
Supplementary Material for Learning to Route: Per-Sample Adaptive Routing for Multimodal Multitask Prediction

Anonymous Author(s)

Affiliation

Address

email

1 Code reproducing the results in the paper is available as open source under the MIT license on
2 GitHub: [link will appear here; for now see code provided in accompanying code folder].

3 A Real-world data collection and usage

4 **Realistic Synthetic Data.** The clinical outcomes described here were not used directly for model
5 training or evaluation in the main manuscript. Instead, they served as the basis of a highly realistic
6 synthetic dataset that captures key distributional and structural characteristics of the real data. Al-
7 though synthetic, enough real details about patients are preserved in the data that it cannot be shared
8 publicly, as it would be considered identifiable, HIPAA-protected data.

9 **Ethical oversight.** All procedures were reviewed and approved by the Institutional Review Board
10 (IRB #*****) at the authors' institution.¹ Informed consent was obtained from all participants.
11 The original dataset contains protected health information and is governed by U.S. HIPAA regulations;
12 it therefore cannot be shared outside the institution or reproduced in any form that might compromise
13 patient confidentiality.

14 Intervention and Data Collection

- 15 • **Context.** A four-session, remote, skills-based psychotherapy program was launched at a
16 large academic medical center during the first peak of the COVID-19 pandemic (March
17 2020–April 2021).
- 18 • **Participants.** $N = 534$ health care workers (HCWs) self-referred for emotional support.
19 Role categories (percentages approximated): 35.2% nursing, 24.3% patient support, 22.8%
20 administrative, 13.8% medical trainees/faculty, 2.4% facilities, 1.3% family members. 70%
21 were on-site (frontline); 19% worked remotely; 11% unspecified.
- 22 • **Clinicians.** Sixty-seven trained providers (licensed psychologists, psychiatrists, social
23 workers, and supervised trainees) delivered a total of 1,423 telehealth sessions.
- 24 • **Measures.**
 - 25 – Patient Health Questionnaire–9 (PHQ-9) and Generalized Anxiety Disorder–7 (GAD-7)
26 at sessions 1 and 4.
 - 27 – PHQ-4 at sessions 2 and 3 for interim symptom tracking.
 - 28 – Columbia Suicide Severity Rating Scale (C-SSRS) at intake and as needed for suicide
29 risk.
- 30 • **Safety.** Participants with severe symptoms or safety concerns were referred to appropriate
31 emergency or long-term psychiatric care.

¹The full IRB number and study name will be unblinded upon acceptance.

32 A linear mixed-effects model (random intercept and slope per subject; fixed effect of time) showed
 33 significant reduction in overall symptom burden:

$$\text{PHQ-4}_{\text{baseline}} = 5.65 \pm 2.95 \rightarrow \text{PHQ-4}_{\text{final}} = 3.32 \pm 2.46, \\ F_{3,823} = 109.23, p < .001, \eta^2_{\text{partial}} = 0.27.$$

34 For participants with clinically elevated symptoms at baseline ($\text{PHQ-4} \geq 6$), the effect size was
 35 stronger ($\eta^2_{\text{partial}} = 0.46$). Response rates were 42% on GAD-7 and 43% on PHQ-9 ($\geq 50\%$ symptom
 36 reduction).

37 B Data-Driven (Empirical) Synthetic Data

38 To support controlled experimentation and model probing, we generated synthetic samples that reflect
 39 “clean” examples from the original dataset. These samples were designed to preserve strong predictive
 40 signal across both GAD and PHQ outcomes. The process is structured to maintain cross-modal
 41 coherence between structured numerical features and free-text clinical notes. The generation process
 42 proceeds in three stages: (1) Sampling structured numerical features from a smoothed approximation
 43 of high-confidence empirical distributions. (2) Generating text conditionally based on these numerical
 44 features using a large language model (LLM) with custom prompting. (3) Filtering generated samples
 45 using our trained multimodal multitask model to retain only those with low predictive uncertainty
 46 and high task-specific confidence.

47 B.1 Synthetic Numerical Data Generation

48 We generated synthetic numerical data using three methods, each designed to preserve the statistical
 49 structure of the original dataset.

50 **Gaussian Synthesis.** We estimated the mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ from
 51 the original dataset. To ensure numerical stability, a small constant 10^{-6} was added to the diagonal
 52 of Σ . Synthetic samples were then drawn from a multivariate normal distribution:

$$\mathbf{X}_{\text{synthetic}} \sim \mathcal{N}(\mu, \Sigma), \quad \mathbf{X}_{\text{synthetic}} \in \mathbb{R}^{n_{\text{samples}} \times d}.$$

53 **Copula-Based Synthesis.** We applied a rank-based transformation to map each feature to a standard
 54 normal distribution. The empirical correlation matrix was computed on the transformed data. Syn-
 55 thetic samples were drawn from a multivariate normal distribution with this correlation structure and
 56 subsequently mapped back to the original marginal distributions using a quantile transform. This
 57 method preserved nonlinear dependencies among features.

58 **Kernel Density Estimation (KDE) Synthesis.** Continuous features were standardized using
 59 `StandardScaler`. A Gaussian kernel density estimator was fitted with bandwidth $h = n^{-1/(d+4)}$,
 60 where n is the number of samples and d is the feature dimensionality. New samples were drawn using
 61 KDE-based resampling. Binary features (e.g., *dissociate*, *anger*, *fear_contam*) were thresholded
 62 post-generation:

$$\mathbf{X}_{\text{synthetic}}[\text{binary}] = (\mathbf{X}_{\text{synthetic}}[\text{binary}] > 0.5).astype(int).$$

63 For each method, we generated 200 synthetic samples while maintaining the original class distribution
 64 of PHQ-9 and GAD-7 binary outcomes. The synthetic datasets closely matched the original statistical
 65 characteristics, with a mean absolute difference in feature correlation of less than 0.1 and a KL
 66 divergence in class distribution of less than 0.05.

67 Synthetic Text Data Generation

68 To generate synthetic text data that reflect realistic psychotherapy session notes, we employed a
 69 prompt-based generation pipeline using small, open-source large language models (LLMs). The goal
 70 was to create natural language samples that are consistent with the patterns observed in the original
 71 dataset, while maintaining data privacy and avoiding memorization of sensitive content.

72 **Numerical-to-Text Conversion.** We first transformed structured numerical features into short natural
 73 language descriptions. For each synthetic numerical instance $x \in \mathbb{R}^d$, we constructed a template-
 74 based summary containing key symptom indicators and severity scores (e.g., PHQ-9, GAD-7). An
 75 example of this intermediate representation is:

76 The patient reported a PHQ-9 score of 15 and a GAD-7 score of 13. They en-
 77 dorsed symptoms such as dissociation and irritability, with no signs of fear of
 78 contamination.

79 **LLM-Based Natural Language Expansion.** We used an instruction-tuned language model (Flan-T5
 80 or Phi-2) to expand the structured summaries into fluent and contextually appropriate psychotherapy
 81 notes. Each prompt followed the format:

82 Patient data: PHQ-9 = 15, GAD-7 = 13, symptoms = [dissociation, irritabil-
 83 ity]. Write a brief therapist note summarizing the patient’s emotional state and
 84 challenges.

85 The model generated coherent text such as:

86 The patient presented with moderate symptoms of depression and anxiety, including
 87 dissociative experiences and heightened irritability. They expressed difficulty
 88 managing emotional stressors and reported low energy and trouble sleeping.

89 **Post-Processing and Filtering.** We generated one therapist-style note per synthetic numerical input,
 90 resulting in 200 synthetic text samples. To ensure linguistic diversity and clinical plausibility, we
 91 applied basic heuristics to filter out degenerate outputs (e.g., overly repetitive or off-topic content).
 92 The vocabulary and sentence structure were qualitatively consistent with those in the original data,
 93 and generated texts maintained semantic alignment with the associated synthetic numerical features.

94 **Privacy Considerations.** All models were run locally without external API calls to ensure HIPAA
 95 compliance. We used only open-access, instruction-tuned models with small memory footprints
 96 (Phi-2), which allowed controlled offline generation and ensured that no real patient data were
 97 exposed or used during synthesis.

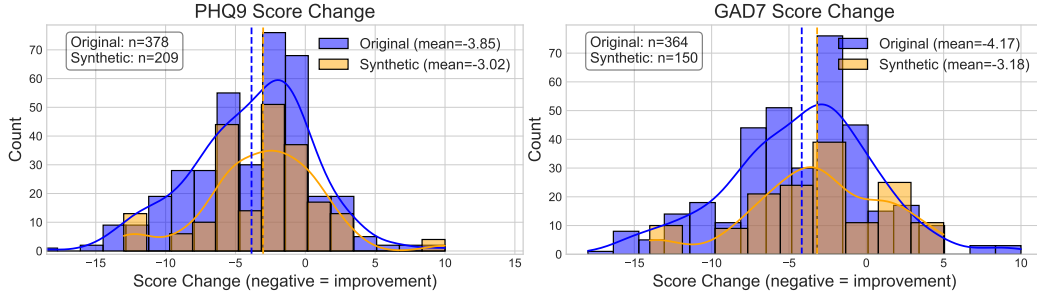


Figure 1: Distribution of synthetic and original data

98 C Equation-Driven (Analytical) Synthetic Data

99 Our synthetic benchmark is designed to rigorously test input-dependent routing in multitask, multi-
 100 modal regression. Each sample consists of two modalities, a numeric vector and a text vector, and
 101 two regression targets. Both modalities contribute to both tasks, but the degree and nature of their
 102 informativeness is heterogeneous and input-dependent, simulating real-world complexity.

103 **Input Features.** For each sample, we generate:

- 104 • Numeric features $\mathbf{x}^{(\text{num})} \in \mathbb{R}^{d_{\text{num}}}$, sampled as $\mathcal{N}(0, I)$.
- 105 • Text features $\mathbf{x}^{(\text{text})} \in \mathbb{R}^{d_{\text{text}}}$, also sampled as $\mathcal{N}(0, I)$.

106 We use $d_{\text{num}} = d_{\text{text}} = 16$ unless otherwise specified.

107 **Random Fourier Feature Maps.** To introduce nonlinear, cross-modal dependencies, we use
 108 random Fourier feature (RFF) maps:

$$\phi(\mathbf{x}^{(\text{text})}) = \sqrt{\frac{2}{D}} \cos(W_\phi \mathbf{x}^{(\text{text})} + b_\phi),$$

$$\psi(\mathbf{x}^{(\text{num})}) = \sqrt{\frac{2}{D}} \cos(W_\psi \mathbf{x}^{(\text{num})} + b_\psi),$$

109 where $W_\phi, W_\psi \in \mathbb{R}^{D \times 16}$ have entries drawn from $\mathcal{N}(0, 1)$, $b_\phi, b_\psi \in \mathbb{R}^D$ are drawn uniformly from
 110 $[0, 2\pi]$, and $D = 32$.

111 **Target Construction.** For each sample, we generate two targets:

$$y_1 = \alpha_1^\top \mathbf{x}^{(\text{num})} + \beta_1^\top \phi(\mathbf{x}^{(\text{text})}) + \gamma_1 \cdot \sin(\omega_1^\top \mathbf{x}^{(\text{num})}) + \epsilon_1,$$

$$y_2 = \alpha_2^\top \mathbf{x}^{(\text{text})} + \beta_2^\top \psi(\mathbf{x}^{(\text{num})}) + \gamma_2 \cdot \cos(\omega_2^\top \mathbf{x}^{(\text{text})}) + \epsilon_2,$$

112 where:

- 113 • $\alpha_k, \beta_k \sim \mathcal{N}(0, I)$ (dimensions match their arguments)
- 114 • $\omega_k \sim \mathcal{N}(0, I)$ (dimension 16)
- 115 • $\gamma_k \sim \text{Uniform}[0.5, 1.5]$
- 116 • $\epsilon_k \sim \mathcal{N}(0, 0.1^2)$.

117 All parameters are independently sampled for each trial, and fixed for all samples within a trial.

118 **Design Rationale.** This construction ensures:

- 119 • Both modalities are relevant to both tasks, but in different, nonlinear, and cross-modal ways.
- 120 • The use of RFFs simulates learned embeddings and increases the complexity of the mapping.
- 121 • Sinusoidal nonlinearities further challenge the model, requiring it to capture nontrivial
- 122 dependencies.
- 123 • The setup allows for controlled ablations (e.g., by zeroing coefficients or removing nonlinear
- 124 terms).

125 **Implementation Notes.**

- 126 • All random seeds are fixed for reproducibility.
- 127 • The code for data generation is provided in the supplementary repository.
- 128 • The synthetic dataset can be easily extended to more modalities or tasks by following the
- 129 same recipe.

Algorithm 1 Synthetic Data Generation

- 1: Sample $\mathbf{x}^{(\text{num})}, \mathbf{x}^{(\text{text})} \sim \mathcal{N}(0, I_{16})$
 - 2: Compute $\phi(\mathbf{x}^{(\text{text})}), \psi(\mathbf{x}^{(\text{num})})$ via RFFs
 - 3: Sample $\alpha_k, \beta_k, \omega_k, \gamma_k$ as above
 - 4: Compute y_1, y_2 as above, add noise ϵ_k
-

```

130 def rff(x, W, b):
131     return np.sqrt(2 / W.shape[0]) * np.cos(W @ x + b)
132
133 x_num = np.random.randn(16)
134 x_text = np.random.randn(16)
135 W_phi, b_phi = np.random.randn(32, 16), np.random.uniform(0, 2*np.pi, 32)
136 W_psi, b_psi = np.random.randn(32, 16), np.random.uniform(0, 2*np.pi, 32)
137 phi_x_text = rff(x_text, W_phi, b_phi)
138 psi_x_num = rff(x_num, W_psi, b_psi)
139 # ... sample coefficients and compute y1, y2 as above

```

140 C.1 Scenario 1: Sinusoidal/Cosine, Both Modalities

```
141 Number of samples: 1000 (train), 1000 (test)
142 Feature dimensions: d_num = 16, d_text = 16, D (RFF output) = 32
143
144 Feature maps:
145  $\phi(x_{\text{text}}) = \sqrt{2/32} * \cos(W_{\phi} x_{\text{text}} + b_{\phi})$ 
146  $\psi(x_{\text{num}}) = \sqrt{2/32} * \cos(W_{\psi} x_{\text{num}} + b_{\psi})$ 
147  $W_{\phi}, W_{\psi} \sim N(0, 1), b_{\phi}, b_{\psi} \sim \text{Uniform}[0, 2\pi]$ 
148
149 Target equations:
150  $y_1 = \alpha_1^T x_{\text{num}} + \beta_1^T \phi(x_{\text{text}}) + \gamma_1 * \sin(\omega_1^T x_{\text{num}}) + \epsilon_1$ 
151  $y_2 = \alpha_2^T x_{\text{text}} + \beta_2^T \psi(x_{\text{num}}) + \gamma_2 * \cos(\omega_2^T x_{\text{text}}) + \epsilon_2$ 
152
153 Parameter distributions:
154 -  $\alpha_k, \beta_k \sim N(0, I)$ 
155 -  $\omega_k \sim N(0, I)$ 
156 -  $\gamma_k \sim \text{Uniform}[0.5, 1.5]$ 
157 -  $\epsilon_k \sim N(0, 0.1^2)$ 
```

158 In this scenario, as expected, the model learns that using MTL yields the best performance (see Figure
159 2).

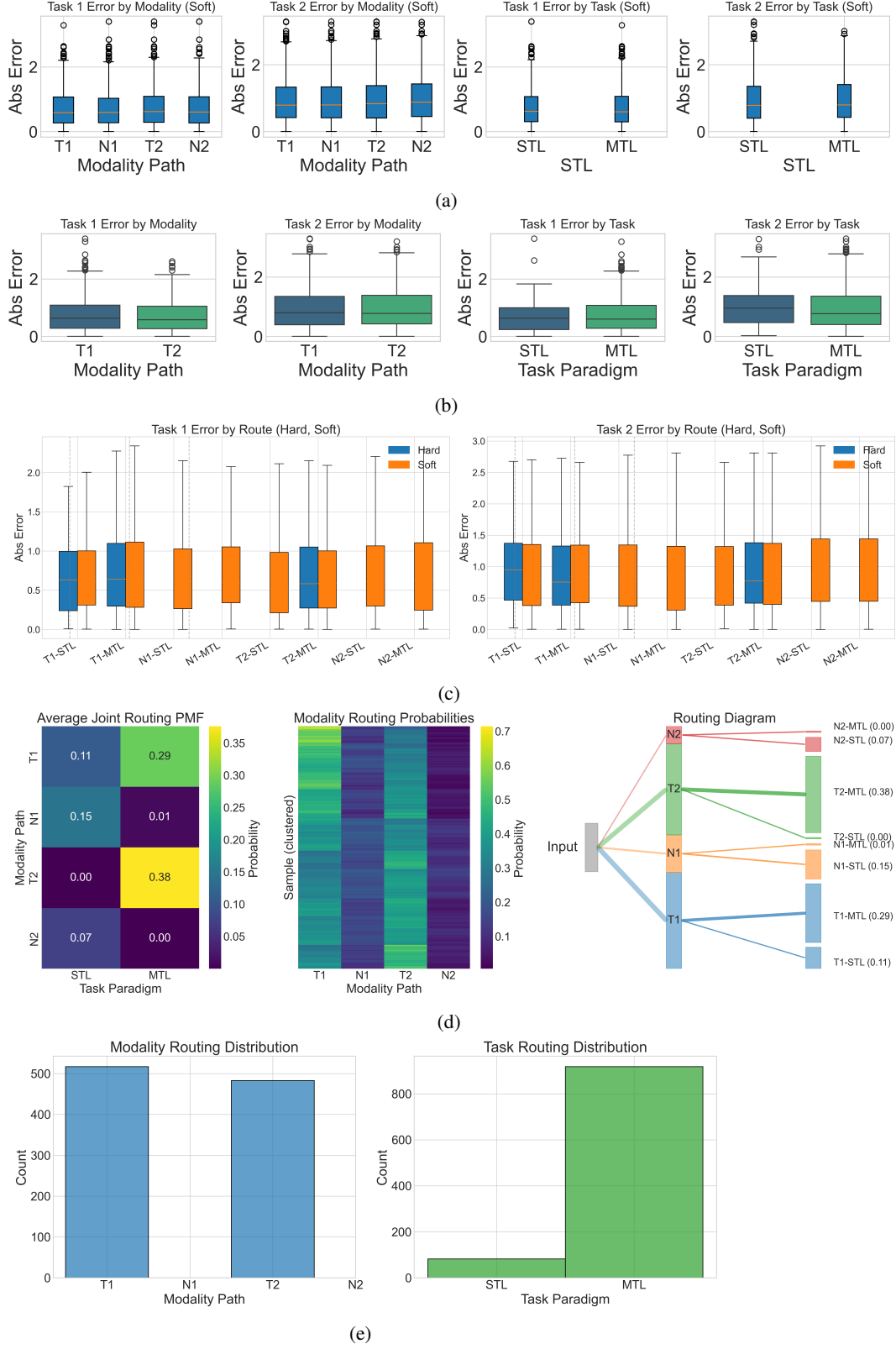
160 C.2 Scenario 2: STL Preferred

```
161 Number of samples: 1000 (train), 1000 (test)
162 Feature dimensions: d_num = 16, d_text = 16, D (RFF output) = 32
163
164 Feature maps:
165  $\phi(x_{\text{text}}) = \sqrt{2/32} * \cos(W_{\phi} x_{\text{text}} + b_{\phi})$ 
166  $\psi(x_{\text{num}}) = \sqrt{2/32} * \cos(W_{\psi} x_{\text{num}} + b_{\psi})$ 
167  $W_{\phi}, W_{\psi} \sim N(0, 1), b_{\phi}, b_{\psi} \sim \text{Uniform}[0, 2\pi]$ 
168
169 Target equations:
170  $y_1 = \alpha_1^T x_{\text{num}} + \gamma_1 * \sin(\omega_1^T x_{\text{num}}) + \epsilon_1$ 
171  $y_2 = \alpha_2^T x_{\text{text}} + \gamma_2 * \cos(\omega_2^T x_{\text{text}}) + \epsilon_2$ 
172
173 Parameter distributions:
174 -  $\alpha_k \sim N(0, I)$ 
175 -  $\omega_k \sim N(0, I)$ 
176 -  $\gamma_k \sim \text{Uniform}[0.5, 1.5]$ 
177 -  $\epsilon_k \sim N(0, 0.1^2)$ 
```

178 In this scenario, each task is generated from a single modality: y_1 depends only on the numeric
179 features and their nonlinear transformation, while y_2 depends only on the textual features and their
180 nonlinear transformation. Since each task is generated independently from its own modality, there
181 is no shared information or benefit to learning the tasks jointly. As a result, the model learns that
182 treating each task separately by using STL yields the best performance (see Figure 3).

183 C.3 Scenario 3: Fusion-Dominant Routing

```
184 Number of samples: 1000 (train), 1000 (test)
185 Feature dimensions: d_num = 16, d_text = 16, D (RFF output) = 32
186
187 Feature maps:
188  $\phi(x_{\text{text}}) = \sqrt{2/32} * \cos(W_{\phi} x_{\text{text}} + b_{\phi})$ 
189  $\psi(x_{\text{num}}) = \sqrt{2/32} * \cos(W_{\psi} x_{\text{num}} + b_{\psi})$ 
190  $W_{\phi}, W_{\psi} \sim N(0, 1), b_{\phi}, b_{\psi} \sim \text{Uniform}[0, 2\pi]$ 
191
```



```

192 Target equations:
193  $y_1 = \alpha_1^T x_{\text{num}} + \beta_1^T \phi(x_{\text{text}}) + \epsilon_1$ 
194  $y_2 = \alpha_2^T x_{\text{text}} + \beta_2^T \psi(x_{\text{num}}) + \epsilon_2$ 
195
196 Parameter distributions:
197 -  $\alpha_k, \beta_k \sim N(0, I)$ 
198 -  $\omega_k \sim N(0, I)$ 
199 -  $\epsilon_k \sim N(0, 0.1^2)$ 

```

200 The results in Figure 4 show that the model prefers the T2 (fusion) modality path, especially in
 201 combination with the MTL (multi-task learning) paradigm. The N1 and N2 paths are rarely or never
 202 used for MTL, indicating that the model has learned to avoid these routes in favor of more effective
 203 ones.

204 **D Broader Impact**

205 This work introduces a flexible machine learning framework for adaptively routing data through
 206 multimodal and multitask pathways, with primary application to psychological outcome prediction.
 207 By enabling models to select personalized computation paths based on both input availability and task
 208 structure, this approach has the potential to improve prediction accuracy and robustness in real-world
 209 settings where data is heterogeneous and incomplete.

210 While our evaluation is framed in the context of mental health prediction, the methodology is broadly
 211 applicable to domains such as clinical decision support, education, and human-centered AI systems
 212 where structured and unstructured inputs coexist and multiple outcomes must be considered jointly.

213 At the same time, predictive models in healthcare and mental health raise significant ethical concerns.
 214 These include the risk of reinforcing biases present in clinical documentation, the opacity of model
 215 decisions, and the potential for overreliance on algorithmic outputs in high-stakes scenarios. Our
 216 model attempts to mitigate some of these risks by producing interpretable routing decisions, which
 217 may offer insight into modality usefulness and task interactions. Nonetheless, interpretability and
 218 fairness should be further studied before deployment.

219 Our work uses synthetic data and carefully preprocessed clinical data to demonstrate technical
 220 contributions, and does not aim to inform clinical decisions directly. Future use of this method
 221 in real-world applications must be coupled with appropriate clinical validation, governance, and
 222 safeguards to ensure equitable, transparent, and accountable outcomes.

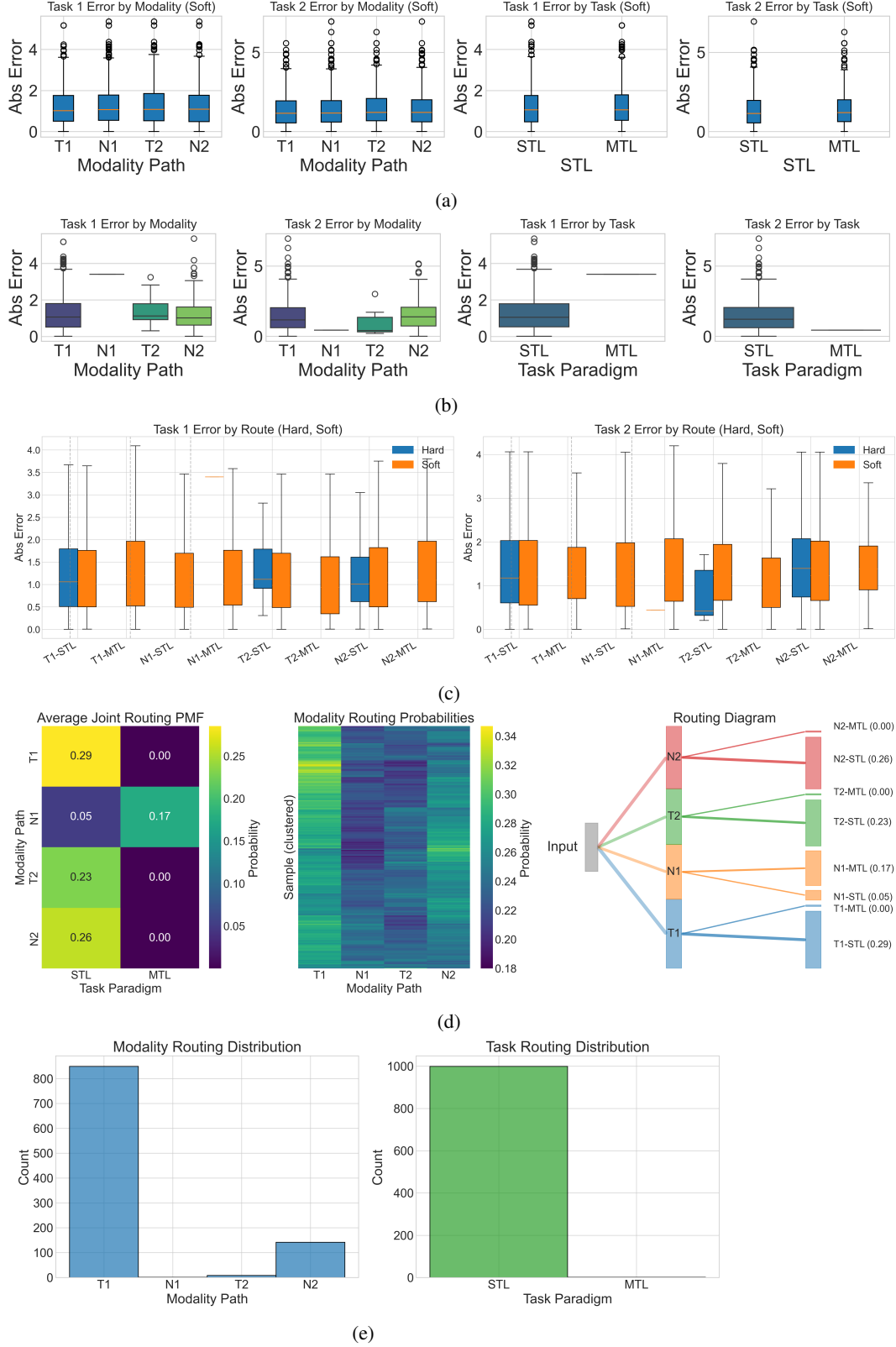


Figure 3: **Scenario 2: STL Preferred** (a) Absolute error by route (soft routing, weighted by probabilities). (b) Absolute error by route (hard routing, based on most probable path). (c) Comparison of absolute errors for each route: hard vs. soft routing (note that in hard routing, boxes do not appear in every case). (d) Summary: joint routing PMF, clustered heatmap, and routing Sankey diagram. (e) Distribution of selected modality and task by the router.

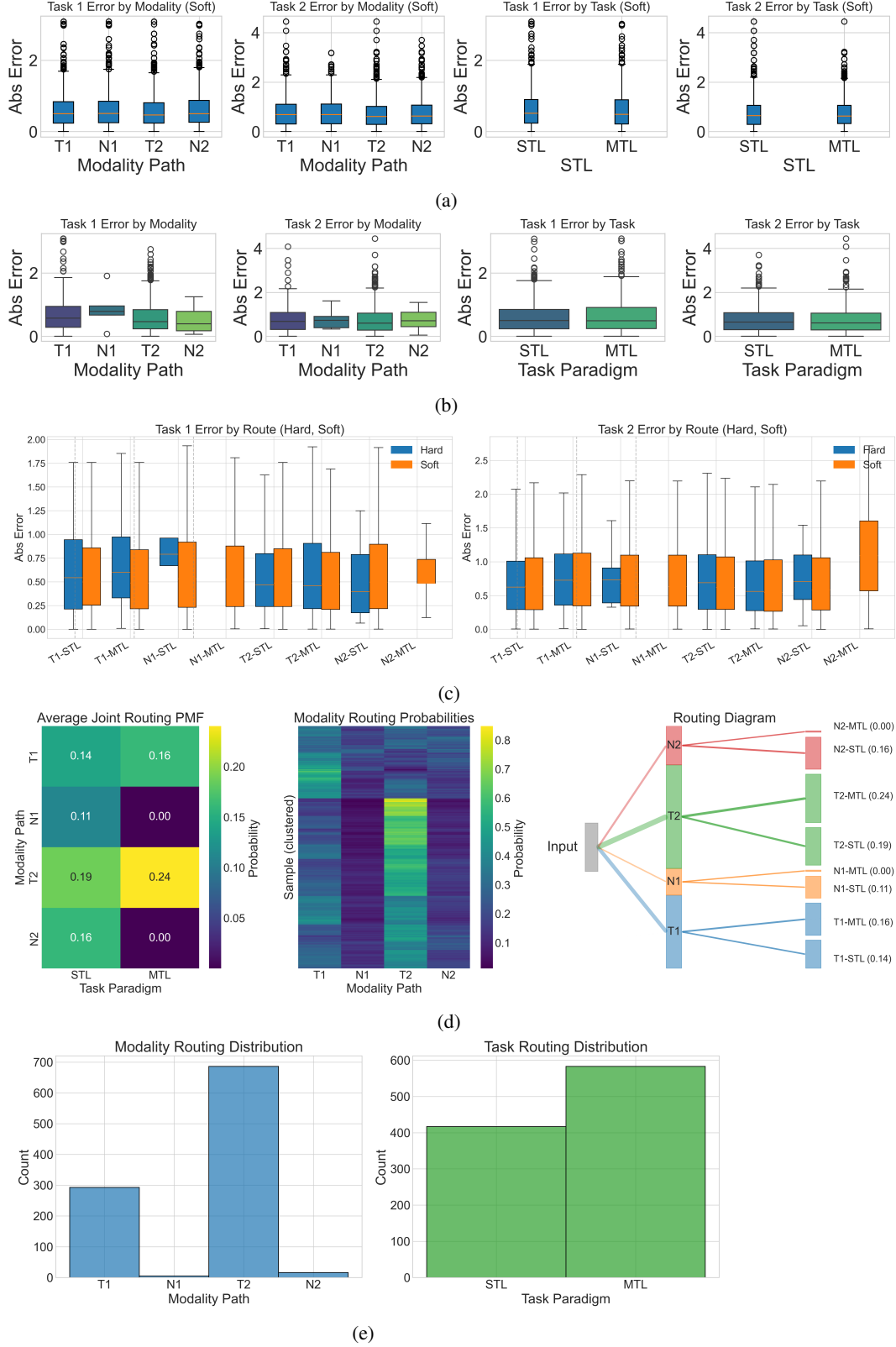


Figure 4: **Scenario 3: Fusion-Dominant Routing** (a) Absolute error by route (soft routing, weighted by probabilities). (b) Absolute error by route (hard routing, based on most probable path). (c) Comparison of absolute errors for each route: hard vs. soft routing (note that in hard routing, boxes do not appear in every case). (d) Summary: joint routing PMF, clustered heatmap, and routing Sankey diagram. (e) Distribution of selected modality and task paradigm by the router.