FuseMoE: Mixture-of-Experts Transformers for Fleximodal Fusion

Xing Han[1] Huy Nguyen[*2] Carl Harris[*3] Nhat Ho[+2] Suchi Saria[+1 4]

Abstract

As machine learning models in critical fields increasingly grapple with multimodal data, they face the dual challenges of handling a wide array of modalities, often incomplete due to missing elements, and the temporal irregularity and sparsity of collected samples. Successfully leveraging this complex data, while overcoming the scarcity of high-quality training samples, is key to improving these models' predictive performance. We introduce "FuseMoE", a mixture-of-experts framework incorporated with an innovative gating function. Designed to integrate a diverse number of modalities, FuseMoE is effective in managing scenarios with missing modalities and irregularly sampled data trajectories. Theoretically, our unique gating function contributes to enhanced convergence rates, leading to better performance in multiple downstream tasks. The practical utility of FuseMoE in real world is validated by a challenging set of clinical risk prediction tasks.

1. Introduction

Multimodal fusion is a critical and extensively studied problem in many significant domains (Shaik et al., 2023; Yang et al., 2007; Tsai et al., 2019; Cao et al., 2023), such as sentiment analysis (Han et al., 2021b; Majumder et al., 2018), image and video captioning (Karpathy & Fei-Fei, 2015; Johnson et al., 2016b), and medical prediction (Huang et al., 2020b; Soenksen et al., 2022). Previous research has shown that embracing multimodality can improve predictive performance by capturing complementary information across modalities, outperforming single-modality approaches in similar tasks (Potamianos et al., 2003; Huang et al., 2020a). However, an ongoing challenge is the creation of scalable frameworks for fusing multimodal data, and in creating reliable models that consistently surpass their single-modal counterparts.

Handling a variable number of input modalities remains an open challenge in multimodal fusion, due to challenges with scalability and lack of a unified approaches for addressing missing modalities. Many existing multimodal fusion methods are designed for only two modalities (Han et al., 2021a; Zhou et al., 2019; Zhang et al., 2023), rely on costly pairwise comparisons between modalities (Tsai et al., 2019), or employ simple concatenation approaches (Soenksen et al., 2022), rendering them unable to scale to settings with a large number of input modalities or adequately capture inter-modal interactions. Similarly, existing works are either unable to handle missing modalities entirely (Zhang et al., 2023; Zhan et al., 2021) or use imputation approaches (Tran et al., 2017; Liu et al., 2023; Soenksen et al., 2022) of varying sophistication. The former methods restrict usage to cases where all modalities are completely observed, significantly diminishing their utility in settings where this is often not the case (such as in clinical applications); the latter can lead to suboptimal performance due to the inherent limitations of imputed data. In addition, the complex and irregular temporal dynamics present in multimodal data have often been overlooked (Zhang et al., 2023; Tipirneni & Reddy, 2022), with existing methods often ignoring irregularity entirely (Soenksen et al., 2022) or relying on positional embedding schemes (Tsai et al., 2019) that may not be appropriate when modalities display a varying degree of temporal irregularity. Consequently, there is a pressing need for more advanced and scalable multimodal fusion techniques that can efficiently handle a broader set of modalities, effectively manage missing and irregular data, and capture the nuanced inter-modal relationships necessary for robust and accurate prediction. We use the term **FlexiModal Data** to capture several of these key aspects, which haven't been well-addressed by prior works:

> *"Flexi" suggests flexibility, indicating the possibility of having any combination of modalities, even with arbitrary missingness or irregularity.*

A practical example of FlexiModal data can be seen in clinical applications, where extensive monitoring results in the

[*]Equal Contribution, [+]Equal Advising. [1]Department of Computer Science, Johns Hopkins University, Baltimore, MD. [2]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, TX. [3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD. [4]Bayesian Health, New York, NY. Correspondence to: Xing Han <xhan56@jhu.edu>.

Table 1. We evaluated the characteristics of FuseMoE against various established benchmarks. The pipeline approach by Soenksen et al. (2022) relies on a simple feature extraction scheme for each modality, followed by concatenation and classification. The approach doesn't incorporate irregularities or missing data in its process, but its use of concatenation and zero-imputation for missing modalities allows it to be adapted to FlexiModal settings. Both Zhang et al. (2023) and Zadeh et al. (2017) tackle multi-modality fusion issues, but as modalities increase, their method demands exponentially more cross-modal computations and significant model architecture modifications. Finally, Mustafa et al. (2022) presents MoE for language-image alignment in multimodal learning, yet it also requires substantial adjustments for the more intricate and universal FlexiModal context we explore.

Method	Type	Irregularity	Missingness	Num of Mods	Theory	Can Adapt to FlexiModal?
Soenksen et al. (2022)	Data Pipeline	✗	✗	≥4	✗	✓
Zhang et al. (2023)	Modality Fusion	✓	✗	2	✗	✗
Zadeh et al. (2017)	Modality Fusion	✗	✗	3	✗	✗
Mustafa et al. (2022)	Multimodal MoE	✗	✗	2	✗	✗
FuseMoE	This Paper	✓	✓	≥4	✓	Adapted

accumulation of comprehensive electronic health records (EHRs) for each patient. A typical EHR encompasses diverse data types, including tabular data (e.g., age, demographics, gender), image data (such as X-rays, magnetic resonance imaging, and photographs), clinical notes, physiological time series (e.g., ECG and EEG) and vital signs (blood chemistry, heart rate, etc.). In this setting, we observe variety of modalities, sampled with varying irregularity and a high degree of missingness. These challenges, coupled with the relevance of predictive models to clinical settings, render ICU predictions an ideal use case to demonstrate our approach for handling FlexiModal data.

Contributions In this paper, we introduce a novel mixture-of-experts (MoE) framework, which we call FuseMoE, specifically designed to enhance the multimodal fusion of FlexiModal Data. FuseMoE incorporates sparsely gated MoE layers in its fusion component, which are adept at managing distinct tasks and learning optimal modality partitioning. In addition, FuseMoE surpasses previous transformer-based methods in scalability, accommodating an unlimited array of input modalities. Furthermore, FuseMoE routes each modality to designated experts that specialize in those specific data types. This allows FuseMoE to adeptly handle scenarios with missing modalities by dynamically adjusting the influence of experts primarily responsible for the absent data, while still utilizing the available modalities. Lastly, another key innovation in FuseMoE is the integration of a novel Laplace gating function, which not only theoretically ensures better convergence rates compared to Softmax functions, but also demonstrates better predictive performance. We demonstrate that our approach shows superior ability, as compared to existing methods, to integrate diverse input modality types with varying missingness and irregular sampling on three challenging ICU prediction tasks.

2. Related Works

Multimodal Fusion Initial approaches to multimodal fusion incorporated techniques such as kernel-based methods (Bucak et al., 2013; Chen et al., 2014; Poria et al., 2015), graphical models (Nefian et al., 2002; Garg et al., 2003; Reiter et al., 2007), and neural networks (Ngiam et al., 2011; Gao et al., 2015; Nojavanasghari et al., 2016). With the diverse evolution of deep learning models, numerous advanced methods have now been employed in the fusion of multimodal data. In the realm of sentiment analysis, Zadeh et al. (2017); Liu et al. (2018) employ a low-rank Tensor Fusion method that leverages both language and video content. Attention-gating mechanisms are used by Rahman et al. (2020); Yang & Wu (2021) to generate displacement vectors through cross-modal self-attention, which are then added to the input vectors from the primary modality. Tsai et al. (2019) takes an alternative approach by integrating multiple layers of cross-modal attention blocks in a word-level vision/language/audio alignment task.

In the context of clinical prediction, Khadanga et al. (2019); Deznabi et al. (2021) adopt a late fusion approach to combining vital sign and text data by concatenating embeddings from pre-trained feature extractors. Soenksen et al. (2022) developed a generalizable data preprocessing and modeling pipeline for EHR encompassing four data modalities, albeit through a direct concatenation of existing feature embeddings for each modality followed by an XGBoost classifier (Chen & Guestrin, 2016). Recently, Zhang et al. (2023) expanded on the work of Tsai et al. (2019) by introducing a discretized multi-time attention (mTAND) module (Shukla & Marlin, 2021) to encode temporal irregularities in time series and text data. Their fusion approach involves layering sets of self- and cross-modal attention blocks. However, this approach is limited to just two modalities and is not easily extendable to include additional modal components or handle missing modalities. To the best of our knowledge, existing works are tailored to application-specific settings that necessitate the computation of pairwise cross-modal relationships, which are not scalable to more general settings with arbitrary modalities. Moreover, these studies typically do not account for scenarios where modalities are missing, or rely on imputation

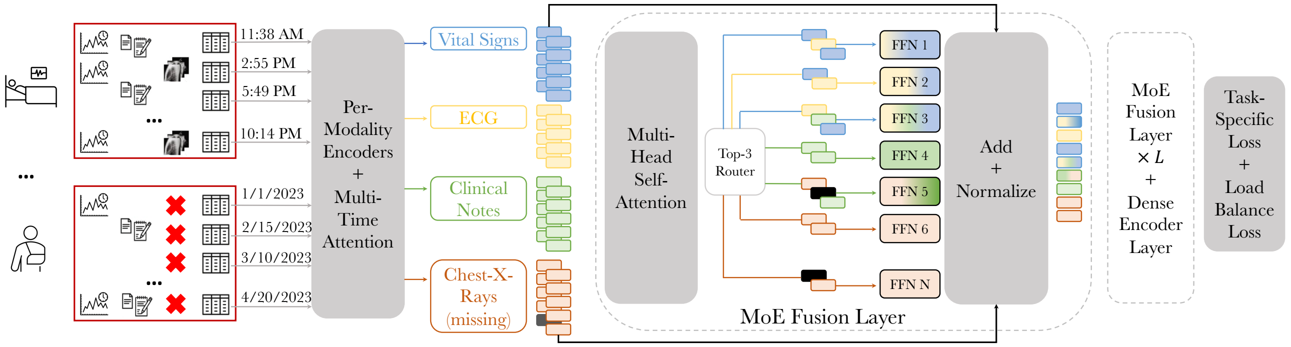


Figure 1. Addressing the challenge of FlexiModal Data in clinical scenarios: patients in ICUs often have extensive and highly irregular health status measurements over time; patients with milder conditions only require monitoring across fewer categories. `FuseMoE` is adept at handling EHRs featuring any combination of modalities, including those with missing elements. It starts by encoding EHRs using modality-specific feature extractors, followed by employing a multi-time attention mechanism (Shukla & Marlin, 2021) to address temporal irregularities. The core of `FuseMoE` lies the MoE Fusion Layer, where a routing mechanism is trained to categorize multimodal inputs and direct them to the appropriate combinations of MLPs, each tailored to process specific types of input. The outputs from these MLPs are weighted through a gating function, resulting in fused embeddings, which are subsequently utilized for further processing.

approaches based on observed data.

Mixture-of-Experts MoE (Jacobs et al., 1991; Xu et al., 1994) has gained significant popularity for managing complex tasks since its introduction three decades ago. Unlike traditional models that reuse the same parameters for all inputs, MoE selects distinct parameters for each specific input. This results in a sparsely activated layer, enabling a substantial scaling of model capacity without a corresponding increase in computational cost. Recent studies have demonstrated the effectiveness of integrating MoE with cutting-edge models across a diverse range of tasks (Shazeer et al., 2017; Fedus et al., 2022; Zhou et al., 2023). These works have also tackled key challenges such as accuracy and training instability (Nie et al., 2021; Zhou et al., 2022; Puigcerver et al., 2023). Given its ability to assign input partitions to specialized experts, MoE naturally lends itself to multimodal applications. This approach has been explored in fields such as vision-language modeling (Mustafa et al., 2022; Shen et al., 2023) and dynamic image fusion (Cao et al., 2023). However, the application of MoE in complex real-world settings, such as those involving FlexiModal Data, remains largely unexplored. This gap presents an opportunity to leverage MoE's potential in handling its intricate and multifaceted nature such as multimodal EHR, where reliable multimodal integration is crucial.

MoE Theory While MoE has been widely employed to scale up large models, its theoretical foundations have remained nascent. Recently, Nguyen et al. (2023b) provided convergence rates for both density and parameter estimation of Softmax gating Gaussian MoE. They connected these rates to the solvability of systems of polynomial equations under Voronoi-based loss functions. Later, Nguyen et al. (2024b) extended these theories to top-K sparse softmax gating MoE. Their theories further characterize the effect of the sparsity of gating functions on the behaviors of parameter estimation and verify the benefits of using top-1 sparse softmax gating MoE in practice. Other theoretical results include estimation rates of parameters and experts for multinomial logistic MoE (Nguyen et al., 2023a), for dense-to-sparse gating MoE (Nguyen et al., 2024a), and for input-independent gating MoE (Ho et al., 2022).

3. FuseMoE: Enhance Predictive Performance for FlexiModal Data

In this section, we delve into the fundamental components of `FuseMoE`, illustrated in Figure 1. We focus on two critical elements: the modality and irregularity encoder, and the MoE fusion layer. These components are pivotal in handling the unique characteristics of FlexiModal data.

3.1. Sparse MoE Backbone

The main components of a sparse MoE layer are a network G as a sparse gate and an expert network E. Shazeer et al. (2017) proposed a Top-K gating function that takes as an input a token representation $x \in \mathbb{R}^D$ and then routes it to the Top-K experts out of the set $\{E_i\}_{i=1}^{S}$. The gating network variable $W \in \mathbb{R}^{D \times N}$ produces logits $h_s(x) = \text{Top K}(W \cdot x)$, which are normalized via Softmax:

$$G(x)_i = \frac{\exp(h_s(x)_i)}{\sum_j^K \exp(h_s(x)_j)} \tag{1}$$

Each expert network ($E_i : \mathbb{R}^D \to \mathbb{R}^D$) contains a feed-forward layer (FFN) and its parameters are independent of other models. The final output of the expert network y is the linearly weighted combination of each expert's output on the token by the gate's output: $y = \sum_{i=1}^{S} G(x)_i E_i(x)$.

Gating Function Softmax gating, commonly used as a gating function in various tasks, is contrasted with Gaussian gated MoE, a popular alternative noted for its distinct advantages. Initially introduced by Xu et al. (1994), Gaussian gating is recognized for its superior qualities in certain scenarios, as compared to Softmax gating. The nonlinearity of the Softmax function complicates the application of the expectation-maximization (EM) algorithm. In contrast, Gaussian gating overcomes this difficulty with an analytical solution during the M-step. Additionally, it offers enhanced localization properties, crucial for minimizing interference among experts, a benefit that becomes increasingly significant as the number of experts in the model grows (Ramamurti & Ghosh, 1998). The logits of the Gaussian gating function are formulated as follows:

$$h_g(x) = \text{Top K}(-\|W - x\|_2^2). \tag{2}$$

The incorporation of Gaussian gating facilitates the convergence of the EM algorithm. Building upon this, our paper introduces an innovative Laplace gating function. This new function's logit is formulated as follows:

$$h_l(x) = \text{Top K}(-\|W - x\|_2). \tag{3}$$

The Laplace gating function, characterized by its Euclidean term $\exp(-\|W - x\|_2)$, is less prone to converge towards extreme weight distributions due to the bounded nature of this term. In subsequent sections, we will illustrate how this gating function facilitates faster parameter estimation rates compared to Gaussian and Softmax gating. Moreover, our empirical findings indicate that the Laplace gating exhibits enhanced performance in managing multimodal data.

3.2. Modality and Irregularity Encoder

To encode the irregularity of sampling in each modality, we utilize a discretized multi-time attention (mTAND) module (Shukla & Marlin, 2021), which leverages a time attention mechanism (Kazemi et al., 2019; Vaswani et al., 2017) to discretize irregularly sampled observations into discrete intervals. Specifically, given a set of l_k continuous time points, $t \in \mathbb{R}^{l_k}$, corresponding to the k^{th} dimensionality of a given modality, we employ H embedding functions $\phi_h(\tau)$ to embed each $\tau_k \in t_k$ in a d_h dimensional space. The i^{th} dimension of the h^{th} embedding is defined as

$$\phi_h(\tau)[i] = \begin{cases} w_i \tau_k, & \text{if } i = 1 \\ \sin(w_i \tau_k + \phi_i), & \text{if } 1 < i \leq d_h, \end{cases}$$

where $\{w_i, \phi_i\}_{i=1}^{d_h}$ are learnable parameters. By performing this for each continuous time point in t_k, we create a d_h dimensional representation of each time point in H different embedding spaces. We then leverage these embeddings to discretize the irregularly sampled observations into discretized bins. Specifically, we seek to discretize x_k (with l_k corresponding observation times t_k) into γ regularly sampled intervals $\boldsymbol{\gamma}$. We do this via an attention mechanism, which, for each embedding function $\phi_h(\tau)$, takes $\boldsymbol{\gamma}$ as queries, t_k as keys, and x_k as values and produces $\hat{x}_{k,h} \in \mathbb{R}^{\gamma}$ embeddings for each sequence. Formally,

$$\hat{x}_{k,h} = \text{softmax}\left(\frac{\phi_h(\boldsymbol{\gamma})\mathbf{Q}_h\mathbf{K}_h^\top \phi_h(t_k)^\top}{\sqrt{l_k}}\right) x_k,$$

where $\mathbf{Q}_h$ and $\mathbf{K}_h$ are learnable parameters. This formulation allows us to discretize univariate observations x_k into γ regularly-sampled bins. To model irregularity across a multivariate set of observations for a given modality with d_m dimensions, we repeat this process for each dimension of the input. This allows us to obtain an interpolation matrix $\hat{X}_h = [\hat{x}_{1,h}, \hat{x}_{2,h}, ..., \hat{x}_{d_m,h}] \in \mathbb{R}^{\gamma \times d_m}$ for each of the h embedding functions. We then concatenate the interpolation matrices across all H embedding functions (i.e., $I = [\hat{X}_1, \hat{X}_2, ..., \hat{X}_h] \in \mathbb{R}^{\gamma \times (H \cdot d_m)}$) and employ a linear projection to achieve a final, discretized embedding for each modality, $Z \in \mathbb{R}^{\gamma \times d_e}$, where d_e denotes the desired dimensionality of each modality's representation. Appendix C discusses further details of the embedding.

Encoding Multiple Modalities The process described above allows us to discretize an arbitrarily long irregular, multivariate sequence into a regularly sampled, discretized embedding with length γ and dimensionality d_e. We repeat this for each of the M modalities, to create M embeddings, $\{Z_j\}_{j=1}^{M}$, which are then combined to generate predictions.

3.3. MoE Fusion Layer

Router Design Study Upon obtaining embeddings from each of the j modalities, we propose multiple complementary approaches for processing multimodal inputs. Figure 2 illustrates a range of router design options. The most straightforward strategy involves employing a common router that handles the concatenated embeddings of all j modalities, without imposing any gating constraints. As the complexity increases with additional modalities, we consider more sophisticated alternatives: deploying separate routers for each modality's embedding and assigning these embeddings to a shared pool of experts. This allows for distinct processing while maintaining a unified expert framework. Additionally, we further segregate these common expert pools, allowing each router to direct its respective embedding to dedicated experts skilled in handling such specific inputs. These varied router design choices offer users enhanced flexibility, enabling more fine-grained control of both inter-modal and intra-modal relationships.

We implement an entropy regularization loss to ensure balanced and stable expert utilization, a concept supported by various previous studies (Mustafa et al., 2022; Meis-

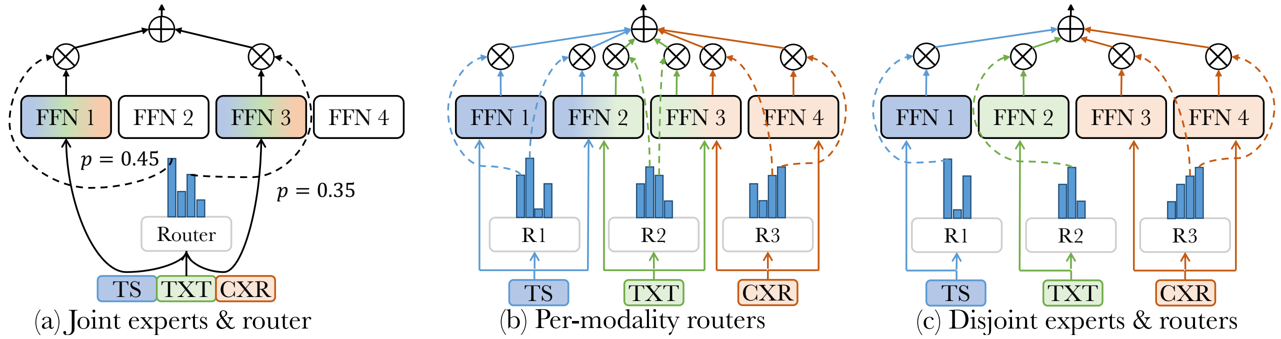


Figure 2. We present three exemplary designs of the Tok-K router for effective multimodal fusion, considering an input scenario with three modalities: Time-Series (TS), Text (TXT), and Chest X-Rays (CXR). (a) The joint router design utilizes a concatenated embedding of all modalities, directing this combined input to selected experts. (b) In the modality-specific router design, each modality's embedding is independently assigned to a shared pool of experts. (c) The third design variant also uses modality-specific routers but assigns each modality's embedding to separate pools of experts, each pool uniquely tailored to process a specific modality type.

ter et al., 2020; Genevay, 2019). It maximizes the mutual information between modalities and experts and serves as an auxiliary loss function in addition to task-specific loss. Given a total of M modalities, and denoting $\mathcal{H}$ as the entropy, we define the loss function $\mathcal{E}$ as follows:

$$\mathcal{E}(x) = \frac{1}{M}\sum_{j=1}^{M}\mathcal{H}(\hat{p}_{m_j}(E)) - \mathcal{H}(\frac{1}{M}\sum_{j=1}^{M}\hat{p}_{m_j}(E)), \quad (4)$$

where $\hat{p}_{m_j}(E)$ is the probability distribution over the experts $\{E_i\}_{i=1}^{S}$ for the j^{th} modality. This distribution can be approximated by $\hat{p}_{m_j}(E) = \frac{1}{l^j}\sum_{i=1}^{l^j} p_{m_j}(E \mid x_i^{m_j})$, where l^j is the number of observations of the j^{th} modality. Intuitively, we actively encourage the input embeddings to diminish the uncertainty in selecting experts. By incorporating the loss $\mathcal{E}$, we aim to stabilize the experts' preferences within each modality, while concurrently promoting a diverse range of expert selections across modalities.

Missing Modalities In scenarios where certain modalities are missing throughout the data trajectories, we substitute the original embedding Z_{missing} with a learnable embedding $\mathcal{Z}$, acting as a "missing indicator". This strategy is facilitated by employing per-modality routers, which, in conjunction with entropy regularization, guide $\mathcal{Z}$ predominantly toward a specific group of less-utilized experts. The new embeddings $\mathcal{Z}$ are dynamically adjusted throughout the model training process to simultaneously minimize the task-specific loss and the entropy regularization loss. As a result, the router will assign lower weights to the experts responsible for processing these embeddings.

4. Theoretical Justification

In this section, we provide a theoretical guarantee of the benefits of the Laplace gating function over the standard softmax gating function in MoE. In particular, we conduct a convergence analysis for maximum likelihood estimation (MLE) under the Laplace gating Gaussian MoE and prove that the MLE in Laplace gating MoE has better convergence behaviors than that in softmax gating MoE.

Notations We denote $[n] := \{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$. Additionally, the notation $|S|$ indicates the cardinality of a given set S. For any vector $v \in \mathbb{R}^d$, $\|v\|$ stands for its 2-norm value. Finally, for any two probability densities p, q dominated by the Lebesgue measure μ, we denote $V(p, q) = \frac{1}{2}\int |p - q| d\mu$ as their Total Variation distance.

Problem Setup Assume that $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ are i.i.d. samples drawn from the Laplace gating Gaussian MoE of order k_* whose conditional density function $p_{G_*}(Y|X)$ is given by:

$$p_{G_*}(Y|X) = \sum_{i=1}^{k_*}\text{softmax}(\text{TopK}(-\|W_i^* - X\|, K; \beta_i^*)) \times f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*), \quad (5)$$

where $f(\cdot|\mu, \sigma)$ denotes a univariate Gaussian density function with mean μ and variance σ. For the simplicity of the presentation, we denote $G_* := \sum_{i=1}^{k_*}\exp(\beta_i^*)\delta_{(W_i^*, a_i^*, b_i^*, \nu_i^*)}$ as a true but unknown *mixing measure* associated with true parameters $(\beta_i^*, W_i^*, a_i^*, b_i^*, \nu_i^*)$ for $i \in \{1, 2, \ldots, k_*\}$. Here, δ denotes the Dirac delta measure. In the paper, we specifically consider two settings of the true number of experts k_*: (i) *Exact-specified setting*: when k_* is known; (ii) *Over-specified setting*: when k_* is unknown, and we over-specify the model in equation 5 by a Laplace gating MoE model with $k > k_*$ experts. However, due to the space limit, we present only the latter setting, and defer the former setting to Appendix G.

Maximum Likelihood Estimation We use the maximum likelihood method to estimate the unknown mixing measure G_*. Notably, under the over-specified setting, as the

Table 2. Summary of parameter estimation rates under the softmax and Laplace gating Gaussian MoE models. In this table, the function $\widetilde{r}(\cdot)$ represents for the solvability of a system of polynomial equations considered in (Nguyen et al., 2023b) with a note that $\widetilde{r}(\cdot) \leq \bar{r}(\cdot)$ and $\widetilde{r}(2) = 4, \widetilde{r}(3) = 6$. Additionally, $\mathcal{A}_j^n := \mathcal{A}_j(\widehat{G}_n)$ denotes a Voronoi cell defined in equation 7.

Gates	$\exp(\beta_j^*)$	W_j^*	a_j^*	b_j^*	ν_j^*
Softmax (Nguyen et al., 2023b)	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/2\widetilde{r}(\lvert\mathcal{A}_j^n\rvert)})$	$\mathcal{O}(n^{-1/\widetilde{r}(\lvert\mathcal{A}_j^n\rvert)})$	$\mathcal{O}(n^{-1/2\widetilde{r}(\lvert\mathcal{A}_j^n\rvert)})$	$\mathcal{O}(n^{-1/\widetilde{r}(\lvert\mathcal{A}_j^n\rvert)})$
Laplace (Ours)	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/4})$	$\mathcal{O}(n^{-1/4})$	$\mathcal{O}(n^{-1/2\bar{r}(\lvert\mathcal{A}_j^n\rvert)})$	$\mathcal{O}(n^{-1/\bar{r}(\lvert\mathcal{A}_j^n\rvert)})$

true density $p_{G_*}(Y|X)$ is associated with top K experts which are possibly approximated by more than K fitted experts, we need to select $\overline{K} > K$ experts in the formulation of density estimation to guarantee its convergence to $p_{G_*}(Y|X)$. In particular, the MLE is given by

$$\widehat{G}_n \in \arg\max_{G \in \mathcal{G}_k(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \log(\overline{p}_G(Y_i|X_i)), \tag{6}$$

where $\mathcal{G}_k(\Theta) := \{G = \sum_{i=1}^{k} \exp(\beta_i)\delta_{(W_i,a_i,b_i,\nu_i)} : (W_i, a_i, b_i, \nu_i) \in \Theta\}$ denotes the set of all mixing measures with at most k components and

$$\overline{p}_G(Y|X) = \sum_{i=1}^{k_*} \mathrm{softmax}(\mathrm{TopK}(-\|W_i - X\|, \overline{K}; \beta_{0i})) \\ \times f(Y|(a_i)^\top X + b_i, \nu_i).$$

Given the MLE defined in equation 6, we first demonstrate that the conditional density function $p_{\widehat{G}_n}$ also converges to its true counterpart p_{G_*} at the parametric rate.

Theorem 4.1. *The density estimation $\overline{p}_{\widehat{G}_n}(Y|X)$ converges to the true density $p_{G_*}(Y|X)$ under the Total Variation distance V at the following rate:*

$$\mathbb{E}_X[V(\overline{p}_{\widehat{G}_n}(\cdot|X), p_{G_*}(\cdot|X))] = \mathcal{O}(\sqrt{\log(n)/n}).$$

Proof of Theorem 4.1 is in Appendix H.3. The parametric rate of the conditional density function $p_{\widehat{G}_n}$ indicates that we only need to determine a loss function between the MLE and the true mixing measure to lower bound the total variation distance between $p_{\widehat{G}_n}$ and p_{G_*}.

Voronoi Loss We now define a loss function between the MLE and the true mixing measure. Given some mixing measure $G := \sum_{i=1}^{k} \exp(\beta_i)\delta_{(W_i,a_i,b_i,\nu_i)}$, we distribute its components to the following Voronoi cells $\mathcal{A}_j \equiv \mathcal{A}_j(G)$ which are generated by the support of the true mixing measure $G_* := \sum_{j=1}^{k_*} \exp(\beta_j^*)\delta_{(W_j^*,a_j^*,b_j^*,\nu_j^*)}$:

$$\mathcal{A}_j := \{i \in [k] : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_\ell^*\|, \forall \ell \neq j\}, \tag{7}$$

where $\theta_i := (W_i, a_i, b_i, \nu_i)$ and $\theta_j^* := (W_j^*, a_j^*, b_j^*, \nu_j^*)$ for any $1 \leq i \leq k$ and $1 \leq j \leq k_*$. Based on these Voronoi cells, we propose the following Voronoi loss function for the over-specified setting:

$$\mathcal{D}_2(G, G_*) := \max \Bigg[\Big| \sum_{i \in \mathcal{A}_{\tau_j}} \exp(\beta_i) - \exp(\beta_{\tau_j}^*) \Big| \\ + \sum_{j \in [K]: |\mathcal{A}_{\tau_j}| > 1} \sum_{i \in \mathcal{A}_{\tau_j}} \exp(\beta_i) \Phi_{i\tau_j}\Big(2, 2, \bar{r}(|\mathcal{A}_{\tau_j}|), \frac{\bar{r}(|\mathcal{A}_{\tau_j}|)}{2}\Big) \\ + \sum_{j \in [K]: |\mathcal{A}_{\tau_j}| = 1} \sum_{i \in \mathcal{A}_{\tau_j}} \exp(\beta_i) \Phi_{i\tau_j}(1, 1, 1, 1) \Bigg], \tag{8}$$

where the maximum in the definition of Voronoi loss function $\mathcal{D}_2$ is for $\{\tau_1, \tau_2, \ldots, \tau_K\} \subseteq [k_*]$. Furthermore, for any $(\rho_1, \rho_2, \rho_3, \rho_4) \in \mathbb{R}^4$, we define $\Phi_{i\tau_j}(\rho_1, \rho_2, \rho_3, \rho_4) = \|W_i - W_{\tau_j}^*\|^{\rho_1} + \|a_i - a_{\tau_j}^*\|^{\rho_2} + |b_i - b_{\tau_j}^*|^{\rho_3} + |\nu_i - \nu_{\tau_j}^*|^{\rho_4}$ for any $i \in \mathcal{A}_{\tau_j}$ and $j \in [K]$. Additionally, the notation $\bar{r}(|\mathcal{A}_{\tau_j}|)$ stands for the minimum value of $r \in \mathbb{N}$ such that the following system of polynomial equations:

$$\sum_{i=1}^{|\mathcal{A}_{\tau_j}|} \sum_{\substack{m_1 + 2m_2 = s, \\ 1 \leq m_1 + m_2 \leq r}} \frac{q_{3i}^2 q_{1i}^{m_1} q_{2i}^{m_2}}{m_1! m_2!} = 0, \text{ for each } s = 1, 2, \ldots, r, \tag{9}$$

does not have any non-trivial solutions for the unknown variables $\{(q_{1i}, q_{2i}, q_{3i})\}_{i=1}^{|\mathcal{A}_{\tau_j}|}$. A solution to the above system is regarded as non-trivial if at least among variables q_{1i} is different from zero, whereas all the variables q_{3i} are non-zero. It is worth noting that the function $\bar{r}(\cdot)$ was previously studied in Ho & Nguyen (2016) to characterize the convergence behavior of parameter estimation under the location-scale Gaussian mixture models. Ho & Nguyen (2016) also gave some specific values of that function, namely $\bar{r}(2) = 4$ and $\bar{r}(3) = 6$. Meanwhile, they claimed that it was non-trivial to determine the value of $\bar{r}(m)$ when $m \geq 4$, and further techniques should be developed for that purpose. Since Gaussian MoE models are generalization of the Gaussian mixture models, we also involve the function $\bar{r}(\cdot)$ in our convergence analysis.

Theorem 4.2. *When $k > k_*$ becomes unknown, the following Total Variation bound holds true for any $G \in \mathcal{G}_k(\Theta)$:*

$$\mathbb{E}_X[V(\overline{p}_G(\cdot|X), p_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_2(G, G_*).$$

Consequently, we obtain $\mathcal{D}_2(\widehat{G}_n, G_*) = \mathcal{O}(\sqrt{\log(n)/n})$.

Table 3. Comparison of FuseMoE-based methods (gray) and baselines, utilizing vital signs and clinical notes of MIMIC-IV. The best results are highlighted in bold font, and the second-best results are underlined. All results are averaged across 3 random experiments.

Task \ Method		MISTS	MulT	MAG	TF	HAIM	Softmax	Gaussian	Laplace
48-IHM	AUROC	75.06 ± 1.03	75.95 ± 0.84	75.82 ± 0.73	78.76 ± 0.79	79.65 ± 0.00	79.49 ± 0.83	80.76 ± 0.56	**81.03** ± **0.25**
	F1	45.61 ± 0.34	38.81 ± 0.22	42.55 ± 0.82	40.61 ± 0.41	39.79 ± 0.00	42.86 ± 0.44	**46.86** ± **0.24**	46.53 ± 0.57
LOS	AUROC	80.56 ± 0.33	81.36 ± 1.32	81.13 ± 0.66	80.71 ± 0.45	82.58 ± 0.00	82.11 ± 0.39	81.92 ± 0.73	**82.91** ± **1.02**
	F1	73.01 ± 0.52	73.45 ± 0.59	72.51 ± 0.27	73.84 ± 0.61	73.18 ± 0.00	74.43 ± 0.88	74.46 ± 0.52	**74.58** ± **0.63**
25-PHE	AUROC	69.45 ± 0.72	66.58 ± 0.41	69.55 ± 0.67	69.18 ± 0.32	63.39 ± 0.00	70.54 ± 0.47	70.42 ± 0.26	**71.23** ± **0.53**
	F1	28.59 ± 0.46	28.55 ± 0.31	27.86 ± 0.29	28.52 ± 0.22	**42.13** ± **0.00**	31.25 ± 0.18	30.44 ± 0.27	31.33 ± 0.19

Table 4. Incorporating CXR and ECG into FuseMoE leads to a noticeable enhancement as compared to their two-modality counterparts.

Task \ Method		Vital & Notes & CXR				Vital & Notes & CXR & ECG			
		HAIM	Softmax	Gaussian	Laplace	HAIM	Softmax	Gaussian	Laplace
48-IHM	AUROC	78.87 ± 0.00	83.13 ± 0.36	83.64 ± 0.47	**83.87** ± **0.33**	78.87 ± 0.00	82.92 ± 0.22	83.03 ± 0.85	**83.55** ± **0.49**
	F1	39.78 ± 0.00	**46.82** ± **0.28**	38.87 ± 0.26	45.36 ± 0.46	39.78 ± 0.00	46.87 ± 0.17	44.04 ± 0.26	**46.88** ± **0.42**
LOS	AUROC	82.46 ± 0.00	**83.76** ± **0.59**	83.64 ± 0.52	83.51 ± 0.51	82.46 ± 0.00	83.53 ± 0.34	83.47 ± 0.37	**83.58** ± **0.78**
	F1	72.75 ± 0.00	74.32 ± 0.44	**76.59** ± **0.74**	75.18 ± 0.77	72.75 ± 0.00	75.01 ± 0.63	74.43 ± 0.64	**75.11** ± **0.65**
25-PHE	AUROC	63.57 ± 0.00	**73.87** ± **0.71**	72.68 ± 0.61	73.65 ± 0.39	63.82 ± 0.00	73.64 ± 0.89	**73.74** ± **0.41**	73.67 ± 0.71
	F1	**42.80** ± **0.00**	35.96 ± 0.23	35.09 ± 0.15	36.01 ± 0.42	**43.20** ± **0.00**	36.06 ± 0.17	36.46 ± 0.55	35.81 ± 0.34

Proof of Theorem 4.2 is in Appendix H.4. The results of Theorem 4.2 together with the formulation of the loss function $\mathcal{D}_2$ in equation 8 reveal that (see also Table 2):

(i) Parameters $W_i^*, a_i^*, b_i^*, \nu_i^*$ which are fitted by exactly one component, i.e. $|\mathcal{A}_i^n| := |\mathcal{A}_i(\widehat{G}_n)| = 1$, enjoy the same estimation rate of order $\mathcal{O}(n^{-1/2})$ (up to some logarithmic factor), which match those in Nguyen et al. (2023b).

(ii) The rates for estimating parameters $W_i^*, a_i^*, b_i^*, \nu_i^*$ which are fitted by more than one component, i.e. $|\mathcal{A}_i^n| > 1$, are no longer homogeneous. On the one hand, the estimation rates for parameters b_i^* and ν_i^* are of orders $\mathcal{O}(n^{-1/2\bar{r}(|\mathcal{A}_i^n|)})$ and $\mathcal{O}(n^{-1/\bar{r}(|\mathcal{A}_i^n|)})$, respectively, both of which are determined by the function $\bar{r}(\cdot)$ and vary with the number of fitted components $|\mathcal{A}_i^n|$. Those rates are comparable to their counterparts in Nguyen et al. (2023b). On the other hand, gating parameters W_i^* and expert parameters a_i^* share the same estimation rate of order $\mathcal{O}(n^{-1/4})$, which remains constant with respect to the number of fitted components. Meanwhile, those rates in Nguyen et al. (2023b) depend on the solvability of a different system of polynomial equations from that in equation 9, which are significantly slower.

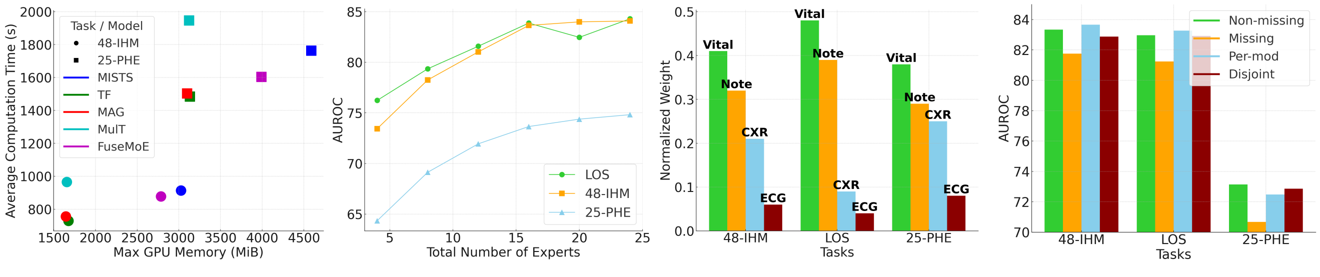
5. Experiments

In this section, we demonstrate empirically that FuseMoE is capable of providing accurate and efficient predictions when applied to FlexiModal datasets. The broad applicability of FlexiModal data means that real-world applications are plentiful. We note that ICU data are particularly rich in this context, featuring a diverse range of measurements across multiple modalities, motivating our focus on three tasks highly relevant to critical care settings (Keuning et al., 2020). The urgency of care in the ICU necessitates swift and accurate assessments of patient conditions to guide decisions, but current methods fail to incorporate information beyond acute physiology (Awad et al., 2017a).

Experimental Setup We leveraged data from MIMIC-IV (Johnson et al., 2020) and its predecessor, MIMIC-III (Johnson et al., 2016a). Our tasks of interest include the 48-hour in-hospital mortality prediction (48-IHM), 25-type phenotype classification (25-PHE), and length-of-stay (LOS) prediction. The baselines include HAIM (Soenksen et al., 2022), multimodal fusion via cross-attention (MulT) (Tsai et al., 2019), cross-attention combined with irregular sequences (MISTS) (Zhang et al., 2023), tensor fusion (TF) (Zadeh et al., 2017), and multimodal adaptation gate (MAG) (Rahman et al., 2020). All details of the experimental setup can be found in Appendices A - E.

5.1. Primary Results

Table 3 shows the outcomes of combining irregular vital signs and clinical notes from the MIMIC-IV dataset. The FuseMoE-based methods surpass baselines in most scenarios, often by a non-trivial margin. The Laplace gating function also outperforms its counterparts, aligning with our theoretical claims. Furthermore, we observe that HAIM shows considerable efficacy in extracting features from time series, resulting in a strong performance in the 48-IHM and LOS tasks, which are heavily reliant on such data. However, its performance appears more moderate on the 25-PHE task. Note that, the results derive from the MIMIC-IV dataset without missing modalities. We used the "joint experts and router" configuration, given that alterations to the router design did not significantly impact performance in this context. Overall, the results have shown the efficacy of integrating the irregularity encoder with the MoE fusion layer. Additional results on (1) the MIMIC-III data and (2) combining vital signs with CXR



(a) (b) (c) (d)

Figure 3. Results of ablation studies: (a) The computational efficiency and resource utilization of each method when applied to vital signs and clinical notes from the MIMIC-IV dataset; (b) The relationship between the number of experts and task performance across different modalities, including vital signs, clinical notes, and CXR; (c) The impact of each modality on the top-k experts within a disjoint router structure; (d) The role of per-modality routers and the entropy loss $\mathcal{E}$ in mitigating the impact of missing modalities.

can be found in Appendix F.

Additional Modalities We then expand the FuseMoE framework to incorporate additional modalities. Note that, except for HAIM, the baselines were not designed to be agnostic to the quantity and variety of input modalities. Therefore, adapting them to manage extra and missing modalities requires considerable model changes, which might compromise their performance. Table 4 presents the revised outcomes after integrating extra modalities, employing the per-modality router and the entropy loss $\mathcal{E}$ within FuseMoE. This setup was chosen as it slightly outperformed the joint router with an increase in modalities. Relative to their two-modality versions, FuseMoE has effectively harnessed additional information (notably from CXR), resulting in a significant enhancement in performance. Conversely, the addition of new modalities did not benefit the HAIM method, possibly due to its reliance on vital signs and clinical notes without adequately addressing the dynamics between different modalities. Furthermore, HAIM's notably high F1 scores on the 25-PHE task can be attributed to XGBoost's proficiency in managing missing minority classes.

5.2. Ablation Studies

We then conducted ablation studies to explore the various attributes of FuseMoE and baselines. Figure 3(a) examines the computational efficiency and resource utilization, positioning FuseMoE approximately in the middle of the comparison. Despite the increase in model parameters due to the incorporation of the MoE layer, its sparse nature does not significantly escalate the computational load. Figure 3(b) illustrates the correlation between the number of experts and task performance across different modalities. Generally, performance improves with the addition of more experts, plateauing once the count exceeds 16. To achieve a compromise between performance and computational expense, we opted to utilize the top 4 experts out of 16 in our experiments. Figure 3(c) studies the influence of each modality on the top-k chosen experts. For every expert selected, we calculate the number of samples that include a specific modality, weighted by corresponding weight factors from the gating functions. The outcomes are subsequently normalized across modalities. The analysis reveals that predictions across all tasks heavily depend on vital signs and clinical notes. This reliance is attributed to the abundant samples in these two modalities. Despite the notably smaller quantity of CXR, they play more significant roles in the 25-PHE and 48-IHM tasks, which aligns with our findings in Table 4. Lastly, Figure 3(d) illustrates the effectiveness of utilizing per-modality routers and the entropy loss $\mathcal{E}$ in addressing missing modalities. Initially, we compare the performance of FuseMoE on patients with fully available modalities against those with missing components, employing a joint router mechanism with the importance loss function (Shazeer et al., 2017), to ensure load balancing. The inclusion of datasets with missing modalities, while expanding the sample size, resulted in a decrease in overall performance due to the compromised data quality. However, an enhancement in performance was observed upon integrating per-modality or disjoint routers with $\mathcal{E}$. Notably, the outcomes for the 48-IHM and LOS tasks with missing modalities surpassed those obtained from datasets without any missing data. This is because the per-modality structure can better separate the present and missing modalities, reducing the influence of experts responsible for processing the absent inputs. Therefore, this leads to a more efficient exploitation of a broader array of samples.

6. Conclusion and Future Works

In this paper, we introduced FuseMoE, a model adept at managing multimodal data characterized by random missingness or irregularity—a crucial yet relatively unexplored challenge. FuseMoE integrates MoE fusion layers with modal embeddings and offers multiple router configurations to adeptly handle multimodal inputs across different complexity levels. FuseMoE also employs an innovative

Laplace gating function, which provides better theoretical results. Through empirical evaluation, FuseMoE has demonstrated superior performance across diverse scenarios. We will focus on identifying more effective modality encoders that seamlessly integrate with the MoE fusion layer and extend the application of FuseMoE to other critical domains.

Impact Statements This paper presents research aimed at propelling advancements in the broad domain of machine learning. The implications of our findings are wide-ranging, with potential applications in sectors including healthcare, autonomous driving, and recommendation systems. Based on our current understanding, this research does not warrant an ethics review, and a detailed discussion of the potential societal impacts is not required at the current stage.

References

Agarwal, V., Podchiyska, T., Banda, J. M., Goel, V., Leung, T. I., Minty, E. P., Sweeney, T. E., Gyang, E., and Shah, N. H. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, 23(6):1166–1173, 2016.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Arbabi, A., Adams, D. R., Fidler, S., Brudno, M., et al. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, 7 (2):e12596, 2019.

Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., Pellikka, P. A., Enriquez-Sarano, M., Noseworthy, P. A., Munger, T. M., et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature medicine*, 25(1):70–74, 2019.

Awad, A., Bader-El-Den, M., and McNicholas, J. Patient length of stay and mortality prediction: a survey. *Health services management research*, 30(2):105–120, 2017a.

Awad, A., Bader-El-Den, M., McNicholas, J., and Briggs, J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International journal of medical informatics*, 108:185–195, 2017b.

Bertsimas, D., Pauphilet, J., Stevens, J., and Tandon, M. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, 24(6):2809–2824, 2022.

Bucak, S. S., Jin, R., and Jain, A. K. Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1354–1369, 2013.

Butler, R. R. Icd-10 general equivalence mappings: Bridging the translation gap from icd-9. *Journal of AHIMA*, 78(9):84–86, 2007.

Cao, B., Sun, Y., Zhu, P., and Hu, Q. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23555–23564, 2023.

Chen, J., Chen, Z., Chi, Z., and Fu, H. Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 508–513, 2014.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pp. 231–249. PMLR, 2022.

Deznabi, I., Iyyer, M., and Fiterau, M. Predicting in-hospital mortality by combining clinical notes with time-series data. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pp. 4026–4031, 2021.

Elixhauser, A. Clinical classifications software (ccs) 2009. *http://www. hcug-us. ahrq. gov/toolssoft-ware/ccs/ccs. jsp*, 2009.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., and Xu, W. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015.

Garg, A., Pavlovic, V., and Rehg, J. M. Boosted learning in dynamic bayesian networks for multimodal speaker detection. *Proceedings of the IEEE*, 91(9):1355–1369, 2003.

Genevay, A. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE), 2019.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Berkowitz, S., Moukheiber, D., Eslami, P., et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. 2022.

Han, W., Chen, H., Gelbukh, A., Zadeh, A., Morency, L.-p., and Poria, S. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 6–15, 2021a.

Han, W., Chen, H., and Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021b.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1): 96, 2019.

Ho, N. and Nguyen, X. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016.

Ho, N., Yang, C.-Y., and Jordan, M. I. Convergence rates for Gaussian mixtures of experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022.

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3 (1):136, 2020a.

Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I., and Lungren, M. P. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, 10(1):22147, 2020b.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., and Horng, S. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 2019a.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

Johnson, A., Pollard, T., Horng, S., Celi, L. A., and Mark, R. Mimic-iv-note: Deidentified free-text clinical notes, 2023.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016a.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019b.

Johnson, J., Karpathy, A., and Fei-Fei, L. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4565–4574, 2016b.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., Wu, S., Smyth, C., Poupart, P., and Brubaker, M. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Keuning, B. E., Kaufmann, T., Wiersema, R., Granholm, A., Pettilä, V., Møller, M. H., Christiansen, C. F., Castela Forte, J., Snieder, H., Keus, F., et al. Mortality prediction models in the adult critically ill: A scoping review. *Acta Anaesthesiologica Scandinavica*, 64(4): 424–442, 2020.

Khadanga, S., Aggarwal, K., Joty, S., and Srivastava, J. Using clinical notes with time series data for icu management. *arXiv preprint arXiv:1909.09702*, 2019.

Lin, K., Hu, Y., and Kong, G. Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model. *International journal of medical informatics*, 125:55–61, 2019.

Liu, J., Capurro, D., Nguyen, A., and Verspoor, K. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of Biomedical Informatics*, 145:104466, 2023.

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., and Morency, L.-P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

Lovaasen, K. R. and Schwerdtfeger, J. *ICD-9-CM Coding: Theory and Practice with ICD-10, 2013/2014 Edition-E-Book*. Elsevier Health Sciences, 2012.

Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., and Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133, 2018.

Meister, C., Salesky, E., and Cotterell, R. Generalized entropy regularization or: There's nothing special about label smoothing. *arXiv preprint arXiv:2005.00820*, 2020.

Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., and Houlsby, N. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

Nefian, A. V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., and Murphy, K. A coupled hmm for audio-visual speech recognition. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. II–2013. IEEE, 2002.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

Nguyen, H., Akbarian, P., Nguyen, T., and Ho, N. A general theory for softmax gating multinomial logistic mixture of experts. *arXiv preprint arXiv:2310.14188*, 2023a.

Nguyen, H., Nguyen, T., and Ho, N. Demystifying softmax gating function in Gaussian mixture of experts. In *Advances in Neural Information Processing Systems*, 2023b.

Nguyen, H., Akbarian, P., and Ho, N. Is temperature sample efficient for softmax Gaussian mixture of experts? *arXiv preprint arXiv:2401.13875*, 2024a.

Nguyen, H., Akbarian, P., Yan, F., and Ho, N. Statistical perspective of top-k sparse softmax gating mixture of experts. In *International Conference on Learning Representations*, 2024b.

Nie, X., Miao, X., Cao, S., Ma, L., Liu, Q., Xue, J., Miao, Y., Liu, Y., Yang, Z., and Cui, B. Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate. *arXiv preprint arXiv:2112.14397*, 2021.

Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., and Morency, L.-P. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 284–288, 2016.

Poria, S., Cambria, E., and Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2539–2544, 2015.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. W. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91 (9):1306–1326, 2003.

Puigcerver, J., Riquelme, C., Mustafa, B., and Houlsby, N. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

Rahman, W., Hasan, M. K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., and Hoque, E. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, pp. 2359. NIH Public Access, 2020.

Ramamurti, V. and Ghosh, J. Use of localized gating in mixture of experts networks. In *Applications and Science of Computational Intelligence*, volume 3390, pp. 24–35. SPIE, 1998.

Reiter, S., Schuller, B., and Rigoll, G. Hidden conditional random fields for meeting segmentation. In *2007 IEEE International Conference on Multimedia and Expo*, pp. 639–642. IEEE, 2007.

Shaik, T., Tao, X., Li, L., Xie, H., and Velásquez, J. D. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, pp. 102040, 2023.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Shen, S., Yao, Z., Li, C., Darrell, T., Keutzer, K., and He, Y. Scaling vision-language models with sparse mixture of experts. *arXiv preprint arXiv:2303.07226*, 2023.

Shukla, S. N. and Marlin, B. M. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H. M., Li, M. L., Fuentes, I., and Bertsimas, D. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

Teicher, H. On the mixture of distributions. *Annals of Statistics*, 31:55–73, 1960.

Teicher, H. Identifiability of mixtures. *Annals of Statistics*, 32:244–248, 1961.

Teicher, H. Identifiability of finite mixtures. *Ann. Math. Statist.*, 32:1265–1269, 1963.

Tipirneni, S. and Reddy, C. K. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–17, 2022.

Tran, L., Liu, X., Zhou, J., and Jin, R. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1405–1414, 2017.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, pp. 6558. NIH Public Access, 2019.

van de Geer, S. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xu, L., Jordan, M., and Hinton, G. E. An alternative model for mixtures of experts. *Advances in neural information processing systems*, 7, 1994.

Yang, B. and Wu, L. How to leverage multimodal ehr data for better medical predictions? *arXiv preprint arXiv:2110.15763*, 2021.

Yang, B., Mei, T., Hua, X.-S., Yang, L., Yang, S.-Q., and Li, M. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 73–80, 2007.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

Zhan, X., Wu, Y., Dong, X., Wei, Y., Lu, M., Zhang, Y., Xu, H., and Liang, X. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11782–11791, 2021.

Zhang, X., Li, S., Chen, Z., Yan, X., and Petzold, L. R. Improving medical predictions by irregular multimodal electronic health records modeling. In *International Conference on Machine Learning*, pp. 41300–41313. PMLR, 2023.

Zhou, T., Ruan, S., and Canu, S. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004, 2019.

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A. M., Le, Q. V., Laudon, J., et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Zhou, Y., Du, N., Huang, Y., Peng, D., Lan, C., Huang, D., Shakeri, S., So, D., Dai, A. M., Lu, Y., et al. Brainformers: Trading simplicity for efficiency. In *International Conference on Machine Learning*, pp. 42531–42542. PMLR, 2023.

Appendix for
"FuseMoE: Mixture-of-Experts Transformers for Fleximodal Fusion"

In this appendix, we provide additional details on the datasets and tasks of interest (in Appendix A), data preprocessing pipelines (in Appendix B), irregularity encoder (in Appendix C), baseline methods (in Appendix D), computational resources and hyperparameters (in Appendix E), additional experimental results (in Appendix F), exact-specified setting of Laplace gating (in Appendix G), and proofs for all theoretical results (in Appendix H).

A. Tasks of Interest

In the ICU, where rapid and informed decisions are crucial, accurate mortality prediction is essential to provide clinicians with advanced warnings of patient deterioration, aiding in critical decision-making processes (Awad et al., 2017b). Similarly, the prediction of patient length-of-stay is indispensable for optimizing treatment plans, resource allocation, and discharge processes (Bertsimas et al., 2022). Further, phenotyping of critical care conditions is highly relevant to comorbidity detection and risk adjustment and presents a more challenging task than binary classification, due to the heterogeneous presentation of conditions and the larger number of prediction tasks (Zhang et al., 2023).

- **48-IHM** In this binary classification task, we predict in-hospital mortality based on the first 48 of the ICU stay for patients who stayed in the ICU for at least 48 hours.
- **LOS** We formulate our length-of-stay task similar to that of 48-IHM: for patients who spent at least 48 hours in the ICU, we predict ICU discharge without expiration within the following 48 hours.
- **25-PHE** In this multilabel classification problem, we attempt to predict one of 25 acute care conditions (Elixhauser, 2009; Lovaasen & Schwerdtfeger, 2012) (e.g., congestive heart failure, pneumonia, shock, etc.) at the *end* each each patient's ICU stay. Because the original task was designed for diagnoses based on ICD-9 codes, but MIMIC-IV includes both ICD-9 and ICD-10 codes, we map patients with diagnoses coded using ICD-10 using the conversion database provided by Butler (2007).

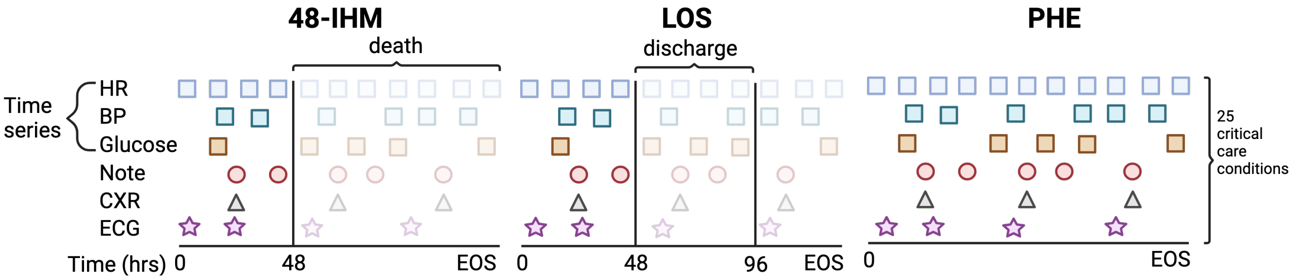


Figure 4. **Schematic of tasks of interest.** Plotted are example vitals/labs, radiological notes, X-rays, and ECGs sampled over the course of a patient's ICU stay. The first three rows represent example observations from a single modality consisting of three irregularly sampled vital signs (HR, BP), and lab values (Glucose). The following three rows represent irregularly sampled radiological notes, X-rays, and ECGs. Opaque shapes denote observations falling within the observation window (i.e., observations that are used to generate predictions), while translucent shapes are not used to generate predictions. For the **48-IHM** task, we use the first 48 hours of observations to predict death at any time during the ICU stay. For the **LOS** task, we use the first 48 hours of observations to predict whether the patient will be discharged (alive) during the following 48 hours. And in the phenotyping task (**PHE**), we use all observations to predict one of 25 critical care conditions.

We implement an in-hospital mortality prediction (**48-IHM**) task to evaluate our method's ability to predict short-term patient deterioration. Similarly, an accurate determination of patient discharge times is crucial for optimizing patient outcomes and hospital resource allocation (Bertsimas et al., 2022), which motivates our length-of-stay (**LOS**) task. We frame 48-IHM and LOS as binary classification problems and use a 48-hour observation window (for patients who spent at least 48 hours in the ICU) to predict in-hospital mortality (48-IHM) and discharge (without expiration) within the 48 hours following the observation window (LOS). Lastly, identifying the presence of specific acute care conditions in patient records is essential for various clinical objectives, including the construction of cohorts for clinical studies and the detection of comorbidities (Agarwal et al., 2016). Traditional methods, often reliant on manual chart reviews or simple billing code-based definitions, are increasingly being supplemented by machine learning techniques (Harutyunyan et al., 2019); automating this process requires high-fidelity classifications, motivating our 25-type phenotype classification (**25-PHE**) task. In this multilabel classification problem, we attempt to predict one of 25 acute care conditions using data from the entire ICU stay.

Evaluation In our initial analysis, we focused on patients with no missing modalities, resulting in a dataset comprised of 8,770 ICU stays for the 48-IHM and LOS tasks, and 14,541 stays for the 25-PHE task. For our analyses *with* missing observations, we include a total of 35,129 stays for 48-IHM and LOS, and 71.173 for 25-PHE. To evaluate the single-label tasks, 48-IHM and LOS, we employ the F1-score and AUROC as our primary metrics. In line with previous studies (Zhang et al., 2023; Lin et al., 2019; Arbabi et al., 2019), we use macro-averaged F1-score and AUROC to assess the 25-PHE task.

A.1. Dataset

We leveraged data from MIMIC-IV (Johnson et al., 2020), a comprehensive database with records from nearly $300k$ patients admitted to a medical center from 2008 to 2019, focusing on the subset of 73,181 ICU stays. We were able to link core ICU records (containing lab results and vital signs) to corresponding chest X-rays (Johnson et al., 2019b), radiological notes (Johnson et al., 2023), and electrocardiogram (ECG) data (Gow et al., 2022) taking place during a given ICU stay. We allocated 70 percent of the data for model training, with the remaining 30 percent evenly split between validation and testing.

A.2. Missingness rates

The total number of samples for each of our three tasks (i.e., those in which *at least one* vital sign was recorded in the specified observation window), along with the total number of observations per-modality, are shown in Table 5.

Table 5. We present the total number of ICU stays in each task, taking into account observations with missing modalities. The total number of stays with *at least one* observation of the corresponding modality are shown in the three right-most columns.

Task(s)	Total	Text	CXR	ECG
48-IHM & LOS	35,129	32,038	8,781	18,271
25-PHE	73,173	56,824	14,568	35,925

B. Data Preprocessing

In the preprocessing stage, we focused on 30 pertinent lab and chart events from each patient's ICU record for vital sign measurements. For chest X-rays, we utilized a pre-trained DenseNet-121 model (Cohen et al., 2022), which was fine-tuned on the CheXpert dataset (Irvin et al., 2019), to extract 1024-dimensional image embeddings. For radiological notes, we obtained 768-dimensional embeddings using the BioClinicalBERT model (Alsentzer et al., 2019). ECG signals were processed using a convolutional autoencoder, adapted from Attia et al. (2019), to generate a 256-dimensional embedding for each ECG.

Time series We selected 30 time series events for inclusion, following Soenksen et al. (2022). Nine of these were vital signs: heart rate, mean/systolic/diastolic blood pressure, respiratory rate, oxygen saturation, and Glascow Coma Scale (GCS) verbal, eye, and motor response. We also included 21 lab values: potassium, sodium, chloride, creatinine, urea nitrogram, bicarbonate, anion gap, hemoglobin, hematocrit, magnesium, platlet count, phosphate, white blood cell count, total calcium, MCH, red blood cell count, MCHC, MCV, RDW, platlet count, neutrophil count, and vancomycin. We standard scale each time series value to have mean 0 and standard deviation 1, based on the values in the training set.

Chest X-rays To incorporate a medical imaging modality into our analyses, we use the MIMIC-CXR-JPG (Johnson et al., 2019a) module available from Physionet (Goldberger et al., 2000), which includes 377,110 JPG format images derived from the DICOM-based MIMIC-CXR database (Johnson et al., 2019b). Following Soenksen et al. (2022), for each image, we resize each JPG image to 224 × 224 pixels and then extract embeddings from the last layer of the Densenet121 model. We identify X-rays taken while the patient was in the ICU by first matching subject IDs in MIMIC-CXR-JPG with the core MIMIC-IV database, then limiting these matched X-rays to those with a chart time occuring between an ICU admission and discharge.

Radiological notes To incorporate text data, we use the MIMIC-IV-Note module (Johnson et al., 2023), which contains 2,321,355 deidentified radiology reports for 237,427 patients that can be matched with patients in the main MIMIC-IV dataset via a similar approach to chest X-rays. We note that we were unable to obtain *intermediate* clinical notes (i.e., notes made by clinicians throughout a patient stay), as those have not yet been publicly released. We extract note embeddings using Bio-Clinical BERT (Alsentzer et al., 2019).

Electrocardiograms (ECGs) To include ECGs as an additional modality in our models, we utilize the MIMIC-IV-ECG (Gow et al., 2022) module, which includes approximately 800,000 ECGs (10 seconds, sampled at 500 Hz) collected from nearly 160,000 unique patients. To transform the ECGs so that they are suitable for input to our model, we adopt a convolutional autoencoder approach, adapted from Attia et al. (2019), that compresses each ECG into a 256-dimensional vector. Specifically, each diagnostic ECG contains a 5000×12 dimensional vector (5000 time points $\times$ 12 ECG leads). To prepare the ECG for input to the autoencoder, we only include the first 4096 time points. We then train the autoencoder to compress the ECG into a 256-dimensional latent vector, and then reconstruct the original ECG using upsampling layers, using mean squared error as our loss function. The architecture is shown in Figure 5.

We train the autoencoder with 90% of the ECGs available in the MIMIC-IV-ECG projection and use the rest for validation. We selected a batch size of 2048, and reduced the learning rate by a factor of 0.5 is the validation loss had plateaued for 3 epochs. Training stopped if the validation loss had not decreased for 6 epochs. For our encoder, we use filter numbers of $[16, 16, 32, 32, 64, 64]$, kernel widths of $[5, 5, 5, 3, 3, 3]$ and a dropout rate of 0.1. For the decoder, we use the same filter numbers and kernel widths in reverse, and maintain a dropout rate of 0.1.

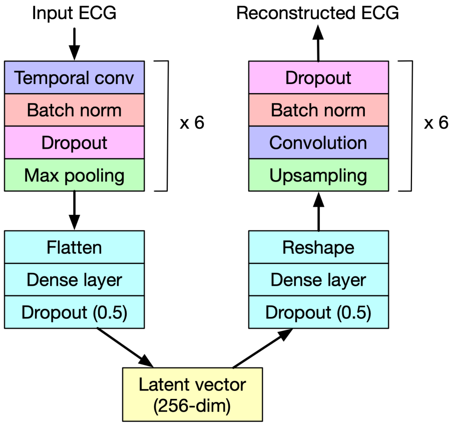


Figure 5. **CNN autoencoder architecture.** The encoder consists of six convolutional blocks (temporal convolution, batch normalization, dropout, and max pooling layers), followed by a dense layer that reduces the dimensionality of the representation of 256. The decoder then reconstructs the input ECG (dimensionality 4096×12) from this latent vector via a dense layer, followed by six upsampling convolutional blocks (upsampling, convolutional, batch normalization, and dropout layers).

C. Modeling Irregularity

C.1. Unified Temporal Discretization Embeddings

Unlike the embeddings in chest X-rays, clinical notes, and ECGs, vitals/lab values present temporal irregularity *across dimensions*. That is, for the former three modalities, each dimension of the corresponding is observed at each irregular time point $\boldsymbol{\tau}$. By contrast, the sampling for vitals/labs is irregular both *within* and *across* dimensions. For example, we might observe heart rate values sampled at times $\boldsymbol{\tau}_{\text{HR}} = \{0, 0.2, 0.8, 1.2, 2.8\}$ and glucose values sampled at time $\boldsymbol{\tau}_{\text{Glu}} = \{0.1, 0.7, 3.4\}$. Given this unique challenge present in vitals/labs, we adapt the Unified Temporal Discretization

Embedding (UTDE) approach described in Zhang et al. (2023), which combines the mTAND approach described in Section 3.2 with a simpler imputation-based discretization scheme. Specifically, given a set of t observations $\mathbf{x} \in \mathbb{R}^t$ observed at irregular times $\boldsymbol{\tau} \in \mathbb{R}^t$, we a simple imputation scheme to discretize $\mathbf{x}$ into target bins $\boldsymbol{\gamma}$ (e.g., $\boldsymbol{\gamma} = \{0, 1, 2, ..., \gamma\}$. Specifically, given bin value $\gamma_i \in \boldsymbol{\gamma}$, we apply the following rules:

- If there exists a previously observed value of $\mathbf{x}$ (i.e., $\exists \tau \in \boldsymbol{\tau}\ st.\ \tau \leq \gamma_i$), we set the imputed value of $\mathbf{x}$ at time γ, $\hat{x}_{\gamma_i}$, to the closest previously observed value.
- If no previously observed value exists, we set the value of $\hat{x}_{\gamma_i}$ to the global mean of $\mathbf{x}$.

We do this for each possible vitals/lab, to generate a matrix of imputation embeddings $\mathbf{I} \in \mathbb{R}^{\gamma \times d_{\text{vitals}}}$, were d_{vitals} is the number of vitals/labs. We then input this embedding into a 1D causal convolutional layer with stride 1 to obtain our final imputation embeddings with hidden dimension d_h, $\mathbf{E}_{\text{Imp}} \in \mathbb{R}^{\gamma \times d_h}$.

C.2. Unifying imputation and mTAND embeddings

We combined simple imputation and mTAND embeddings via a gating function $\mathbf{g}$. Following Zhang et al. (2023), we let $\mathbf{E}_{\text{mTAND}} \in \mathbb{R}^{\gamma \times d_h}$ denote the mTAND embeddings for vitals/labs derived from the process described in Section 3.2 and let $\mathbf{E}_{\text{Imp}} \in \mathbb{R}^{\gamma \times d_h}$ denote the simple imputations from the process described above. We use each of these discretization embeddings to derive a final set of embeddings for vitals/labs $\mathbf{E}_{\text{vitals}}$ via a one-layer MLP gating function f. Specifically, we let $\mathbf{g} = f(\mathbf{E}_{\text{Imp}} \oplus \mathbf{E}_{\text{mTAND}})$, where $\oplus$ denotes the concatenation operator. We then calculate $\mathbf{E}_{\text{vitals}}$ as

$$\mathbf{E}_{\text{vitals}} = \mathbf{g} \odot \mathbf{E}_{\text{Imp}} + (1 - \mathbf{g}) \odot \mathbf{E}_{\text{mTAND}} \in \mathbb{R}^{\gamma \times d_h},$$

where $\odot$ denotes point-wise multiplication.

D. Baseline Comparison

Considering the relatively recent release of MIMIC-IV, there have been limited applications to this dataset. In our study, we evaluated both FuseMoE and various baselines using the earlier MIMIC-III version, as well as on our processed data from MIMIC-IV. To our knowledge, the HAIM approach described in Soenksen et al. (2022) represents the only comprehensive attempt at multimodal modeling on the MIMIC-IV dataset to date.

D.1. MISTS

This approach, from Zhang et al. (2023), casts time series and clinical notes as multivariate, irregularly-sampled time series (MISTS) and uses layers of self- and cross-attention to fuse modalities. The method uses a Time2Vec (Kazemi et al., 2019) encoding scheme to represent the irregularity of observation times. We use the same hyperparameters as in the original paper (e.g., 3 self- and cross-attention blocks, 128-dimensional time embedding, etc.).

D.2. MultT

This model from Tsai et al. (2019) relies on multiple stacks of pairwise and bidirectional cross-modal attention blocks (without a self-attention mechanism) to attend to low-level features. The results of cross-modal attention are then sent to modality-specific transformers, concatenated, and used to make predictions.

D.3. MAG

This method introduces the Multimodal Adaptation Gate (MAG) as an extension to BERT and XLNet, allowing these pre-trained models to incorporate visual and acoustic data during fine-tuning. By generating a modality-conditioned shift in their internal representations, MAG enables enhanced sentiment analysis performance on multimodal datasets, achieving human-level accuracy in the field (Rahman et al., 2020).

D.4. TF

The proposed Tensor Fusion approach (TF) integrates three core components: Modality Embedding Subnetworks for generating rich embeddings from unimodal inputs, a Tensor Fusion Layer for capturing all levels of modality interactions through a 3-fold Cartesian product, and a Sentiment Inference Subnetwork tailored to perform sentiment analysis based on the fusion layer's output (Zadeh et al., 2017).

D.5. HAIM

The multimodal fusion approach detailed by Soenksen et al. (2022) extracts a single set of features for each ICU stay, and uses this to predict the outcome of interest (in-hospital mortality, etc.). For vitals/lab values, the authors extract a set of 11 generic time series features: signal length, maximum, minimum, mean, median, SD, variance, number of peaks, and average time-series slope and piece-wise change over time of these metrics. This is done independently for each of the 30 events, leading to $30 \times 11 = 330$ vital/lab features per ICU stay. To provide a fair comparison with our method, we only include the most recent five notes and 128 vitals measurements in calculating embeddings. We only include entries for which all modalities are observed. For note/X-ray/ECG embeddings, we compute the mean embedding across all observations occurring during the specified time frame (i.e., first 48 hours of 48-IHM and LOS, the entire stay for PHE). As with our method, we standard scale values based on the training set.

Soenksen et al. (2022) uses an XGBoost (Chen & Guestrin, 2016) classifier to predict the outcomes of interest. We follow the hyperparameter optimization approach described in the paper. Specifically, we conduct a grid search across the following sets of hyperparameters: max depth $= \{5, 6, 7, 8\}$, number of estimators $= \{200, 300\}$, learning rate $= \{0.3, 0.1, 0.05\}$. Hyperparameters are selected based on the maximum AU-ROC from five-fold cross-validation

D.6. Implementation

We integrate D.1 through D.4 into our workflow using the implementation provided by (Zhang et al., 2023). For D.5, we adapt the time series (e.g., series variance, mean, etc.) feature extraction and model fitting code from the repository released by the corresponding paper. The original paper doesn't use ECG waveforms, so we adopt a similar approach to ECG embeddings as with image and note embeddings, and take the mean value of the latent vector across all included observations.

Table 6. Hyperparameters used for MoE framework and general architecture.

Hyper-Parameter Type	Parameter Name	Value
MoE	Number of experts	16
	FFN hidden size	512
	Top k	4
	Disjoint top k	2
	Hidden activation function	GeLU
	Number of MoE layers	3
Other Parameters	Random seed	[32, 42, 52]
	Training epochs	8
	Training batch size	2
	Eval batch size	8
	CNN kernel size	1
	Gradient accumulation steps	16
	BERT update epochs	2
	BERT learning rate	2.00E-05
	Time series encoder learning rate	4.00E-04
	Number of notes to include for a patient	5
	Get notes from beginning or last	Last
	Attention embedding dimension	128
	Number of attention heads	8
	Maximum time for irregular time series	48
	Time embedding dimension	64

E. Computational Resources and Hyper-Parameters

E.1. Computational resources

We train models using a Lambda Workstation with four A550 GPUs with 24 GB of memory. We are able to train models using a single GPU. An analysis of computation time and memory requirements is shown in Figure 3.

E.2. Hyper-Parameters

All parameters can be found in Table 6 and the codebase submitted in the zip file.

F. Additional Results

Below are additional results on comparing `FuseMoE` with baselines on the MIMIC-III data (Table 7) that only has vital signs and clinical notes, and the vital signs and CXR of the MIMIC-IV data (Table 8). All experiments are based on the "joint experts and router" configuration. `FuseMoE` also shows noticeable advantages in these settings.

Table 7. Comparison of `FuseMoE`-based methods (gray) and baselines, utilizing vital signs and clinical notes of MIMIC-III. The best results are highlighted in bold font, and the second-best results are underlined. All results are averaged across 3 random experiments. Since HAIM is not designed for the MIMIC-III dataset, we use the concatenation method from e.g. Khadanga et al. (2019) as a replacement.

Task \ Method		MISTS	MulT	MAG	TF	Concat	Softmax	Gaussian	Laplace
48-IHM	AUROC	89.14 ± 0.57	87.26 ± 0.35	86.53 ± 1.21	87.22 ± 0.89	86.72 ± 0.76	90.25 ± 0.74	90.77 ± 0.18	**91.19 ± 0.52**
	F1	56.45 ± 1.30	54.13 ± 1.20	53.20 ± 2.13	51.44 ± 0.66	52.77 ± 0.70	56.41 ± 0.98	56.21 ± 0.17	**57.36 ± 0.73**
25-PHE	AUROC	86.06 ± 0.06	85.96 ± 0.07	85.94 ± 0.07	84.74 ± 0.16	85.94 ± 0.21	86.41 ± 0.75	85.96 ± 0.64	**86.72 ± 0.27**
	F1	54.84 ± 0.31	54.20 ± 0.33	53.73 ± 0.37	49.84 ± 0.83	53.30 ± 0.35	55.02 ± 0.23	55.29 ± 0.45	**55.38 ± 0.69**

Table 8. Comparison of `FuseMoE`-based methods (gray) and baselines, utilizing vital signs and CXR of MIMIC-IV. The best results are highlighted in bold font, and the second-best results are underlined. All results are averaged across 3 random experiments.

Task \ Method		MISTS	MulT	MAG	TF	HAIM	Softmax	Gaussian	Laplace
48-IHM	AUROC	81.36 ± 0.24	77.70 ± 0.44	81.19 ± 1.25	76.92 ± 0.65	80.87 ± 0.00	82.08 ± 0.26	81.26 ± 0.18	**82.97 ± 0.49**
	F1	43.35 ± 0.39	28.40 ± 0.75	39.59 ± 0.43	46.59 ± 0.33	40.88 ± 0.00	38.14 ± 0.31	44.59 ± 0.24	**47.48 ± 0.23**
LOS	AUROC	82.07 ± 0.82	81.94 ± 0.26	81.86 ± 0.76	81.47 ± 0.89	81.69 ± 0.00	82.96 ± 0.47	82.74 ± 0.85	**83.22 ± 0.68**
	F1	74.07 ± 0.18	74.46 ± 0.17	73.89 ± 0.93	73.39 ± 0.14	72.93 ± 0.00	**75.67 ± 0.59**	75.16 ± 0.42	75.43 ± 0.19
25-PHE	AUROC	**71.50 ± 0.22**	71.20 ± 0.76	70.89 ± 0.47	70.55 ± 0.29	63.43 ± 0.00	71.38 ± 0.31	70.87 ± 0.67	71.44 ± 0.24
	F1	33.52 ± 0.39	32.80 ± 0.18	33.14 ± 0.61	33.56 ± 0.74	**42.45 ± 0.00**	33.49 ± 0.15	31.94 ± 0.09	34.13 ± 0.56

G. Exact-Specified Setting

In this appendix, we study the theoretical behaviors of the MLE under the exact-specified setting, i.e., $k = k_*$, of the Laplace gating Gaussian MoE. We demonstrate that under the exact-specified setting, the rate of estimated conditional density function $p_{\widehat{G}_n}$ to p_{G_*} is parametric $O(n^{-1/2})$ (up to some logarithmic factor).

Theorem G.1. *The density estimation $p_{\widehat{G}_n}(Y|X)$ converges to the true density $p_{G_*}(Y|X)$ under the Total Variation distance V at the following rate:*

$$\mathbb{E}_X[V(p_{\widehat{G}_n}(\cdot|X), p_{G_*}(\cdot|X))] = \mathcal{O}(\sqrt{\log(n)/n}).$$

Proof of Theorem G.1 is in Appendix H.1. The result of Theorem G.1 indicates that as long as we can establish the lower bound of the total variation distance between $p_{\widehat{G}_n}$ and p_{G_*} based on certain loss function between the MLE $\widehat{G}_n$ and the true mixing measure G_*, we directly achieve the rate of the MLE under that loss function.

Voronoi Loss We now define that loss function between the MLE and the true mixing measure. Given some mixing measure $G := \sum_{i=1}^{k} \exp(\beta_i)\delta_{(W_i,a_i,b_i,\nu_i)}$, we distribute its components to the following Voronoi cells which are generated by the support of the true mixing measure $G_* := \sum_{j=1}^{k_*} \exp(\beta_j^*)\delta_{(W_j^*, a_j^*, b_j^*, \nu_j^*)}$:

$$\mathcal{A}_j \equiv \mathcal{A}_j(G) := \{i \in [k] : \|\theta_i - \theta_j^*\| \leq \|\theta_i - \theta_\ell^*\|, \ \forall \ell \neq j\},$$

where $\theta_i := (W_i, a_i, b_i, \nu_i)$ and $\theta_j^* := (W_j^*, a_j^*, b_j^*, \nu_j^*)$ for any $1 \leq i \leq k$ and $1 \leq j \leq k_*$. Based on these Voronoi cells, we propose the following Voronoi loss function for the exact-specified setting:

$$\mathcal{D}_1(G, G_*) := \max \left[\Big| \sum_{i \in \mathcal{A}_{\tau_j}} \exp(\beta_i) - \exp(\beta_{\tau_j}^*) \Big| + \sum_{j \in [K] : |\mathcal{A}_j| = 1} \sum_{i \in \mathcal{A}_{\tau_j}} \exp(\beta_i) \Phi_{i\tau_j}(1, 1, 1, 1) \right], \tag{10}$$

where the maximum in the definition of Voronoi loss function $\mathcal{D}_1$ is for $\{\tau_1, \tau_2, \ldots, \tau_K\} \subseteq [k_*]$. Furthermore, for any $(\rho_1, \rho_2, \rho_3, \rho_4) \in \mathbb{R}^4$, we define $\Phi_{i\tau_j}(\rho_1, \rho_2, \rho_3, \rho_4) = \|W_i - W_{\tau_j}^*\|^{\rho_1} + \|a_i - a_{\tau_j}^*\|^{\rho_2} + |b_i - b_{\tau_j}^*|^{\rho_3} + |\nu_i - \nu_{\tau_j}^*|^{\rho_4}$ for any $i \in \mathcal{A}_{\tau_j}$ and $j \in [K]$. We demonstrate in the following theorem that the rate of MLE to the true mixing measure under the Voronoi loss function $\mathcal{D}_1$ is $\mathcal{O}(n^{-1/2})$ (up to some logarithmic factor).

Theorem G.2 (Exact-specified setting)**.** *When $k = k_*$ is known, the following Total Variation bound holds guarantetrue for any $G \in \mathcal{G}_k(\Theta)$:*

$$\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_1(G, G_*).$$

Therefore, we have $\mathcal{D}_1(\widehat{G}_n, G_*) = \mathcal{O}(\sqrt{\log(n)/n})$.

Proof of Theorem G.2 is in Appendix H.2. The convergence rate of MLE under the Voronoi loss function $\mathcal{D}_1$ implies that the rates of estimating the true parameters $W_i^*, a_i^*, b_i^*, \nu_i^*$ are also $\mathcal{O}(n^{-1/2})$ (up to logarithmic factors). These rates are comparable to those under the exact-specified setting of softmax gating Gaussian MoE (cf. Theorem 1 in (Nguyen et al., 2023b)).

H. Proof of Theoretical Results

In this appendix, we provide proofs for all theoretical results in the paper. Throughout this appendix, for any vector $v \in \mathbb{R}^d$ and $\alpha := (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}^d$, we denote $v^\alpha = v_1^{\alpha_1} v_2^{\alpha_2} \ldots v_d^{\alpha_d}$, $|v| := v_1 + v_2 + \ldots + v_d$ and $\alpha! := \alpha_1! \alpha_2! \ldots \alpha_d!$.

H.1. Proof of Theorem G.1

In this appendix, we employ results for M-estimators in (van de Geer, 2000) to establish the density estimation rate under the top-K sparse Laplace gating Gaussian mixture of experts (MoE).

Firstly, we introduce some necessary notations and fundamental results. In particular, let $\mathcal{P}_{k_*}(\Theta) := \{p_G(Y|X) : G \in \mathcal{G}_{k_*}(\Omega)\}$ be the set of all conditional density functions w.r.t mixing measures in $\mathcal{G}_{k_*}(\Omega)$. Next, we denote by $N(\varepsilon, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ the covering number of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$. Meanwhile, $H_B(\varepsilon, \mathcal{P}_{k_*}(\Omega), h)$ stands for the bracketing entropy of $\mathcal{P}_{k_*}(\Omega)$ under the Hellinger distance h where $h(p, q) := \left(\frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2 d\mu\right)^{1/2}$ for any probability densities p, q dominated by the Lebesgue measure μ. Then, we provide in the following lemma the upper bounds of those terms.

Lemma H.1. *If Ω is a bounded set, then the following inequalities hold for any $0 < \eta < 1/2$:*

(i) $\log N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta)$*;*

(ii) $H_B(\eta, \mathcal{P}_{k_*}(\Omega), h) \lesssim \log(1/\eta)$.

Proof of Lemma H.1 is in Appendix H.1.2. Subsequently, we denote

$$\widetilde{\mathcal{P}}_{k_*}(\Omega) := \{p_{(G+G_*)/2}(Y|X) : G \in \mathcal{G}_{k_*}(\Omega)\};$$
$$\widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega) := \{p^{1/2}_{(G+G_*)/2}(Y|X) : G \in \mathcal{G}_{k_*}(\Omega)\}.$$

In addition, for each $\delta > 0$, we define a Hellinger ball centered around the conditional density function $p_{G_*}(Y|X)$ and intersected with the set $\widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega)$ as

$$\widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta) := \{p^{1/2} \in \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega) : h(p, p_{G_*}) \leq \delta\}.$$

To capture the size of the above Hellinger ball, (van de Geer, 2000) suggest using the following quantity:

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta)) := \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, t), \|\cdot\|)\mathrm{d}t \vee \delta, \tag{11}$$

where $t \vee \delta := \max\{t, \delta\}$. Given those notations, let us recall a standard result for density estimation in (van de Geer, 2000).

Lemma H.2 (Theorem 7.4, (van de Geer, 2000))**.** *Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta))$ such that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then, for some sequence (δ_n) and universal constant c which satisfy $\sqrt{n}\delta_n^2 \geq c\Psi(\delta)$, we obtain that*

$$\mathbb{P}\left(\mathbb{E}_X\left[h(p_{\widehat{G}_n}(\cdot|X), p_{G_*}(\cdot|X))\right] > \delta\right) \leq c\exp(-n\delta^2/c^2),$$

for any $\delta \geq \delta_n$

Proof of Lemma H.2 can be found in (van de Geer, 2000). Now, we are ready to provide the proof for convergence rate of density estimation in Theorem G.1 in Appendix H.1.1.

H.1.1. Main Proof

It is worth noting that for any $t > 0$, we have

$$H_B(t, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, t), \|\cdot\|) \leq H_B(t, \mathcal{P}_{k_*}(\Omega, t), h).$$

Then, the integral in equation 11 is upper bounded as follows:

$$\mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta)) \leq \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(t, \mathcal{P}_{k_*}(\Omega, t), h)\mathrm{d}t \vee \delta \lesssim \int_{\delta^2/2^{13}}^{\delta} \log(1/t)\mathrm{d}t \vee \delta, \tag{12}$$

where the second inequality follows from part (ii) of Lemma H.1.

As a result, by choosing $\Psi(\delta) = \delta \cdot \sqrt{\log(1/\delta)}$, we can verify that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Furthermore, the inequality in equation 12 indicates that $\Psi(\delta) \geq \mathcal{J}_B(\delta, \widetilde{\mathcal{P}}^{1/2}_{k_*}(\Omega, \delta))$. Next, let us consider a sequence (δ_n) defined as $\delta_n := \sqrt{\log(n)/n}$. This sequence can be validated to satisfy the condition $\sqrt{n}\delta_n^2 \geq c\Psi(\delta)$ for some universal constant c. Therefore, by Lemma H.2, we reach the conclusion of Theorem G.1:

$$\mathbb{P}\Big(\mathbb{E}_X[h(p_{\widehat{G}_n}(\cdot|X), p_{G_*}(\cdot|X))] > C\sqrt{\log(n)/n}\Big) \lesssim n^{-c},$$

for some universal constant C depending only on Ω.

H.1.2. Proof of Lemma H.1

Part (i). In this part, we will derive the following upper bound for the covering number of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ for any $0 < \eta < 1/2$ given the bounded set Ω:

$$\log N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta).$$

To start with, we denote $\Theta := \{(a, b, \nu) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ : (\beta, W, a, b, \nu) \in \Omega\}$. As Ω is a bounded set, the set Θ is also bounded. Therefore, we can find an η-cover of Θ, denoted by $\overline{\Theta}_\eta$. Additionally, we also define $\Delta := \{(\beta, W) \in \mathbb{R} \times \mathbb{R}^d : (\beta, W, a, b, \nu) \in \Omega\}$, and $\overline{\Delta}_\eta$ be an η-cover of Δ. Then, it can be validated that

$$|\overline{\Theta}_\eta| \leq \mathcal{O}(\eta^{-(d+2)k_*}), \quad |\overline{\Delta}_\eta| \leq \mathcal{O}(\eta^{-(d+1)k_*}).$$

Next, for each mixing measure $G = \sum_{i=1}^{k_*} \exp(\beta_i)\delta_{(W_i, a_i, b_i, \nu_i)} \in \mathcal{G}_{k_*}(\Omega)$, we take into account two other mixing measures. The first measure is $G' = \sum_{i=1}^{k_*} \exp(\beta_i)\delta_{(W_i, \overline{a}_i, \overline{b}_i, \overline{\nu}_i)}$, where $(\overline{a}_i, \overline{b}_i, \overline{\nu}_i) \in \overline{\Theta}_\eta$ is the closest points to (a_i, b_i, ν_i) in this set for all $i \in [k_*]$. The second one is $\overline{G} := \sum_{i=1}^{k_*} \exp(\overline{\beta}_i)\delta_{(\overline{W}_i, \overline{a}_i, \overline{b}_i, \overline{\nu}_i)}$ in which $(\overline{\beta}_i, \overline{W}_i) \in \overline{\Delta}_\eta$ for any $i \in [k_*]$. Next, let us define

$$\mathcal{T} := \{p_{\overline{G}} \in \mathcal{P}_{k_*}(\Omega) : (\overline{\beta}_i, \overline{W}_i) \in \overline{\Delta}_\eta,\ (\overline{a}_i, \overline{b}_i, \overline{\nu}_i) \in \overline{\Theta}_\eta, \forall i \in [k_*]\},$$

then it is obvious that $p_{\overline{G}} \in \mathcal{T}$. Now, we will show that $\mathcal{T}$ is an η-cover of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ with a note that it is not necessarily the smallest cover. Indeed, according to the triangle inequality,

$$\|p_G - p_{\overline{G}}\|_1 \leq \|p_G - p_{G'}\|_1 + \|p_{G'} - p_{\overline{G}}\|_1. \tag{13}$$

Since the softmax function is no greater than one, the first term in the right hand side can be upper bounded as follows:

$$\begin{aligned}
\|p_G - p_{G'}\|_1 &\leq \sum_{i=1}^{k_*} \int \Big| f(Y|a_i^\top X + b_i, \nu_i) - f(Y|\overline{a}_i^\top X + \overline{b}_i, \overline{\nu}_i) \Big| \mathrm{d}(X, Y) \\
&\lesssim \sum_{i=1}^{k_*} \int \Big(\|a_i - \overline{a}_i\| + \|b_i - \overline{b}_i\| + \|\nu_i - \overline{\nu}_i\| \Big) \mathrm{d}(X, Y) \\
&= \sum_{i=1}^{k_*} \Big(\|a_i - \overline{a}_i\| + \|b_i - \overline{b}_i\| + \|\nu_i - \overline{\nu}_i\| \Big) \\
&\lesssim \eta.
\end{aligned} \tag{14}$$

Subsequently, we show that $\|p_{G'} - p_{\overline{G}}\|_1 \lesssim \eta$. For that purpose, we consider $q := \binom{k_*}{K}$ K-element subsets of $\{1, \ldots, k_*\}$, which are assumed to take the form $\{\tau_1, \tau_2, \ldots, \tau_K\}$ for any $\tau \in [q]$. Additionally, we also denote $\{\tau_{K+1}, \ldots, \tau_{k_*}\} := \{1, \ldots, k_*\} \setminus \{\tau_1, \ldots, \tau_K\}$ for any $\tau \in [q]$. Then, we define

$$\begin{aligned}
\mathcal{X}_\tau &:= \{x \in \mathcal{X} : -\|W_i - x\| \geq -\|W_{i'} - x\| : i \in \{\tau_1, \ldots, \tau_K\}, i' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}\}, \\
\widetilde{\mathcal{X}}_\tau &:= \{x \in \mathcal{X} : -\|\overline{W}_i - x\| \geq -\|\overline{W}_{i'} - x\| : i \in \{\tau_1, \ldots, \tau_K\}, i' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}\}.
\end{aligned}$$

By using the same arguments as in the proof of Lemma I.1 in Appendix I, we achieve that either $\mathcal{X}_\tau = \widetilde{\mathcal{X}}_\tau$ or $\mathcal{X}_\tau$ has measure zero for any $\tau \in [q]$. As the $\mathrm{softmax}$ function is differentiable, it is a Lipschitz function with some Lipschitz constant $L \geq 0$. Next, we denote

$$\pi_\tau(X) := \Big(-\|W_{\tau_i} - X\| + \beta_{\tau_i} \Big)_{i=1}^{K}; \qquad \overline{\pi}_\tau(X) := \Big(-\|\overline{W}_{\tau_i} - X\| + \overline{\beta}_{\tau_i} \Big)_{i=1}^{K},$$

for any K-element subset $\{\tau_1, \ldots \tau_K\}$ of $\{1, \ldots, k_*\}$. Then, we get

$$\begin{aligned}
\|\mathrm{softmax}(\pi_\tau(X)) - \mathrm{softmax}(\overline{\pi}_\tau(X))\| &\leq L \cdot \|\pi_\tau(X) - \overline{\pi}_\tau(X)\| \\
&\leq L \cdot \sum_{i=1}^{K} \Big(\Big| \|W_{\tau_i} - X\| - \|\overline{W}_{\tau_i} - X\| \Big| + |\beta_{\tau_i} - \overline{\beta}_{\tau_i}| \Big) \\
&\leq L \cdot \sum_{i=1}^{K} \Big(\|W_{\tau_i} - \overline{W}_{\tau_i}\| + |\beta_{\tau_i} - \overline{\beta}_{\tau_i}| \Big) \\
&\leq L \cdot \sum_{i=1}^{K} \Big(\eta + \eta \Big) \\
&\lesssim \eta.
\end{aligned}$$

Back to the proof for $\|p_{G'} - p_{\overline{G}}\|_1 \lesssim \eta$, it follows from the above results that

$$\begin{aligned}\|p_{G'} - p_{\overline{G}}\|_1 &= \int |p_{G'}(Y|X) - p_{\overline{G}}(Y|X)| \,\mathrm{d}(X,Y) \\ &\leq \sum_{\tau=1}^{q} \int_{\mathcal{X}_\tau \times \mathcal{Y}} |p_{G'}(Y|X) - p_{\overline{G}}(Y|X)| \,\mathrm{d}(X,Y) \\ &\leq \sum_{\tau=1}^{q} \int_{\mathcal{X}_\tau \times \mathcal{Y}} \sum_{i=1}^{K} \Big|\mathrm{softmax}(\pi_\tau(X)_i) - \mathrm{softmax}(\overline{\pi}_\tau(X)_i)\Big| \cdot \Big| f(Y|\overline{a}_{\tau_i}^\top X + \overline{b}_{\tau_i}, \overline{\nu}_{\tau_i})\Big| \,\mathrm{d}(X,Y) \\ &\lesssim \eta, \end{aligned} \tag{15}$$

It follows from the results in equation 13, equation 14 and equation 15 that $\|p_G - p_{\overline{G}}\|_1 \lesssim \eta$. This result indicates that $\mathcal{T}$ is an η-cover of the metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$. As a consequence, we obtain that

$$N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim |\overline{\Delta}_\eta| \times |\overline{\Theta}_\eta| \leq \mathcal{O}(1/\eta^{(2d+3)k}),$$

which leads to the conclusion of this part: $\log N(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta)$.

Part (ii). In this part, we provide an upper bound for the bracketing entropy of $\mathcal{P}_{k_*}(\Omega)$ under the Hellinger distance h:

$$H_B(\eta, \mathcal{P}_{k_*}(\Omega), h) \lesssim \log(1/\eta).$$

Since Ω and $\mathcal{X}$ are bounded sets, there exist positive constants γ, ℓ, u such that $-\gamma \leq a^\top X + b \leq \gamma$ and $\ell \leq \nu \leq u$. Let us define

$$B(Y|X) := \begin{cases} \frac{1}{\sqrt{2\pi\ell}} \exp\Big(-\frac{Y^2}{8u}\Big), & \text{for } |Y| \geq 2\gamma \\ \frac{1}{\sqrt{2\pi\ell}}, & \text{for } |Y| < 2\gamma \end{cases}$$

Then, it can be validated that $f(Y|a^\top X + b, \nu) \leq B(Y|X)$ for any $(X,Y) \in \mathcal{X} \times \mathcal{Y}$.

Next, let $\zeta \leq \eta$ which will be chosen later and $\{p_1, \ldots, p_N\}$ be an ζ-cover of metric space $(\mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$ with the covering number $N := N(\tau, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1)$. Additionally, we also consider brackets of the form $[\Psi_i^L(Y|X), \Psi_i^U(Y|X)]$ where

$$\begin{aligned}\Psi_i^L(Y|X) &:= \max\{p_i(Y|X) - \tau, 0\} \\ \Psi_i^U(Y|X) &:= \max\{p_i(Y|X) + \tau, B(Y|X)\}.\end{aligned}$$

Then, we can check that $\mathcal{P}_{k_*}(\Omega) \subseteq \bigcup_{i=1}^N [\Psi_i^L(Y|X), \Psi_i^U(Y|X)]$ and $\Psi_i^U(Y|X) - \Psi_i^L(Y|X) \leq \min\{2\eta, B(Y|X)\}$.

Let $S := \max\{2\gamma, \sqrt{8u}\} \log(1/\zeta)$, we have for any $i \in [N]$ that

$$\begin{aligned}\|\Psi_i^U - \Psi_i^L\|_1 &= \int_{|Y|<2\gamma} [\Psi_i^U(Y|X) - \Psi_i^L(Y|X)] \,\mathrm{d}X\mathrm{d}Y + \int_{|Y|\geq 2\gamma} [\Psi_i^U(Y|X) - \Psi_i^L(Y|X)] \,\mathrm{d}X\mathrm{d}Y \\ &\leq S\zeta + \exp\Big(-\frac{S^2}{2u}\Big) \leq S'\zeta,\end{aligned}$$

where S' is some positive constant. This inequality indicates that

$$H_B(S'\zeta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \leq \log N(\zeta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \leq \log(1/\tau).$$

By setting $\zeta = \eta/S'$, we obtain that $H_B(\eta, \mathcal{P}_{k_*}(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta)$. Finally, since the norm $\|\cdot\|_1$ is upper bounded by the Hellinger distance, we reach the conclusion of this part:

$$H_B(\eta, \mathcal{P}_{k_*}(\Omega), h) \lesssim \log(1/\eta).$$

Hence, the proof is completed.

H.2. Proof of Theorem G.2

First of all, we need to establish the following bound:

$$\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_1(G, G_*).$$

For that sake, it is sufficient to demonstrate two following inequalities:

- **Inequality A.** $\inf_{G\in\mathcal{G}_{k_*}(\Omega):\mathcal{D}_1(G,G_*)\leq\varepsilon'} \dfrac{\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0$;
- **Inequality B.** $\inf_{G\in\mathcal{G}_{k_*}(\Omega):\mathcal{D}_1(G,G_*)>\varepsilon'} \dfrac{\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0$,

for some constant $\varepsilon' > 0$.

Proof of inequality A: The inequality A is equivalent to

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_{k_*}(\Omega):\mathcal{D}_1(G,G_*)\leq\varepsilon} \frac{\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0.$$

Assume that the above inequality is not true, then, there exists a sequence of mixing measure $G_n := \sum_{i=1}^{k_*}\exp(\beta_i^n)\delta_{(W_i^n,a_i^n,b_i^n,\nu_i^n)} \in \mathcal{G}_{k_*}(\Omega)$ such that both $\mathcal{D}_1(G_n, G_*)$ and $\mathbb{E}_X[V(p_{G_n}(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_1(G_n, G_*)$ go to zero as $n\to\infty$. Now, we define

$$\mathcal{A}_j^n = \mathcal{A}_j(G_n) := \{i\in[k_*] : \|\theta_i^n - \theta_j^*\| \leq \|\theta_i^n - \theta_\tau^*\|,\ \forall \tau\neq j\},$$

for any $j\in[k_*]$ as Voronoi cells with respect to the mixing measure G_n, where we denote $\theta_i^n := (W_i^n, a_i^n, b_i^n, \nu_i^n)$ and $\theta_j^* := (W_j^*, a_j^*, b_j^*, \nu_j^*)$. In this proof, since our arguments are assymptotic, we can assume without loss of generality (WLOG) that these Voronoi cells does not depend on n, that is, $\mathcal{A}_j = \mathcal{A}_j^n$. Next, it follows from the hypothesis $\mathcal{D}_{1n} := \mathcal{D}_1(G_n, G_*)\to 0$ as $n\to\infty$ that each Voronoi cell contains only one element. Therefore, we may assume WLOG that $\mathcal{A}_j = \{j\}$ for any $j\in[k_*]$, which implies that $(W_j^n, a_j^n, b_j^n, \nu_j^n)\to(W_j^*, a_j^*, b_j^*, \nu_j^*)$ and $\exp(\beta_j^n)\to\exp(\beta_j^*)$ as $n\to\infty$.

Subsequently, to specify the top K selection in the formulations of $p_{G_n}(Y|X)$ and $p_{G_*}(Y|X)$, we divide the covariate space $\mathcal{X}$ into some subsets in two ways. In particular, we first consider $q := \binom{k_*}{K}$ different K-element subsets of $[k_*]$, which are assumed to take the form $\{\tau_1,\ldots,\tau_K\}$, for $\tau\in[q]$. Additionally, we denote $\{\tau_{K+1},\ldots,\tau_{k_*}\} := [k_*]\setminus\{\tau_1,\ldots,\tau_K\}$. Then, we define for each $\tau\in[q]$ two following subsets of $\mathcal{X}$:

$$\mathcal{X}_\tau^n := \Big\{x\in\mathcal{X} : -\|W_j - x\| \geq -\|W_{j'} - x\| : \forall j\in\{\tau_1,\ldots,\tau_K\}, j'\in\{\tau_{K+1},\ldots,\tau_{k_*}\}\Big\},$$
$$\mathcal{X}_\tau^* := \Big\{x\in\mathcal{X} : -\|W_j^* - x\| \geq -\|W_{j'}^* - x\| : \forall j\in\{\tau_1,\ldots,\tau_K\}, j'\in\{\tau_{K+1},\ldots,\tau_{k_*}\}\Big\}.$$

Since $(W_j^n, \beta_j^n)\to(W_j^*, \beta_j^*)$ as $n\to\infty$ for any $j\in[k_*]$, we have for any arbitrarily small $\eta_j > 0$ that $\|W_j^n - W_j^*\|\leq\eta_j$ and $|\beta_j^n - \beta_j^*|\leq\eta_j$ for sufficiently large n. By applying Lemma I.1, we obtain that $\mathcal{X}_\tau^n = \mathcal{X}_\tau^*$ for any $\tau\in[q]$ for sufficiently large n.

WLOG, we assume that

$$\mathcal{D}_1(G_n, G_*) = \sum_{i=1}^{K}\Big[\exp(\beta_i^n)\Big(\|\Delta W_i^n\| + \|\Delta a_i^n\| + \|\Delta b_i^n\| + \|\Delta\nu_i^n\|\Big) + \Big|\exp(\beta_i^n) - \exp(\beta_i^*)\Big|\Big],$$

where we denote $\Delta\beta_{1i}^n := \beta_{1i}^n - \beta_{1i}^*$, $\Delta a_i^n := a_i^n - a_i^*$, $\Delta b_i^n := b_i^n - b_i^*$ and $\Delta\nu_i^n := \nu_i^n - \nu_i^*$.

Let $\tau\in[q]$ such that $\{\tau_1,\ldots,\tau_K\} = \{1,\ldots,K\}$. Then, for almost surely $(X,Y)\in\mathcal{X}_\tau^*\times\mathcal{Y}$, we can rewrite the conditional densities $p_{G_n}(Y|X)$ and $p_{G_*}(Y|X)$ as

$$p_{G_n}(Y|X) = \sum_{i=1}^{K}\frac{\exp(-\|W_i^n - X\| + \beta_i^n)}{\sum_{j=1}^{K}\exp(-\|W_j^n - X\| + \beta_j^n)}\cdot f(Y|(a_i^n)^\top X + b_i^n, \nu_i^n),$$
$$p_{G_*}(Y|X) = \sum_{i=1}^{K}\frac{\exp(-\|W_i^* - X\| + \beta_i^*)}{\sum_{j=1}^{K}\exp(-\|W_j^* - X\| + \beta_j^*)}\cdot f(Y|(a_i^*)^\top X + b_i^*, \nu_i^*).$$

Now, we break the rest of our arguments into three steps:

Stage 1 - Density decomposition:

In this step, we aim to decompose the term $Q_n := \Big[\sum_{i=1}^{K} \exp(-\|W_i^* - X\| + \beta_i^*)\Big] \cdot [p_{G_n}(Y|X) - p_{G_*}(Y|X)]$, which can be represented as follows:

$$\begin{aligned} Q_n &= \sum_{i=1}^{K} \exp(\beta_i^n) \Big[F(Y|X; W_i^n, a_i^n, b_i^n, \nu_i^n) - F(Y|X; W_i^*, a_i^*, b_i^*, \nu_i^*)\Big] \\ &- \sum_{i=1}^{K} \exp(\beta_i^n) \Big[H(Y|X; W_i^n) - H(Y|X; W_i^*)\Big] \\ &+ \sum_{i=1}^{K} \Big[\exp(\beta_i^n) - \exp(\beta_i^*)\Big] \Big[F(Y|X; W_i^*, a_i^*, b_i^*, \nu_i^*) - H(Y|X, W_i^*)\Big] \\ &: = A_n - B_n + E_n, \end{aligned} \tag{16}$$

where we denote $F(Y|X; W, a, b, \nu) := \exp(-\|W - X\|) f(Y|a^\top X + b, \nu)$ and $H(Y|X; W) = \exp(-\|W - X\|) p_{G_n}(Y|X)$. By applying the first-order Taylor expansion, we can rewrite A_n as

$$\begin{aligned} A_n &= \sum_{i=1}^{K} \sum_{|\alpha|=1} \frac{\exp(\beta_i^n)}{\alpha!} \cdot (\Delta W_i^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta \nu_i^n)^{\alpha_4} \cdot \frac{\partial^{|\alpha_1|+|\alpha_2|+\alpha_3+\alpha_4} F}{\partial W^{\alpha_1} \partial a^{\alpha_2} \partial b^{\alpha_3} \partial \nu^{\alpha_4}} (Y|X; W_i^*, a_i^*, b_i^*, \nu_i^*) + R_1(X, Y) \\ &= \sum_{i=1}^{K} \sum_{|\alpha|=1} \frac{\exp(\beta_i^n)}{\alpha!} \cdot (\Delta W_i^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta \nu_i^n)^{\alpha_4} \cdot \frac{\partial^{|\alpha_1|} g}{\partial W^{\alpha_1}} (X; W_i^*) \cdot \frac{\partial^{|\alpha_2|+\alpha_3+\alpha_4} f}{\partial a^{\alpha_2} \partial b^{\alpha_3} \partial \nu^{\alpha_4}} (Y|(a_i^*)^\top X + b_i^*, \nu_i^*) \\ &\qquad + R_1(X, Y), \end{aligned}$$

where $R_1(X, Y)$ is a Taylor remainder that satisfies $R_1(X, Y)/\mathcal{D}_1(X, Y) \to 0$ as $n \to \infty$ and $g(X, W) := \exp(\|W - X\|)$. Recall that f is the univariate Gaussian density, then by denoting $h_1(X; a, b) := a^\top X + b$, we can verify that

$$\frac{\partial^{\alpha_4} f}{\partial \nu^{\alpha_4}} (Y|(a_i^*)^\top X + b_i^*, \nu_i^*) = \frac{1}{2^{\alpha_4}} \cdot \frac{\partial^{2\alpha_4} f}{\partial h_1^{2\alpha_4}} (Y|(a_i^*)^\top X + b_i^*, \nu_i^*).$$

Consequently, we get

$$\begin{aligned} A_n &= \sum_{i=1}^{K} \sum_{|\alpha|=1} \frac{\exp(\beta_i^n)}{2^{\alpha_4} \alpha!} \cdot (\Delta W_i^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta \nu_i^n)^{\alpha_4} \\ &\qquad \times X^{\alpha_2} \cdot \frac{\partial^{|\alpha_1|} g}{\partial W^{\alpha_1}} (X; W_i^*) \cdot \frac{\partial^{|\alpha_2|+\alpha_3+2\alpha_4} f}{\partial h_1^{|\alpha_2|+\alpha_3+2\alpha_4}} (Y|(a_i^*)^\top X + b_i^*, \nu_i^*) + R_1(X, Y) \\ &= \sum_{i=1}^{K} \sum_{|\alpha_1|=0}^{1} \sum_{|\alpha_2|=0}^{1-|\alpha_1|} \sum_{\eta=0}^{2(1-|\alpha_1|-|\alpha_2|)} \sum_{\substack{\alpha_3+2\alpha_4=\eta, \\ 0 \le \alpha_3+\alpha_4 \le 1-|\alpha_1|-|\alpha_2|}} \frac{\exp(\beta_i^n)}{2^{\alpha_4} \alpha!} \cdot (\Delta W_i^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta \nu_i^n)^{\alpha_4} \\ &\qquad \times X^{\alpha_2} \cdot \frac{\partial^{|\alpha_1|} g}{\partial W^{\alpha_1}} (X; W_i^*) \cdot \frac{\partial^{|\alpha_2|+\eta} f}{\partial h_1^{|\alpha_2|+\eta}} (Y|(a_i^*)^\top X + b_i^*, \nu_i^*) + R_1(X, Y), \end{aligned} \tag{17}$$

where we denote $\eta = \alpha_3 + 2\alpha_4 \in \mathbb{N}$.

Subsequently, we also apply the first-order Taylor expansion to the term B_n defined in equation 16 and get that

$$\begin{aligned} B_n &= \sum_{i=1}^{K} \sum_{|\gamma|=1} \frac{\exp(\beta_i^n)}{\gamma!} (\Delta W_i^n)^{\gamma} \cdot \frac{\partial^{|\gamma|} H}{\partial W^{\gamma}} (Y|X; W_i^*) + R_2(X, Y) \\ &= \sum_{i=1}^{K} \sum_{|\gamma|=1} \frac{\exp(\beta_i^n)}{\gamma!} (\Delta W_i^n)^{\gamma} \cdot \frac{\partial^{|\gamma|} g}{\partial W^{\gamma}} (X; W_i^*) p_{G_n}(Y|X) + R_2(X, Y), \end{aligned} \tag{18}$$

where $R_2(X,Y)$ is a Taylor remainder such that $R_2(X,Y)/\mathcal{D}_1(G_n,G_*) \to 0$ as $n \to \infty$.

From the above results, the term Q_n can be rewritten as

$$Q_n = \sum_{i=1}^{K} \sum_{|\alpha_1|=0}^{1} \sum_{|\alpha_2|=0}^{1-|\alpha_1|} \sum_{\eta=0}^{2(1-|\alpha_1|-|\alpha_2|)} S^n_{i,\alpha_1,\alpha_2,\eta} \cdot X^{\alpha_2} \cdot \frac{\partial^{|\alpha_1|} g}{\partial W^{\alpha_1}}(X;W_i^*) \cdot \frac{\partial^{|\alpha_2|+\eta} f}{\partial h_1^{|\alpha_2|+\eta}}(Y|(a_i^*)^\top X + b_i^*, \nu_i^*)$$
$$+ \sum_{i=1}^{K} \sum_{|\gamma|=0}^{1} T^n_{i,\gamma} \cdot \frac{\partial^{|\gamma|} g}{\partial W^{\gamma}}(X;W_i^*) p_{G_n}(Y|X) + R_1(X,Y) + R_2(X,Y), \tag{19}$$

in which we respectively define for each $i \in [K]$ that

$$S^n_{i,\alpha_1,\alpha_2,\eta} := \sum_{\substack{\alpha_3+2\alpha_4=\eta,\\ 0\leq \alpha_3+\alpha_4 \leq 1-|\alpha_1|-|\alpha_2|}} \frac{\exp(\beta_i^n)}{2^{\alpha_4}\alpha!} \cdot (\Delta W_i^n)^{\alpha_1} (\Delta a_i^n)^{\alpha_2} (\Delta b_i^n)^{\alpha_3} (\Delta \nu_i^n)^{\alpha_4},$$
$$T^n_{i,\gamma} := \frac{\exp(\beta_i^n)}{\gamma!} (\Delta W_i^n)^{\gamma},$$

for any $(\alpha_1,\alpha_2,\eta) \neq (\mathbf{0}_d, \mathbf{0}_d, 0)$ and $\gamma \neq \mathbf{0}_d$. Otherwise, $S^n_{i,\mathbf{0}_d,\mathbf{0}_d,0} = T^n_{i,\mathbf{0}_d} := \exp(\beta_i^n) - \exp(\beta_i^*)$.

Stage 2 - Non-vanishing coefficients:

Moving to the second step, we will show that not all the ratios $S^n_{i,\alpha_1,\alpha_2,\eta}/\mathcal{D}_1(G_n,G_*)$ and $T^n_{i,\gamma}/\mathcal{D}_1(G_n,G_*)$ tend to zero as $n \to \infty$. Assume by contrary that all of them approach zero when $n \to \infty$, then for $(\alpha_1,\alpha_2,\eta) = (\mathbf{0}_d,\mathbf{0}_d,0)$, it follows that

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \Big| \exp(\beta_i^n) - \exp(\beta_i^*) \Big| = \sum_{i=1}^{K} \frac{|S^n_{i,\alpha_1,\alpha_2,\eta}|}{\mathcal{D}_1(G_n,G_*)} \to 0. \tag{20}$$

Additionally, for tuples (α_1,α_2,η) where $\alpha_1 \in \{e_1, e_2, \ldots, e_d\}$ with $e_j := (0,\ldots,0,\underbrace{1}_{j-th},0,\ldots,0)$, $\alpha_2 = \mathbf{0}_d$ and $\eta = 0$, we get

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_i^n) \|\Delta W_i^n\|_1 = \sum_{i=1}^{K} \frac{|S^n_{i,\alpha_1,\alpha_2,\eta}|}{\mathcal{D}_1(G_n,G_*)} \to 0.$$

For (α_1,α_2,η) where $\alpha_1 = \mathbf{0}_d$, $\alpha_2 \in \{e_1,e_2,\ldots,e_d\}$ and $\eta = 0$, we have

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_i^n) \|\Delta a_i^n\|_1 = \sum_{i=1}^{K} \frac{|S^n_{i,\alpha_1,\alpha_2,\eta}|}{\mathcal{D}_1(G_n,G_*)} \to 0.$$

For (α_1,α_2,η) where $\alpha_1 = \alpha_2 = \mathbf{0}_d$ and $\eta = 1$, we have

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_i^n) \|\Delta b_i^n\|_1 = \sum_{i=1}^{K} \frac{|S^n_{i,\alpha_1,\alpha_2,\eta}|}{\mathcal{D}_1(G_n,G_*)} \to 0.$$

For (α_1,α_2,η) where $\alpha_1 = \alpha_2 = \mathbf{0}_d$ and $\eta = 2$, we have

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_i^n) \|\Delta \nu_i^n\|_1 = \sum_{i=1}^{K} \frac{|S^n_{i,\alpha_1,\alpha_2,\eta}|}{\mathcal{D}_1(G_n,G_*)} \to 0.$$

As a result, we achieve that

$$\frac{1}{\mathcal{D}_1(G_n,G_*)} \cdot \sum_{i=1}^{K} \exp(\beta_i^n) \Big[\|\Delta W_i^n\|_1 + \|\Delta a_i^n\|_1 + |\Delta b_i^n| + |\Delta \nu_i^n| \Big] \to 0.$$

Due to the topological equivalence between norm-1 and norm-2, the above limit implies that

$$\frac{1}{\mathcal{D}_1(G_n,G_*)}\cdot\sum_{i=1}^{K}\exp(\beta_i^n)\Big[\|\Delta W_i^n\|+\|\Delta a_i^n\|+|\Delta b_i^n|+|\Delta\nu_i^n|\Big]\to 0. \tag{21}$$

Combine equation 20 with equation 21, we deduce that $\mathcal{D}_1(G_n,G_*)/\mathcal{D}_1(G_n,G_*)\to 0$, which is a contradiction. Consequently, at least one among the ratios $S^n_{i,\alpha_1,\alpha_2,\eta}/\mathcal{D}_1(G_n,G_*)$ and $T^n_{i,\gamma}/\mathcal{D}_1(G_n,G_*)$ does not vanish as n tends to infinity.

Stage 3 - Fatou's contradiction:

In this step, we use the Fatou's lemma to point out a contradiction to the results achieved in Step 2. In particular, we denote by m_n the maximum of the absolute values of $S^n_{i,\alpha_1,\alpha_2,\eta}/\mathcal{D}_1(G_n,G_*)$ and $T^n_{i,\gamma}/\mathcal{D}_1(G_n,G_*)$. Since at least one of the previous ratios does not converge to zero, we deduce that $1/m_n\not\to\infty$.

Recall from the hypothesis that $\mathbb{E}_X[V(p_{G_n}(\cdot|X),p_{G_*}(\cdot|X))]/\mathcal{D}_1(G_n,G_*)\to 0$ as $n\to\infty$. According to the Fatou's lemma, we have

$$0=\lim_{n\to\infty}\frac{\mathbb{E}_X[V(p_{G_n}(\cdot|X),p_{G_*}(\cdot|X))]}{\mathcal{D}_1(G_n,G_*)}\geq\frac{1}{2}\cdot\int\liminf_{n\to\infty}\frac{|p_{G_n}(Y|X)-p_{G_*}(Y|X)|}{\mathcal{D}_1(G_n,G_*)}\mathrm{d}(X,Y)\geq 0.$$

This result indicates that $|p_{G_n}(Y|X)-p_{G_*}(Y|X)|/\mathcal{D}_1(G_n,G_*)$ tends to zero as n goes to infinity for almost surely (X,Y). As a result, it follows that

$$\lim_{n\to\infty}\frac{Q_n}{m_n\mathcal{D}_1(G_n,G_*)}=\lim_{n\to\infty}\frac{|p_{G_n}(Y|X)-p_{G_*}(Y|X)|}{m_n\mathcal{D}_1(G_n,G_*)}=0.$$

Next, let us denote $S^n_{i,\alpha_1,\alpha_2,\eta}/[m_n\mathcal{D}_1(G_n,G_*)]\to\xi_{i,\alpha_1,\alpha_2,\eta}$ and $T^n_{i,\gamma}/[m_n\mathcal{D}_1(G_n,G_*)]\to\kappa_{i,\gamma}$ with a note that at least one among them is non-zero. From the formulation of Q_n in equation 19, we deduce that

$$\sum_{i=1}^{K}\sum_{|\alpha_1|=0}^{1}\sum_{|\alpha_2|=0}^{1-|\alpha_1|}\sum_{\eta=0}^{2(1-|\alpha_1|-|\alpha_2|)}\xi_{i,\alpha_1,\alpha_2,\eta}\cdot X^{\alpha_2}\cdot\frac{\partial^{|\alpha_1|}g}{\partial W^{\alpha_1}}(X;W_i^*)\cdot\frac{\partial^{|\alpha_2|+\eta}f}{\partial h_1^{|\alpha_2|+\eta}}(Y|(a_i^*)^\top X+b_i^*,\nu_i^*)$$
$$+\sum_{i=1}^{K}\sum_{|\gamma|=0}^{1}\kappa_{i,\gamma}\cdot\frac{\partial^{|\gamma|}g}{\partial W^{\gamma}}(X;W_i^*)p_{G_n}(Y|X)=0, \tag{22}$$

for almost surely (X,Y). The above equation is equivalent to

$$\sum_{i=1}^{K}\sum_{|\alpha_1|=0}^{1}\left[\sum_{|\alpha_2|=0}^{1-|\alpha_1|}\sum_{\eta=0}^{2(1-|\alpha_1|-|\alpha_2|)}\xi_{i,\alpha_1,\alpha_2,\eta}\cdot X^{\alpha_2}\frac{\partial^{\alpha_2+\eta}f}{\partial h_1^{\alpha_2+\eta}}(Y|(a_i^*)^\top X+b_i^*,\nu_i^*)+\kappa_{i,\alpha_1}p_{G_*}(Y|X)\right]\frac{\partial^{|\alpha_1|}g}{\partial W^{\alpha_1}}(X;W_i^*)=0,$$

for almost surely (X,Y). It is worth noting that parameters $W_1^*,\ldots,W_K^*$ are pair-wise distinct, thus, the set $\left\{\frac{\partial^{|\alpha_1|}g}{\partial W^{\alpha_1}}(X;W_i^*):i\in[K],\ 0\leq|\alpha_1|\leq 1\right\}$ is a linearly independent, which implies that

$$\sum_{|\alpha_2|=0}^{1-|\alpha_1|}\sum_{\eta=0}^{2(1-|\alpha_1|-|\alpha_2|)}\xi_{i,\alpha_1,\alpha_2,\eta}\cdot X^{\alpha_2}\frac{\partial^{\alpha_2+\eta}f}{\partial h_1^{\alpha_2+\eta}}(Y|(a_i^*)^\top X+b_i^*,\nu_i^*)+\kappa_{i,\alpha_1}p_{G_*}(Y|X)=0,$$

for any $i\in[K]$, $0\leq|\alpha_1|\leq 1$ and for almost surely (X,Y). Moreover, since $(a_1^*,b_1^*,\nu_1^*),\ldots,(a_K^*,b_K^*,\nu_K^*)$ have pair-wise distinct values, those of $((a_1^*)^\top X+b_1^*,\nu_1^*),\ldots,((a_K^*)^\top X+b_K^*,\nu_K^*)$ are also pair-wise different. Therefore, the set

$$\left\{X^{\alpha_2}\frac{\partial^{\alpha_2+\eta}f}{\partial h_1^{\alpha_2+\eta}}(Y|(a_i^*)^\top X+b_i^*,\nu_i^*),\ p_{G_*}(Y|X):0\leq|\alpha_2|\leq 1-|\alpha_1|,\ 0\leq\eta\leq 2(1-|\alpha_1|-|\alpha_2|)\right\}$$

is also linearly independent. Consequently, we obtain that $\xi_{i,\alpha_1,\alpha_2,\eta}=\kappa_{i,\gamma}=0$ for any $i\in[K]$, $0\leq|\alpha_1|+\alpha_2\leq 1$, $0\leq\eta\leq 2(1-|\alpha_1|-|\alpha_2|)$ and $0\leq|\gamma|\leq 1$, which contradicts the fact that at least one among those terms is different from zero.

Hence, we can find some constant $\varepsilon' > 0$ such that

$$\inf_{G \in \mathcal{G}_{k_*}(\Omega): \mathcal{D}_1(G, G_*) \leq \varepsilon'} \frac{\mathbb{E}_X[V(p_G(\cdot|X), p_{G_*}(\cdot|X))]}{\mathcal{D}_1(G, G_*)} > 0.$$

Proof of inequality B: Assume by contrary that the inequality B does not hold, then there exists a sequence of mixing measures $G'_n \in \mathcal{G}_{k_*}(\Omega)$ such that $\mathcal{D}_1(G'_n, G_*) > \varepsilon'$ and

$$\lim_{n \to \infty} \frac{\mathbb{E}_X[V(p_{G'_n}(\cdot|X), p_{G_*}(\cdot|X))]}{\mathcal{D}_1(G'_n, G_*)} = 0.$$

This result leads to $\mathbb{E}_X[V(p_{G'_n}(\cdot|X), p_{G_*}(\cdot|X))] \to 0$ as $n \to \infty$. Recall that Ω is a compact set, therefore, we can replace the sequence G'_n by one of its subsequences that converges to a mixing measure $G' \in \mathcal{G}_{k_*}(\Omega)$. Since $\mathcal{D}_1(G'_n, G_*) > \varepsilon'$, this result induces that $\mathcal{D}_1(G', G_*) > \varepsilon'$.

Subsequently, by means of the Fatou's lemma, we achieve that

$$0 = \lim_{n \to \infty} \mathbb{E}_X[2V(p_{G'_n}(\cdot|X), p_{G_*}(\cdot|X))] \geq \int \liminf_{n \to \infty} \left| p_{G'_n}(Y|X) - p_{G_*}(Y|X) \right| \mathrm{d}(X, Y).$$

It follows that $p_{G'}(Y|X) = p_{G_*}(Y|X)$ for almost surely (X, Y). According to Lemma I.3, the noisy top-K sparse softmax gating Gaussian mixture of experts is identifiable, thus, we obtain that $G' \equiv G_*$. As a consequence, we obtain that $\mathcal{D}_1(G', G_*) = 0$, which contradicts to the fact that $\mathcal{D}_1(G', G_*) > \varepsilon' > 0$.

Hence, the proof is completed.

H.3. Proof of Theorem 4.1

In this appendix, we leverage proof techniques in Appendix H.1 to derive the density estimation rate under the over-specified setting in Theorem 4.1. Recall that under this setting, the MLE $\widehat{G}_n$ belongs to $\mathcal{G}_k(\Omega)$, i.e. the set of all mixing measures with at most $k > k_*$ components, where k is unknown.

It is worth noting that if we can adapt the result of part (i) of Lemma H.1 to the over-specified settings, then other results presented in Appendix H.1 will also hold true. Therefore, our main goal is to establish following bound for the covering number of the metric space $(\mathcal{P}_k(\Omega), \|\cdot\|_1)$:

$$\log N(\eta, \mathcal{P}_k(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta),$$

for any $0 < \eta < 1/2$, where $\mathcal{P}_k(\Omega) := \{\overline{p}_G(Y|X) : G \in \mathcal{G}_k(\Omega)\}$.

To facilitate the presentation, we reuse the notations defined in Appendix H.1 with $\mathcal{G}_{k_*}(\Omega)$ being substituted by $\mathcal{G}_k(\Omega)$. Next, let us recall other essential notations for this proof.

Firstly, we define $\Theta = \{(a, b, \nu) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+ : (\beta, W, a, b, \nu) \in \Omega\}$, and $\overline{\Theta}_\eta$ is an η-cover of Θ. Additionally, we also denote $\Delta := \{(\beta, W) \in \mathbb{R} \times \mathbb{R}^d : (\beta, W, a, b, \nu) \in \Omega\}$, and $\overline{\Delta}_\eta$ be an η-cover of Δ. Next, for each mixing measure $G = \sum_{i=1}^k \exp(\beta_i)\delta_{(W_i, a_i, b_i, \nu_i)} \in \mathcal{G}_k(\Omega)$, we denote $G' = \sum_{i=1}^k \exp(\beta_i)\delta_{(W_i, \overline{a}_i, \overline{b}_i, \overline{\nu}_i)}$ in which $(\overline{a}_i, \overline{b}_i, \overline{\nu}_i) \in \overline{\Theta}_\eta$ is the closest point to (a_i, b_i, ν_i) in this set for any $i \in [k]$. We also consider another mixing measure $\overline{G} := \sum_{i=1}^k \exp(\overline{\beta}_i)\delta_{(\overline{W}_i, \overline{a}_i, \overline{b}_i, \overline{\nu}_i)} \in \mathcal{G}_k(\Omega)$ where $(\overline{\beta}_i, \overline{W}_i) \in \overline{\Delta}_\eta$ is the closest point to (β_i, W_i) in this set for any $i \in [k]$.

Subsequently, we define

$$\mathcal{L} := \{\overline{p}_{\overline{G}} \in \mathcal{P}_k(\Omega) : (\overline{\beta}_i, \overline{W}_i) \in \overline{\Delta}_\eta,\ (\overline{a}_i, \overline{b}_i, \overline{\nu}_i) \in \overline{\Theta}_\eta\}.$$

We demonstrate that $\mathcal{L}$ is an η-cover of the metric space $(\mathcal{P}_k(\Omega), \|\cdot\|_1)$, that is, for any $\overline{p}_G \in \mathcal{P}_k(\Omega)$, there exists a density $\overline{p}_{\overline{G}} \in \mathcal{L}$ such that $\|\overline{p}_G - \overline{p}_{\overline{G}}\|_1 \leq \eta$. By the triangle inequality, we have

$$\|\overline{p}_G - \overline{p}_{\overline{G}}\|_1 \leq \|\overline{p}_G - \overline{p}_{G'}\|_1 + \|\overline{p}_{G'} - \overline{p}_{\overline{G}}\|_1. \tag{23}$$

From the formulation of G', we get that

$$\begin{aligned}\|\overline{p}_G - \overline{p}_{G'}\|_1 &\leq \sum_{i=1}^{k} \int_{\mathcal{X}\times\mathcal{Y}} \Big|f(Y|a_i^\top X + b_i, \nu_i) - f(Y|\overline{a}_i^\top X + \overline{b}_i, \overline{\nu}_i)\Big| \mathrm{d}(X,Y) \\ &\lesssim \sum_{i=1}^{k} \int_{\mathcal{X}\times\mathcal{Y}} \Big(\|a_i - \overline{a}_i\| + |b_i - \overline{b}_i| + |\nu_i - \overline{\nu}_i|\Big) \mathrm{d}(X,Y) \\ &\lesssim \eta \qquad (24)\end{aligned}$$

Based on inequalities in equations equation 23 and equation 24, it is sufficient to show that $\|\overline{p}_{G'} - \overline{p}_{\overline{G}}\|_1 \lesssim \eta$. For any $\overline{\tau} \in [\overline{q}]$, let us define

$$\overline{\mathcal{X}}_{\overline{\tau}} := \{x \in \mathcal{X} : -\|W_i - x\| \geq -\|W_{i'} - x\|,\ \forall i \in \{\overline{\tau}_1, \ldots, \overline{\tau}_{\overline{K}}\}, i' \in \{\overline{\tau}_{\overline{K}+1}, \ldots, \overline{\tau}_k\}\},$$
$$\mathcal{X}'_{\overline{\tau}} := \{x \in \mathcal{X} : -\|\overline{W}_i - x\| \geq -\|\overline{W}_{i'} - x\|,\ \forall i \in \{\overline{\tau}_1, \ldots, \overline{\tau}_{\overline{K}}\}, i' \in \{\overline{\tau}_{\overline{K}+1}, \ldots, \overline{\tau}_k\}\}.$$

Since the $\mathrm{softmax}$ function is differentiable, it is a Lipschitz function with some Lipschitz constant $L \geq 0$. Additionally, let us denote

$$\pi_{\overline{\tau}}(X) := \Big(-\|W_{\overline{\tau}_i} - x\| + \beta_{\overline{\tau}_i}^\top\Big)_{i=1}^{\overline{K}}; \qquad \overline{\pi}_{\overline{\tau}}(X) := \Big(-\|\overline{W}_{\overline{\tau}_i} - x\| + \overline{\beta}_{\overline{\tau}_i}^\top\Big)_{i=1}^{\overline{K}},$$

for any $\overline{K}$-element subset $\{\overline{\tau}_1, \ldots \overline{\tau}_K\}$ of $\{1, \ldots, k\}$. Then, we have

$$\begin{aligned}\|\mathrm{softmax}(\pi_{\overline{\tau}}(X)) - \mathrm{softmax}(\overline{\pi}_{\overline{\tau}}(X))\| &\leq L \cdot \|\pi_{\overline{\tau}}(X) - \overline{\pi}_{\overline{\tau}}(X)\| \\ &\leq L \cdot \sum_{i=1}^{\overline{K}} \Big(\Big|\|W_{\overline{\tau}_i} - X\| - \|\overline{W}_{\overline{\tau}_i} - X\|\Big| + |\beta_{\overline{\tau}_i} - \overline{\beta}_{\overline{\tau}_i}|\Big) \\ &\leq L \cdot \sum_{i=1}^{\overline{K}} \Big(\|W_{\overline{\tau}_i} - \overline{W}_{\overline{\tau}_i}\| + \beta_{\overline{\tau}_i} - \overline{\beta}_{\overline{\tau}_i}|\Big) \\ &\leq L \cdot \sum_{i=1}^{\overline{K}} \Big(\eta + \eta\Big) \\ &\lesssim \eta.\end{aligned}$$

By arguing similarly to the proof of Lemma I.2 in Appendix I, we receive that either $\overline{\mathcal{X}}_{\overline{\tau}} = \mathcal{X}'_{\overline{\tau}}$ or $\overline{\mathcal{X}}_{\overline{\tau}}$ has measure zero for any $\overline{\tau} \in [\overline{q}]$. As a result, we deduce that

$$\begin{aligned}\|\overline{p}_{G'} - \overline{p}_{G_*}\|_1 &\leq \sum_{\overline{\tau}=1}^{\overline{q}} \int_{\overline{\mathcal{X}}_{\overline{\tau}}\times\mathcal{Y}} |\overline{p}_{G'}(Y|X) - \overline{p}_{\overline{G}}(Y|X)| \mathrm{d}(X,Y) \\ &\leq \sum_{\overline{\tau}=1}^{\overline{q}} \int_{\overline{\mathcal{X}}_{\overline{\tau}}\times\mathcal{Y}} \sum_{i=1}^{\overline{K}} \Big|\mathrm{softmax}(\pi_{\overline{\tau}}(X)_i) - \mathrm{softmax}(\overline{\pi}_{\overline{\tau}}(X)_i)\Big| \cdot \Big|f(Y|\overline{a}_{\overline{\tau}_i}^\top X + \overline{b}_{\overline{\tau}_i}, \overline{\nu}_{\overline{\tau}_i})\Big| \mathrm{d}(X,Y) \\ &\lesssim \eta.\end{aligned}$$

Thus, $\mathcal{L}$ is an η-cover of the metric space $(\mathcal{P}_k(\Omega), \|\cdot\|_1)$, which implies that

$$N(\eta, \mathcal{P}_k(\Omega), \|\cdot\|_1) \lesssim |\overline{\Delta}_\eta| \times |\overline{\Theta}_\eta| \leq \mathcal{O}(\eta^{-(d+1)k}) \times \mathcal{O}(\eta^{-(d+2)k}) = \mathcal{O}(\eta^{-(2d+3)k}). \qquad (25)$$

Hence, we reach the conclusion that $\log N(\eta, \mathcal{P}_k(\Omega), \|\cdot\|_1) \lesssim \log(1/\eta)$.

H.4. Proof of Theorem 4.2

In order to establish the following Total Variation lower bound under the over-specified settings, i.e. when $k > k_*$ is unknown:

$$\mathbb{E}_X[V(\overline{p}_G(\cdot|X), p_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_2(G, G_*),$$

we need to prove two following inequalities:

- **Inequality A.** $\inf_{G\in\mathcal{G}_k(\Omega):\mathcal{D}_2(G,G_*)\leq\varepsilon'} \dfrac{\mathbb{E}_X[V(p_G(\cdot|X),p_{G_*}(\cdot|X))]}{\mathcal{D}_2(G,G_*)} > 0;$
- **Inequality B.** $\inf_{G\in\mathcal{G}_k(\Omega):\mathcal{D}_2(G,G_*)>\varepsilon'} \dfrac{\mathbb{E}_X[V(p_G(\cdot|X),p_{G_*}(\cdot|X))]}{\mathcal{D}_2(G,G_*)} > 0,$

for some constant $\varepsilon' > 0$. As the inequality B can be achieved in the same fashion as in Appendix H.2, we concentrate on showing the inequality A in this proof. For that purpose, it suffices to prove that

$$\lim_{\varepsilon\to 0} \inf_{G\in\mathcal{G}_k(\Theta):\mathcal{D}_2(G,G_*)\leq\varepsilon} \frac{\mathbb{E}_X[V(\overline{p}_G(\cdot|X),p_{G_*}(\cdot|X))]}{\mathcal{D}_2(G,G_*)} > 0. \tag{26}$$

Assume that the above claim does not hold true, then there exists a sequence of mixing measures $G_n := \sum_{i=1}^{k_n}\exp(\beta_i^n)\delta_{(W_i^n,a_i^n,b_i^n,\nu_i^n)} \in \mathcal{G}_k(\Omega)$ such that both $\mathcal{D}_2(G_n,G_*)$ and $\mathbb{E}_X[V(\overline{p}_{G_n}(\cdot|X),p_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n,G_*)$ go to zero as $n\to\infty$. Since $\mathcal{D}_2(G_n,G_*)\to 0$, we deduce that $\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)\to\exp(\beta_j^*)$ and $(W_i^n,a_i^n,b_i^n,\nu_i^n)\to(W_j^*,a_j^*,b_j^*,\nu_j^*)$ for all $i\in\mathcal{A}_j$ and $j\in[k_*]$. WLOG, we may assume that

$$\begin{aligned}\mathcal{D}_2(G_n,G_*) &= \sum_{\substack{j\in[K],\\|\mathcal{A}_j|>1}}\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)\Big[\|\Delta W_{ij}^n\|^2+\|\Delta a_{ij}^n\|^2+|\Delta b_{ij}^n|^{\bar{r}_j}+|\Delta\nu_{ij}^n|^{\frac{\bar{r}_j}{2}}\Big]\\ &+\sum_{\substack{j\in[K],\\|\mathcal{A}_j|=1}}\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)\Big[\|\Delta W_{ij}^n\|+\|\Delta a_{ij}^n\|+|\Delta b_{ij}^n|+|\Delta\nu_{ij}^n|\Big]+\sum_{j=1}^{K}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)-\exp(\beta_j^*)\Big|.\end{aligned} \tag{27}$$

Regarding the top-K selection in the conditional density p_{G_*}, we partition the covariate space $\mathcal{X}$ in a similar fashion to Appendix H.2. More specifically, we consider $q=\binom{k_*}{K}$ subsets $\{\tau_1,\ldots,\tau_K\}$ of $\{1,\ldots,k_*\}$ for any $\tau\in[q]$, and denote $\{\tau_{K+1},\ldots,\tau_{k_*}\}:=[k_*]\setminus\{\tau_1,\ldots,\tau_K\}$. Then, we define

$$\mathcal{X}_\tau^* := \Big\{x\in\mathcal{X}: -\|W_j^*-x\|\geq-\|W_{j'}^*-x\|, \forall j\in\{\tau_1,\ldots,\tau_K\}, j'\in\{\tau_{K+1},\ldots,\tau_{k_*}\}\Big\},$$

for any $\tau\in[q]$. On the other hand, we need to introduce a new partition method of the covariate space for the weight selection in the conditional density p_{G_n}. In particular, let $\overline{K}\in\mathbb{N}$ such that $\max_{\{\tau_j\}_{j=1}^K\subset[k_*]}\sum_{j=1}^K|\mathcal{A}_{\tau_j}|\leq\overline{K}\leq k$ and $\overline{q}:=\binom{k}{\overline{K}}$. Then, for any $\overline{\tau}\in[\overline{q}]$, we denote $(\overline{\tau}_1,\ldots,\overline{\tau}_k)$ as a subset of $[k]$ and $\{\overline{\tau}_{\overline{K}+1},\ldots,\overline{\tau}_k\}:=[k]\setminus\{\overline{\tau}_1,\ldots,\overline{\tau}_{\overline{K}}\}$. Additionally, we define

$$\mathcal{X}_{\overline{\tau}}^n := \Big\{x\in\mathcal{X}: -\|W_i^n-x\|\geq-\|W_{i'}^n-x\|, \forall i\in\{\overline{\tau}_1,\ldots,\overline{\tau}_{\overline{K}}\}, i'\in\{\overline{\tau}_{\overline{K}+1},\ldots,\overline{\tau}_k\}\Big\}.$$

Let $X\in\mathcal{X}_\tau^*$ for some $\tau\in[q]$ such that $\{\tau_1,\ldots,\tau_K\}=\{1,\ldots,K\}$. If $\{\overline{\tau}_1,\ldots\overline{\tau}_{\overline{K}}\}\neq\mathcal{A}_1\cup\ldots\cup\mathcal{A}_K$ for any $\overline{\tau}\in[\overline{q}]$, then $\mathbb{E}_X[V(\overline{p}_{G_n}(\cdot|X),p_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n,G_*)\not\to 0$ as n tends to infinity. This contradicts the fact that this term must approach zero. Therefore, we only need to consider the scenario when there exists $\overline{\tau}\in[\overline{q}]$ such that $\{\overline{\tau}_1,\ldots\overline{\tau}_{\overline{K}}\}=\mathcal{A}_1\cup\ldots\cup\mathcal{A}_K$. Recall that we have $(W_i^n,\beta_i^n)\to(W_j^*,\beta_j^*)$ as $n\to\infty$ for any $j\in[k_*]$ and $i\in\mathcal{A}_j$. Thus, for any arbitrarily small $\eta_j>0$, we have that $\|W_i^n-W_j^*\|\leq\eta_j$ and $|\beta_i^n-\beta_j^*|\leq\eta_j$ for sufficiently large n. Then, it follows from Lemma I.2 that $\mathcal{X}_\tau^*=\mathcal{X}_{\overline{\tau}}^n$ for sufficiently large n. This result indicates that $X\in\mathcal{X}_{\overline{\tau}}^n$.

Then, we can represent the conditional densities $p_{G_*}(Y|X)$ and $p_{G_n}(Y|X)$ for any sufficiently large n as follows:

$$\begin{aligned}p_{G_*}(Y|X) &= \sum_{j=1}^K\frac{\exp(-\|W_j^*-X\|+\beta_j^*)}{\sum_{j'=1}^K\exp(-\|W_{j'}^*-x\|+\beta_{0j'}^*)}\cdot f(Y|(a_j^*)^\top X+b_j^*,\nu_j^*),\\ \overline{p}_{G_n}(Y|X) &= \sum_{j=1}^K\sum_{i\in\mathcal{A}_j}\frac{\exp(-\|W_i^n-x\|+\beta_i^n)}{\sum_{j'=1}^K\sum_{i'\in\mathcal{A}_{j'}}\exp(-\|W_{i'}^n-x\|+\beta_{0i'}^n)}\cdot f(Y|(a_i^n)^\top X+b_i^n,\nu_i^n).\end{aligned}$$

Now, we reuse the three-step framework in Appendix H.2.

Stage 1 - Density decomposition:

Firstly, by abuse of notations, let us consider the quantity

$$Q_n := \Big[\sum_{j=1}^{K}\exp(-\|W_j^* - X\| + \beta_j^*)\Big]\cdot[\overline{p}_{G_n}(Y|X) - p_{G_*}(Y|X)].$$

Similar to Step 1 in Appendix H.2, we can express this term as

$$\begin{aligned}
Q_n &= \sum_{j=1}^{K}\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)\Big[F(Y|X;W_i^n,a_i^n,b_i^n,\nu_i^n) - F(Y|X;W_j^*,a_j^*,b_j^*,\nu_j^*)\Big]\\
&-\sum_{j=1}^{K}\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)\Big[H(Y|X;W_i^n) - H(Y|X;W_j^*)\Big]\\
&+\sum_{j=1}^{K}\Big[\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n) - \exp(\beta_j^*)\Big]\Big[F(Y|X;W_j^*,a_j^*,b_j^*,\nu_j^*) - H(Y|X,W_j^*)\Big]\\
&:= A_n - B_n + E_n,
\end{aligned}$$

Next, we proceed to decompose A_n based on the cardinality of the Voronoi cells as follows:

$$\begin{aligned}
A_n &= \sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)\Big[F(Y|X;W_i^n,a_i^n,b_i^n,\nu_i^n) - F(Y|X;W_j^*,a_j^*,b_j^*,\nu_j^*)\Big]\\
&+\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta_i^n)\Big[F(Y|X;W_i^n,a_i^n,b_i^n,\nu_i^n) - F(Y|X;W_j^*,a_j^*,b_j^*,\nu_j^*)\Big].
\end{aligned}$$

By applying the Taylor expansions of order 1 and $\bar{r}_j$ to the first and second terms of A_n, respectively, and following the derivation in equation 17, we get that

$$\begin{aligned}
A_n = &\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\sum_{|\alpha_1|=0}^{1}\sum_{|\alpha_2|=0}^{1-|\alpha_1|}\sum_{\eta=0}^{2(1-|\alpha_1|-|\alpha_2|)}\sum_{\substack{\alpha_3+2\alpha_4=\eta,\\0\leq\alpha_3+\alpha_4\leq 1-|\alpha_1|-|\alpha_2|}}\frac{\exp(\beta_i^n)}{2^{\alpha_4}\alpha!}\cdot(\Delta W_{ij}^n)^{\alpha_1}(\Delta a_{ij}^n)^{\alpha_2}(\Delta b_{ij}^n)^{\alpha_3}(\Delta \nu_{ij}^n)^{\alpha_4}\\
&\times X^{\alpha_2}\cdot\frac{\partial^{|\alpha_1|}g}{\partial W^{\alpha_1}}(X;W_j^*)\cdot\frac{\partial^{|\alpha_2|+\eta}f}{\partial h_1^{|\alpha_2|+\eta}}(Y|(a_j^*)^\top X + b_j^*,\nu_j^*) + R_3(X,Y)\\
+&\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\sum_{|\alpha_1|=0}^{\bar{r}_j}\sum_{|\alpha_2|=0}^{\bar{r}_j-|\alpha_1|}\sum_{\eta=0}^{2(\bar{r}_j-|\alpha_1|-|\alpha_2|)}\sum_{\substack{\alpha_3+2\alpha_4=\eta,\\0\leq\alpha_3+\alpha_4\leq \bar{r}_j-|\alpha_1|-|\alpha_2|}}\frac{\exp(\beta_i^n)}{2^{\alpha_4}\alpha!}\cdot(\Delta W_{ij}^n)^{\alpha_1}(\Delta a_{ij}^n)^{\alpha_2}(\Delta b_{ij}^n)^{\alpha_3}(\Delta \nu_{ij}^n)^{\alpha_4}\\
&\times X^{\alpha_2}\cdot\frac{\partial^{|\alpha_1|}g}{\partial W^{\alpha_1}}(X;W_j^*)\cdot\frac{\partial^{|\alpha_2|+\eta}f}{\partial h_1^{|\alpha_2|+\eta}}(Y|(a_j^*)^\top X + b_j^*,\nu_j^*) + R_4(X,Y)
\end{aligned}$$

where $R_i(X,Y)$ is a Taylor remainder such that $R_i(X,Y)/\mathcal{D}_2(G_n,G_*) \to 0$ as $n\to\infty$ for $i\in\{3,4\}$. Next, we apply the Taylor expansions of order 1 and 2 to the first and second terms of B_n, respectively, and following the derivation in equation 18, we get that

$$\begin{aligned}
B_n = &\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\sum_{|\gamma|=1}\frac{\exp(\beta_i^n)}{\gamma!}(\Delta W_{ij}^n)^{\gamma}\cdot\frac{\partial^{|\gamma|}g}{\partial W^{\gamma}}(X;W_j^*)p_{G_n}(Y|X) + R_5(X,Y)\\
&\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\sum_{|\gamma|=1}^{2}\frac{\exp(\beta_i^n)}{\gamma!}(\Delta W_{ij}^n)^{\gamma}\cdot\frac{\partial^{|\gamma|}g}{\partial W^{\gamma}}(X;W_j^*)p_{G_n}(Y|X) + R_6(X,Y),
\end{aligned}$$

where $R_5(X,Y)$ and $R_6(X,Y)$ are Taylor remainders such that their ratios over $\mathcal{D}_2(G_n,G_*)$ approach zero as $n\to\infty$. Subsequently, let us define

$$\begin{aligned} S^n_{j,\alpha_1,\alpha_2,\eta} &:= \sum_{i\in\mathcal{A}_j}\ \sum_{\substack{\alpha_3+2\alpha_4=\eta,\\ 0\leq\alpha_3+\alpha_4\leq\bar{r}_j-|\alpha_1|-|\alpha_2|}} \frac{\exp(\beta^n_i)}{2^{\alpha_4}\alpha!}\cdot(\Delta W^n_{ij})^{\alpha_1}(\Delta a^n_{ij})^{\alpha_2}(\Delta b^n_{ij})^{\alpha_3}(\Delta\nu^n_{ij})^{\alpha_4},\\ T^n_{j,\gamma} &:= \sum_{i\in\mathcal{A}_j}\frac{\exp(\beta^n_i)}{\gamma!}(\Delta W^n_{ij})^{\gamma}, \end{aligned}$$

for any $(\alpha_1,\alpha_2,\eta)\neq(\mathbf{0}_d,\mathbf{0}_d,0)$ and $\gamma\neq\mathbf{0}_d$. Otherwise, $S^n_{j,\mathbf{0}_d,\mathbf{0}_d,0}=T^n_{j,\mathbf{0}_d}:=\sum_{i\in\mathcal{A}_j}\exp(\beta^n_i)-\exp(\beta^*_j)$. As a consequence, it follows that

$$\begin{aligned} Q_n &= \sum_{j=1}^{K}\sum_{|\alpha_1|=0}^{\bar{r}_j}\sum_{|\alpha_2|=0}^{\bar{r}_j-|\alpha_1|}\ \sum_{\eta=0}^{2(\bar{r}_j-|\alpha_1|-|\alpha_2|)} S^n_{j,\alpha_1,\alpha_2,\eta}\cdot X^{\alpha_2}\cdot\frac{\partial^{|\alpha_1|}g}{\partial W^{\alpha_1}}(X;W^*_j)\cdot\frac{\partial^{|\alpha_2|+\eta}f}{\partial h_1^{|\alpha_2|+\eta}}(Y|(a^*_j)^\top X+b^*_j,\nu^*_j)\\ &+\sum_{j=1}^{K}\sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}} T^n_{j,\gamma}\cdot\frac{\partial^{|\gamma|}g}{\partial W^{\gamma}}(X;W^*_j)p_{G_n}(Y|X)+R_3(X,Y)+R_4(X,Y)+R_5(X,Y)+R_6(X,Y). \end{aligned} \tag{28}$$

Stage 2 - Non-vanishing coefficients:

In this step, we demonstrate that not all the ratios $S^n_{j,\alpha_1,\alpha_2,\eta}/\mathcal{D}_2(G_n,G_*)$ and $T^n_{j,\gamma}/\mathcal{D}_2(G_n,G_*)$ converge to zero as $n\to\infty$. Assume by contrary that all these terms go to zero. Then, by employing arguments for deriving equation 20 and equation 21, we get that

$$\begin{aligned} \frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\Big[\sum_{j=1}^{K}\Big|\sum_{i\in\mathcal{A}_j}\exp(\beta^n_i)-\exp(\beta^*_j)\Big|\\ +\sum_{j:|\mathcal{A}_j|=1}\sum_{i\in\mathcal{A}_j}\exp(\beta^n_i)\Big(\|\Delta W^n_{ij}\|+\|\Delta a^n_{ij}\|+|\Delta b^n_{ij}|+|\Delta\nu^n_{ij}|\Big)\Big]\to 0. \end{aligned}$$

Taking the summation of $\sum_{j:|\mathcal{A}_j|>1}\frac{|S^n_{j,\alpha_1,\alpha_2,\eta}|}{\mathcal{D}_2(G_n,G_*)}$ for all (α_1,α_2,η) where $\alpha_1\in\{2e_1,2e_2,\ldots,2e_d\}$, $\alpha_2=\mathbf{0}_d$ and $\eta=0$, we have

$$\frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta^n_i)\|\Delta W^n_{ij}\|^2\to 0.$$

Taking the summation of $\sum_{j:|\mathcal{A}_j|>1}\frac{|S^n_{j,\alpha_1,\alpha_2,\eta}|}{\mathcal{D}_2(G_n,G_*)}$ for all (α_1,α_2,η) where $\alpha_1=\mathbf{0}_d$, $\alpha_2\in\{2e_1,2e_2,\ldots,2e_d\}$ and $\eta=0$, we have

$$\frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta^n_i)\|\Delta a^n_{ij}\|^2\to 0.$$

Combine the above limit with the formulation of $\mathcal{D}_2(G_n,G_*)$ in equation 27, we have that

$$\frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\sum_{j:|\mathcal{A}_j|>1}\sum_{i\in\mathcal{A}_j}\exp(\beta^n_i)\Big(|\Delta b^n_{ij}|^{\bar{r}_j}+|\Delta\nu^n_{ij}|^{\frac{\bar{r}_j}{2}}\Big)\not\to 0.$$

This result implies that we can find some index $j'\in[K]:|\mathcal{A}_{j'}|>1$ that satisfies

$$\frac{1}{\mathcal{D}_2(G_n,G_*)}\cdot\sum_{i\in\mathcal{A}_{j'}}\exp(\beta^n_i)\Big(|\Delta b^n_{ij'}|^{\bar{r}_{j'}}+|\Delta\nu^n_{ij'}|^{\frac{\bar{r}_{j'}}{2}}\Big)\not\to 0.$$

For simplicity, we may assume that $j' = 1$. Since $S^n_{1,\mathbf{0}_d,\mathbf{0}_d,\eta}/\mathcal{D}_2(G_n, G_*)$ vanishes as $n \to \infty$ for any $1 \leq \eta \leq \bar{r}_j$, we divide this term by the left hand side of the above equation and achieve that

$$\frac{\sum_{i\in\mathcal{A}_1}\sum_{\substack{\alpha_3+2\alpha_4=\eta,\\ 1\leq\alpha_3+\alpha_4\leq\bar{r}_1}} \dfrac{\exp(\beta^n_i)}{2^{\alpha_4}\alpha!}(\Delta b^n_{i1})^{\alpha_3}(\Delta\nu^n_{i1})^{\alpha_4}}{\sum_{i\in\mathcal{A}_1}\exp(\beta^n_i)\Big(|\Delta b^n_{i1}|^{\bar{r}_1} + |\Delta\nu^n_{i1}|^{\frac{\bar{r}_1}{2}}\Big)} \to 0, \tag{29}$$

for any $1 \leq \eta \leq \bar{r}_1$.

Subsequently, we define $M_n := \max\{|\Delta b^n_{i1}|, |\Delta\nu^n_{i1}|^{1/2} : i \in \mathcal{A}_1\}$ and $\pi_n := \max\{\exp(\beta^n_i) : i \in \mathcal{A}_1\}$. As a result, the sequence $\exp(\beta^n_i)/\pi_n$ is bounded, which indicates that we can substitute it with its subsequence that admits a positive limit $z^2_{5i} := \lim_{n\to\infty}\exp(\beta^n_i)/\pi_n$. Therefore, at least one among the limits z^2_{5i} equals to one. Furthermore, we also denote

$$(\Delta b^n_{i1})/M_n \to z_{3i},\ (\Delta\nu^n_{i1})/(2M_n) \to z_{4i}.$$

From the above definition, it follows that at least one among the limits z_{3i} and z_{4i} equals to either 1 or -1. By dividing both the numerator and the denominator of the term in equation 29 by $\pi_n M^\eta_n$, we arrive at the following system of polynomial equations:

$$\sum_{i\in\mathcal{A}_1}\sum_{\substack{\alpha_3+2\alpha_4=\eta,\\ 1\leq\alpha_3+\alpha_4\leq\bar{r}_1}} \frac{z^2_{5i}\ z^{\alpha_3}_{3i}\ z^{\alpha_4}_{4i}}{\alpha_3!\ \alpha_4!} = 0,$$

for all $1 \leq \eta \leq \bar{r}_1$. Nevertheless, from the definition of $\bar{r}_1$, we know that the above system does not admit any non-trivial solutions, which is a contradiction. Consequently, not all the ratios $S^n_{j,\alpha_1,\alpha_2,\eta}/\mathcal{D}_2(G_n, G_*)$ and $T^n_{j,\gamma}/\mathcal{D}_2(G_n, G_*)$ tend to zero as $n \to \infty$.

Stage 3 - Fatou's contradiction:

Recall that $\mathbb{E}_X[V(\overline{p}_{G_n}(\cdot|X), p_{G_*}(\cdot|X))]/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$. Then, by applying the Fatou's lemma, we get

$$0 = \lim_{n\to\infty}\frac{\mathbb{E}_X[V(\overline{p}_{G_n}(\cdot|X), p_{G_*}(\cdot|X))]}{\mathcal{D}_2(G_n, G_*)} \geq \frac{1}{2}\cdot\int\liminf_{n\to\infty}\frac{|\overline{p}_{G_n}(Y|X) - p_{G_*}(Y|X)|}{\mathcal{D}_2(G_n, G_*)}\mathrm{d}(X, Y),$$

which implies that $|\overline{p}_{G_n}(Y|X) - p_{G_*}(Y|X)|/\mathcal{D}_2(G_n, G_*) \to 0$ as $n \to \infty$ for almost surely (X, Y).

Next, we define m_n as the maximum of the absolute values of $S^n_{j,\alpha_1,\alpha_2,\eta}/\mathcal{D}_2(G_n, G_*)$. It follows from Step 2 that $1/m_n \not\to \infty$. Moreover, by arguing in the same way as in Step 3 in Appendix H.2, we receive that

$$Q_n/[m_n\mathcal{D}_2(G_n, G_*)] \to 0 \tag{30}$$

as $n \to \infty$. By abuse of notations, let us denote

$$S^n_{j,\alpha_1,\alpha_2,\eta}/[m_n\mathcal{D}_2(G_n, G_*)] \to \xi_{j,\alpha_1,\alpha_2,\eta},$$
$$T^n_{j,\gamma}/[m_n\mathcal{D}_2(G_n, G_*)] \to \kappa_{j,\gamma}.$$

Here, at least one among $\xi_{j,\alpha_1,\alpha_2,\eta}, \kappa_{j,\gamma}$ is non-zero. Then, by putting the results in equation 28 and equation 30 together, we get

$$\sum_{j=1}^{K}\sum_{|\alpha_1|=0}^{\bar{r}_j}\sum_{|\alpha_2|=0}^{\bar{r}_j-|\alpha_1|}\sum_{\eta=0}^{2(\bar{r}_j-|\alpha_1|-|\alpha_2|)} \xi_{j,\alpha_1,\alpha_2,\eta}\cdot X^{\alpha_2}\cdot\frac{\partial^{|\alpha_1|}g}{\partial W^{\alpha_1}}(X; W^*_j)\cdot\frac{\partial^{|\alpha_2|+\eta}f}{\partial h_1^{|\alpha_2|+\eta}}(Y|(a^*_j)^\top X + b^*_j, \nu^*_j)$$
$$+\sum_{j=1}^{K}\sum_{|\gamma|=0}^{1+\mathbf{1}_{\{|\mathcal{A}_j|>1\}}}\kappa_{j,\gamma}\cdot\frac{\partial^{|\gamma|}g}{\partial W^\gamma}(X; W^*_j)p_{G_n}(Y|X) = 0.$$

Arguing in a similar fashion as in Step 3 of Appendix H.2, we obtain that $\xi_{j,\alpha_1,\alpha_2,\eta} = \kappa_{j,\gamma} = 0$ for any $j \in [K]$, $0 \leq |\alpha_1| + |\alpha_2| \leq 2\bar{r}_j$, $0 \leq \eta \leq 2(\bar{r}_j - |\alpha_1| - |\alpha_2|)$ and $0 \leq |\gamma| \leq 1 + \mathbf{1}_{\{|\mathcal{A}_j|>1\}}$. This contradicts the fact that at least one among them is non-zero. Hence, the proof is completed.

I. Auxiliary Results

Lemma I.1. *For any $i \in [k_*]$, let $W_i, W_i^* \in \mathbb{R}^d$ such that $\|W_i - W_i^*\| \leq \eta_i$ for some sufficiently small $\eta_i > 0$. Then, for any $\tau \in [q]$, unless the set $\mathcal{X}_\tau^*$ has measure zero, we obtain that $\mathcal{X}_\tau^* = \mathcal{X}_\tau$ where*

$$\mathcal{X}_\tau := \{x \in \mathcal{X} : -\|W_i - x\| \geq -\|W_{i'} - x\|, \forall i \in \{\tau_1, \ldots, \tau_K\}, i' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}\}.$$

Proof of Lemma I.1. Let $\eta_i = M_i \varepsilon$, where ε is some fixed positive constant and M_i will be chosen later. For an arbitrary $\tau \in [q]$, since $\mathcal{X}$ and Ω are bounded sets, there exists some constant $c_\tau^* \geq 0$ such that

$$\min_{x,i,i'} \Big[-\|W_i^* - x\| + \|W_{i'}^* - x\| \Big] = c_\tau^* \varepsilon, \tag{31}$$

where the minimum is subject to $x \in \mathcal{X}_\tau^*, i \in \{\tau_1, \ldots, \tau_K\}$ and $i' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}$. We will point out that $c_\tau^* > 0$. Assume by contrary that $c_\tau^* = 0$. For $x \in \mathcal{X}_\tau^*$, we may assume for any $1 \leq i < j \leq k_*$ that

$$-\|W_{\tau_i}^* - x\| \geq -\|W_{\tau_j}^* - x\|.$$

Since $c_\tau^* = 0$, it follows from equation 31 that $\|W_{\tau_K}^* - x\| = \|W_{\tau_{K+1}}^* - x\|$. In other words, $\mathcal{X}_\tau^*$ is a subset of $\mathcal{Z} := \{x \in \mathcal{X} : \|W_{\tau_K}^* - x\| = \|W_{\tau_{K+1}}^* - x\|\}$. As $W_{\tau_K}^* \neq W_{\tau_{K+1}}^*$ and the distribution of X is continuous, it follows that the set $\mathcal{Z}$ has measure zero. Since $\mathcal{X}_\tau^* \subseteq \mathcal{Z}$, we can conclude that $\mathcal{X}_\tau^*$ also has measure zero, which contradicts the hypothesis of Lemma I.1. Thus, we must have $c_\tau^* > 0$.

As $\mathcal{X}$ is a bounded set, we assume that $\|x\| \leq B$ for any $x \in \mathcal{X}$. Let $x \in \mathcal{X}_\tau^*$, then we have for any $i \in \{\tau_1, \ldots, \tau_K\}$ and $i' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}$ that

$$\begin{aligned} -\|W_i - x\| &= -\|W_i - W_i^*\| - \|W_i^* - x\| \\ &\geq -M_i\varepsilon - \|W_{i'}^* - x\| + c_\tau^*\varepsilon \\ &= -M_i\varepsilon + c_\tau^*\varepsilon - \|W_{i'}^* - W_{i'}\| - \|W_{i'} - x\| \\ &\geq -2M_i\varepsilon + c_\tau^*\varepsilon - \|W_{i'} - x\|. \end{aligned}$$

By setting $M_i \leq \dfrac{c_\tau^*}{2}$, we get that $x \in \mathcal{X}_\tau$, which means that $\mathcal{X}_\tau^* \subseteq \mathcal{X}_\tau$. Similarly, assume that there exists some constant $c_\tau \geq 0$ that satisfies

$$\min_{x,i,i'} \Big[-\|W_i - x\| + \|W_{i'} - x\| \Big] = c_\tau^* \varepsilon.$$

Here, the above minimum is subject to $x \in \mathcal{X}_\tau$, $i \in \{\tau_1, \ldots, \tau_K\}$ and $i' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}$. If $M_i \leq \dfrac{c_\tau}{2}$, then we also receive that $\mathcal{X}_\tau \subseteq \mathcal{X}_\tau^*$. Hence, if we set $M_i = \dfrac{1}{2}\min\{c_\tau^*, c_\tau\}$, we reach the conclusion that $\mathcal{X}_\tau^* = \mathcal{X}_\tau$. □

Lemma I.2. *For any $j \in [k_*]$ and $i \in \mathcal{A}_j$, let $W_i, W_j^* \in \mathbb{R}^d$ that satisfy $\|W_i - W_j^*\| \leq \eta_j$ for some sufficiently small $\eta_j > 0$. Additionally, for $\max_{\{\tau_j\}_{j=1}^K \subseteq [k_*]} \sum_{j=1}^K |\mathcal{A}_{\tau_j}| \leq \overline{K} \leq k$, we assume that there exist $\tau \in [q]$ and $\overline{\tau} \in [\overline{q}]$ such that $\{\overline{\tau}_1, \ldots, \overline{\tau}_{\overline{K}}\} = \mathcal{A}_{\tau_1} \cup \ldots \cup \mathcal{A}_{\tau_K}$. Then, if the set $\mathcal{X}_\tau^*$ does not have measure zero, we achieve that $\mathcal{X}_\tau^* = \overline{\mathcal{X}}_{\overline{\tau}}$.*

Proof of Lemma I.2. Let $\eta_j = M_j\varepsilon$, where ε is some fixed positive constant and M_i will be chosen later. As $\mathcal{X}$ and Ω are bounded sets, we can find some constant $c_\tau^* \geq 0$ such that

$$\min_{x,j,j'} \Big[-\|W_j^* - x\| + \|W_{j'}^* - x\| \Big] = c_\tau^* \varepsilon,$$

where the above minimum is subject to $x \in \mathcal{X}_\tau^*, j \in \{\tau_1, \ldots, \tau_K\}$ and $j' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}$. By arguing similarly to the proof of Lemma I.1, we deduce that $c_\tau^* > 0$.

Let $x \in \mathcal{X}^*_\tau$ and $\overline{\tau} \in [\overline{q}]$ such that $\{\overline{\tau}_1, \ldots, \overline{\tau}_{\overline{K}}\} = \mathcal{A}_{\tau_1} \cup \ldots \cup \mathcal{A}_{\tau_K}$. Then, for any $i \in \{\overline{\tau}_1, \ldots, \overline{\tau}_{\overline{K}}\}$ and $i' \in \{\overline{\tau}_{\overline{K}+1}, \ldots, \overline{\tau}_k\}$, we have that

$$\begin{aligned}
-\|W_i - x\| &= -\|W_i - W^*_j\| - \|W^*_j - x\| \\
&\geq -M_i\varepsilon - \|W^*_{j'} - x\| + c^*_\tau\varepsilon \\
&= -M_i\varepsilon + c^*_\tau\varepsilon - \|W^*_{j'} - W_{j'}\| - \|W_{i'} - x\| \\
&\geq -2M_i\varepsilon + c^*_\tau\varepsilon - \|W_{i'} - x\|,
\end{aligned}$$

where $j \in \{\tau_1, \ldots, \tau_K\}$ and $j' \in \{\tau_{K+1}, \ldots, \tau_{k_*}\}$ such that $i \in \mathcal{A}_j$ and $i' \in \mathcal{A}_{j'}$. If $M_j \leq \dfrac{c^*_\tau}{2}$, then we get that $x \in \mathcal{X}_{\overline{\tau}}$, which leads to $\mathcal{X}^*_\tau \subseteq \overline{\mathcal{X}}_{\overline{\tau}}$.

Analogously, assume that there exists some constant $c_\tau \geq 0$ such that

$$\min_{x,j,j'} \left[(\beta^*_{1j})^\top x - (\beta^*_{1j'})^\top x\right] = c^*_\tau\varepsilon,$$

where the minimum is subject to $x \in \overline{\mathcal{X}}_{\overline{\tau}}$, $i \in \{\overline{\tau}_1, \ldots, \overline{\tau}_{\overline{K}}\}$ and $i' \in \{\overline{\tau}_{\overline{K}+1}, \ldots, \overline{\tau}_k\}$. Then, if $M_j \leq \dfrac{c_\tau}{2}$, then we receive that $\overline{\mathcal{X}}_{\overline{\tau}} \subseteq \mathcal{X}^*_\tau$. As a consequence, by setting $M_j = \dfrac{1}{2}\min\{c^*_\tau, c_\tau\}$, we achieve the conclusion that $\overline{\mathcal{X}}_{\overline{\tau}} = \mathcal{X}^*_\tau$. □

Lemma I.3. *For any mixing measures G and G_* in $\mathcal{G}_k(\Theta)$ that satisfy $p_G(Y|X) = p_{G_*}(Y|X)$ for almost surely $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, we have that $G \equiv G_*$.*

Proof of Lemma I.3. First, we assume that two mixing measures G and G_* take the following forms: $G = \sum_{i=1}^k \exp(\beta_i)\delta_{(W_i, a_i, b_i, \nu_i)}$ and $G_* = \sum_{i=1}^{k_*} \exp(\beta^*_i)\delta_{(W^*_i, a^*_i, b^*_i, \nu^*_i)}$. Recall that $p_G(Y|X) = p_{G_*}(Y|X)$ for almost surely (X, Y), then we have

$$\begin{aligned}
&\sum_{i=1}^k \mathrm{softmax}(\mathrm{TopK}(-\|W_i - X\|, K; \beta_i)) \cdot f(Y|a_i^\top X + b_i, \nu_i) \\
&\qquad = \sum_{i=1}^{k_*} \mathrm{softmax}(\mathrm{TopK}(-\|W^*_i - X\|, K; \beta^*_i)) \cdot f(Y|(a^*_i)^\top + b^*_i, \nu^*_i).
\end{aligned} \tag{32}$$

Due to the identifiability of the location-scale Gaussian mixtures (Teicher, 1960; 1961; 1963), we get that $k = k_*$ and

$$\Big\{\mathrm{softmax}(\mathrm{TopK}(-\|W_i - X\|, K; \beta_i)) : i \in [k]\Big\} \equiv \Big\{\mathrm{softmax}(\mathrm{TopK}(-\|W^*_i - X\|, K; \beta^*_i)) : i \in [k]\Big\},$$

for almost surely X. WLOG, we may assume that

$$\mathrm{softmax}(\mathrm{TopK}(-\|W_i - X\|, K; \beta_i)) = \mathrm{softmax}(\mathrm{TopK}(-\|W^*_i - X\|, K; \beta^*_i)), \tag{33}$$

for almost surely X for any $i \in [k]$. Since the softmax function is invariant to translations, it follows from equation 33 that $W_i = W^*_i$ and $\beta_i = \beta^*_i + v_0$ for some $v_0 \in \mathbb{R}$. Notably, from the assumption of the model, we have $\beta_{0k} = \beta'_{0k} = 0$, which implies that $v_0 = 0$. As a result, we obtain that $\beta_i = \beta^*_i$ for any $i \in [k_*]$.

Let us consider $X \in \mathcal{X}_\tau$ where $\tau \in [q]$ such that $\{\tau_1, \ldots, \tau_K\} = \{1, \ldots, K\}$. Then, equation 32 can be rewritten as

$$\sum_{i=1}^K \exp(\beta_i)\exp(-\|W_i - X\|) f(Y|a_i^\top X + b_i, \nu_i) = \sum_{i=1}^K \exp(\beta_i)\exp(-\|W_i - X\|) f(Y|(a^*_i)^\top X + b^*_i, \nu^*_i), \tag{34}$$

for almost surely (X, Y). Next, we denote $J_1, J_2, \ldots, J_m$ as a partition of the index set $[k]$, where $m \leq k$, such that $\exp(\beta_i) = \exp(\beta_{i'})$ for any $i, i' \in J_j$ and $j \in [m]$. On the other hand, when i and i' do not belong to the same set J_j, we let $\exp(\beta_i) \neq \exp(\beta_{i'})$. Thus, we can reformulate equation 34 as

$$\sum_{j=1}^m \sum_{i \in J_j} \exp(\beta_i)\exp(-\|W_i - X\|) f(Y|a_i^\top X + b_i, \nu_i) = \sum_{j=1}^m \sum_{i \in J_j} \exp(\beta_i)\exp(-\|W_i - X\|) f(Y|(a^*_i)^\top X + b^*_i, \nu^*_i),$$

for almost surely (X, Y). This results leads to $\{((a_i)^\top X + b_i, \nu_i) : i \in J_j\} \equiv \{((a_i^*)^\top X + b_i^*, \nu_i^*) : i \in J_j\}$, for almost surely X for any $j \in [m]$. Therefore, we have

$$\{(a_i, b_i, \nu_i) : i \in J_j\} \equiv \{(a_i^*, b_i^*, \nu_i^*) : i \in J_j\},$$

for any $j \in [m]$. As a consequence,

$$G = \sum_{j=1}^{m} \sum_{i \in J_j} \exp(\beta_i) \delta_{(W_i, a_i, b_i, \nu_i)} = \sum_{j=1}^{m} \sum_{i \in J_j} \exp(\beta_i') \delta_{(W_i^*, a_i^*, b_i^*, \nu_i^*)} = G_*.$$

Hence, we reach the conclusion of this lemma. □