducing the compute cost specially at inference time.

As a positive societal impact, in another project we are working on a mental health data where these audio-text multimodal models are used on clinical interviews of children to predict various types of disorders. This has great positive impacts in helping psychiatrists in early diagnosis of mental disorders. We advocate for the responsible use of CoPRIME, particularly for potential clinical applications, ensuring consent, de-identification, and human-in-the-loop deployment. Our clinical speech analysis work was approved by the Nova Scotia Health Research Ethics Board (file number: 100266).