

# GRAPHPL: LEVERAGING GNN FOR EFFICIENT AND ROBUST MODALITIES IMPUTATION IN PATCHWORK LEARNING

Xingjian Hu<sup>\*</sup>, Zuoyu Yan<sup>†</sup>, Jianhua Zhu<sup>\*</sup>, Liangcai Gao<sup>\*</sup>, Fei Wang<sup>†</sup>, Tengfei Ma<sup>\*</sup>

<sup>\*</sup> Wangxuan Institute of Computer Technology, Peking University, Beijing  
E-mail: {huxingjian, glc}@pku.edu.cn

<sup>†</sup> Weill Cornell Medicine, Cornell University, New York

<sup>\*</sup> Biomedical Informatics, Stony Brook University, New York

## ABSTRACT

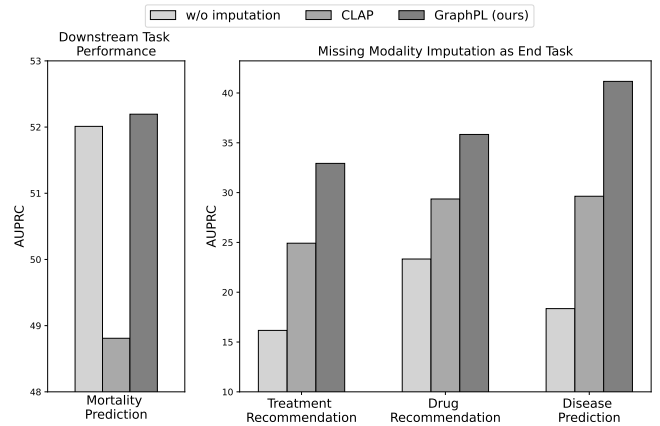
Current research on distributed multi-modal learning typically assumes that clients can access complete information across all modalities, which may not hold in practice. In this paper, we explore patchwork learning, in which the modalities available to different clients vary, and the objective is to impute the missing modalities for each client in an unsupervised manner. Existing methods are shown not to fully utilize the modality information as they tend to rely on only a subset of the observed modalities. To address this issue, we propose GraphPL, which combines graph neural networks with patchwork learning to flexibly integrate all observed modalities and remains robust with noisy inputs. Experimental results show that GraphPL achieves SOTA performance on benchmark datasets. Our results on real-world distributed electronic health record dataset show GraphPL learns strong downstream features and enables tasks like disease prediction via superior modality imputation.

**Index Terms**— graph neural network, missing modality imputation, electronic health record, patchwork learning, health informatics

## 1. INTRODUCTION

Multi-modal learning [3] extracting rich information from various modalities. Diverse modalities offer complementary analytical advantages via varied content, structure, and expression [4]. It typically uses paired multi-modal data, with inter-modal correspondences per sample, helping models learn cross-modal associations for more comprehensive understanding.

However, in real-world scenarios, multiple modalities may be unavailable, and due to privacy concerns, data sharing is not possible [5, 6]. Rajendran *et al.* [7] introduce **patchwork learning** to address the situations where clients have incomplete and differing observed modalities while preserving data privacy. This scenario highlights the importance of missing modality imputation, which involves using the

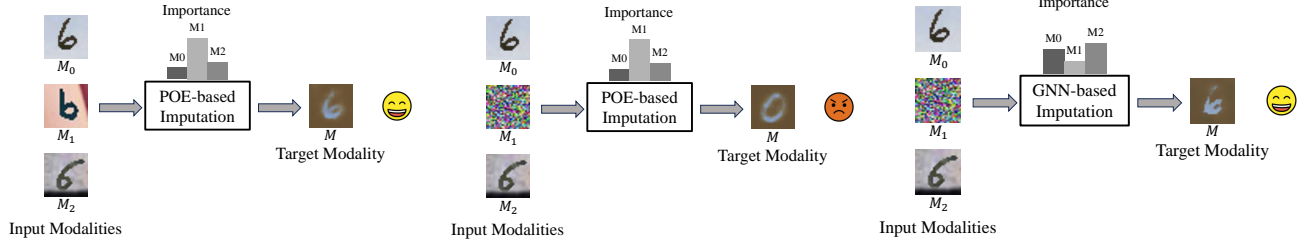


**Fig. 1:** The performance of different methods on the real-world EHR dataset eICU [1]. By incorporating imputation tasks and leveraging GNNs to mitigate the modality collapse issue identified in CLAP [2], GraphPL achieves superior performance in both tasks.

observed modalities to perform cross-generation of the missing modalities within the same data sample. As illustrated in Fig. 1, **imputation is essential for two primary reasons**: it enhances the quality of multi-modal representations by providing more comprehensive information, and it maintains data integrity, sometimes serving as an end task itself (e.g., imputing diagnostic information in healthcare) [8, 9].

A limitation of existing methods [2] is **modality collapse** [10, 11], where over-reliance on partial modality information leads to suboptimal performance and robustness, as shown in Fig. 2. These methods often use product-of-experts (POE) [12] for fusion, which can exacerbate modality collapse [13].

To address the modality collapse issue, we propose **GraphPL**, a framework leveraging Graph Neural Networks (GNNs) to dynamically fuse observed modalities. It constructs a modality-modality graph with nodes representing modalities and edges denoting interactions. A message-passing mechanism aggregates information across modali-



**Fig. 2:** Existing POE-based method faced the **modality collapse issue**, due to its over-reliance on specific modality ( $M_1$  here). In contrast, our GNN-based method merge the input information adaptively, exhibiting stronger robustness.

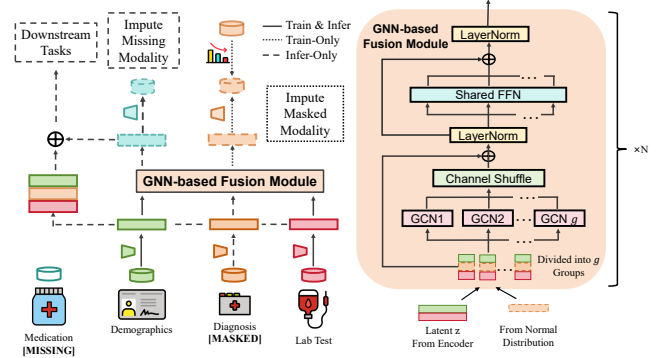
ties, adapting to varying input modalities and noise. Experiment results demonstrate that GraphPL is **effective**: it improves generation quality by **8.8%**, **13.9%**, and **4.8%** on simulated benchmark datasets, and achieves gains of **11.5%**, **6.5%**, and **8.0%** in disease diagnosis, drug recommendation, treatment recommendation, respectively, on the real-world EHR dataset eICU. Our code is available at: <https://github.com/LumionHXJ/GraphPL>. In summary, our key contributions are as follows:

- To address the modality collapse issue, we propose GraphPL, which leverages GNNs for modality fusion, allowing for flexible integration of input information.
- Experimental results demonstrate that GraphPL an average **9.2%** improvement in imputation tasks across multiple simulated benchmark datasets as well as on the real-world distributed EHR dataset eICU. On eICU, it attains an average improvement of **8.7%** over baseline methods in imputation tasks, confirming its practical effectiveness.
- GraphPL exhibits robustness under varying noise conditions, effectively mitigating modality collapse and enhancing overall model stability.

## 2. GRAPHPL

GraphPL is designed for the distributed multi-modal patchwork learning scenario, where clients have incomplete and diverse visible modalities, and data sharing is prohibited by privacy constraints.

The pipeline of GraphPL is illustrated in Fig. 3. We implement an individual VAE [14] for each modality. During the forward pass, the input modalities are categorized into two groups: **conditional modalities**  $x_{\text{cond}}$  (available as input) and **target modalities**  $x_{\text{target}}$  (those to be imputed). The conditional modalities are first encoded into latent representations  $z$  using their respective encoders. These features are then fused by a GNN-based fusion module to generate the features for target modalities. The resulting features can either be passed to decoders to reconstruct target modalities or concatenated with the features from conditional modalities for use in downstream tasks.



**Fig. 3:** GraphPL's overall pipeline (here, the client lacks the medication modality): During training, it masks one modality as target modality and imputes it via other observed modalities as conditional ones. During inference, it uses all observed modalities as conditional modalities to predict the missing modality's features for either imputation or downstream tasks.

### 2.1. GNN-based Fusion Module

Previous methods [2, 13, 15, 16] use POE to fuse features of conditional modalities, the process can be overly influenced by partial distributions [13], facing the modality collapse issue. To overcome the limitations, we employ a GNNs for dynamic and adaptive modality fusion. Each conditional modality is represented as a node in a complete modality-modality graph, enabling full interaction across all available inputs. When imputing a target modality, it is introduced as a virtual node connected to all conditional modalities, allowing customized fusion for each target.

In the GNN fusion module, node embeddings for conditional modalities are derived from their VAE encoders, while target modalities are initialized via sampling from a standard normal distribution. Our fusion module consists of multiple identical modules, each composed of a grouped GCN-Conv [17] followed by two-layer FFN. Channel shuffling [18] is applied after each GCNConv to promote cross-group information flow. LayerNorm stabilizes training and ensures output features remain compatible with the scale of encoder-derived features, facilitating effective modality imputation.

**Table 1:** Generation quality (%) and representation quality (%) on multiple benchmark datasets.

Method	PolyMNIST		MST		Quad-CelebA	
	GQ (↑)	RQ (↑)	GQ (↑)	RQ (↑)	GQ (↑)	RQ (↑)
MVAE	9.7 $\pm$ 0.7	96.8 $\pm$ 0.5	14.3 $\pm$ 1.5	93.4 $\pm$ 2.0	47.8 $\pm$ 0.5	54.7 $\pm$ 0.6
MMVAE	42.8 $\pm$ 4.5	98.2 $\pm$ 0.5	55.2 $\pm$ 5.9	98.6 $\pm$ 0.3	56.1 $\pm$ 0.7	56.4 $\pm$ 0.7
MoPoE	48.8 $\pm$ 1.7	98.4 $\pm$ 0.3	54.9 $\pm$ 3.3	98.4 $\pm$ 0.4	58.9 $\pm$ 0.3	56.0 $\pm$ 0.7
CLAP	46.8 $\pm$ 3.1	96.9 $\pm$ 0.7	51.7 $\pm$ 6.7	98.7 $\pm$ 0.4	60.1 $\pm$ 0.6	57.3 $\pm$ 0.1
GraphPL	<b>57.6</b> $\pm$ 5.0	<b>98.5</b> $\pm$ 0.2	<b>69.1</b> $\pm$ 6.3	<b>99.2</b> $\pm$ 0.4	<b>64.9</b> $\pm$ 0.7	<b>62.5</b> $\pm$ 0.6

## 2.2. Training

**Local Rounds.** During local round on client  $C^i$ , GraphPL iteratively treats each observed modality as the target modality, with rest observed modalities serving as conditional modalities. Unlike methods optimizing the ELBO [14] of  $\log p(x_{\text{target}})$ , we directly model the conditional log-likelihood:

$$\log p(x_{\text{target}}|x_{\text{cond}}) = \mathbb{E}_{p(z|x_{\text{cond}})} \log p(x_{\text{target}}|z). \quad (1)$$

Since direct sampling  $z \sim p(z|x_{\text{cond}})$  is infeasible, we introduce the single-modality reconstruction task and approximate it by VAE encoder output  $q(z|x_{\text{cond}}) \approx p(z|x_{\text{cond}})$  [14]. The imputation loss is thus defined as:

$$\mathcal{L}_{\text{impute}} = -\mathbb{E}_{q(z|x_{\text{cond}})} \log p(x_{\text{target}}|z). \quad (2)$$

Additionally, the single-modality reconstruction task is incorporated for each observed modality using its dedicated VAE, with loss similar to  $\beta$ -VAE [19]. The total local loss combines both objectives:

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{\text{impute}} + \lambda \mathcal{L}_{\text{single}}, \quad (3)$$

with  $\lambda$  as a balancing hyperparameter.

**Global Rounds.** After certain local rounds, each client uploads the encoder and decoder of its observed modalities, as well as the GNN parameters, to the server. The server then uses FedAvg [20] to update the global model parameters.

## 2.3. Inference

During the inference phase, we serve two purposes: **imputing missing modalities and supporting downstream tasks**. For imputation, the missing modalities features output by the GNN module are passed to corresponding decoders trained by clients that possess the corresponding modality and shared via communication. For downstream tasks, the latent features of both observed and imputed modalities are concatenated to form an enriched representation for task-specific models.

# 3. EXPERIMENTS

## 3.1. Experiment Settings

We evaluate GraphPL and baselines (MVAE [15], MMVAE [13], MoPoE [16], CLAP [2]) on non-distributed benchmarks (PolyMNIST, MST, Quad-CelebA [16]) and real-world

Electronic Health Record (EHR) dataset eICU [1]. The patchwork is constructed by randomly dropping some modalities on each client independently. Training uses FedAvg [20] for distributed learning and the Adam optimizer with batch size 256. Each experiment runs 5 times (same patchwork setup, different data splits), with results reported as means and STDs.

**Evaluation Metrics:** 1) **Generation Quality (GQ):** For missing modality imputation, pre-trained classifiers check if imputed class labels match ground truth. 2) **Representation Quality (RQ):** For downstream tasks, a logistic regression classifier is trained on each client using concatenated features; quality is measured via local test set classification accuracy.

## 3.2. Experiments on Non-distributed Benchmarks

Following [2], we conduct experiments on same non-distributed benchmarks. Due to the number of modalities in the Bimodal CelebA dataset is limited, following Nair *et al.* [21], we extracted Canny edges from face images as sketches and generated face segmentation maps using FaRL [22], resulting in the **Quad-CelebA** dataset. To construct statistical heterogeneity for these non-distributed benchmarks, for PolyMNIST and MST, we perform class-imbalanced sampling to ensure that each client contains only some classes; for Quad-CelebA, we assign different proportions of male samples to each client. All of these experiments use 5 clients.

The results are shown in Table 1. MVAE fails as it uses all observed modalities for training and cannot learn missing modality imputation. MoPoE performs better by using all observed modality combinations but incurs higher computational cost. Our method, GraphPL, achieves optimal performance due to its use of GNNs for flexible modality integration, demonstrating its effectiveness.

## 3.3. Experiments on Distributed Dataset eICU

To validate real-world performance, we use the distributed EHR dataset eICU [1], with each hospital as a client. Data preprocessing follows MUSE [11], with input modalities: demographics, diagnosis (ICD-9 codes), treatment, medication (HICL codes), lab values, and vital signals. For diagnosis, treatment, and medication, many of whose categories are lim-

**Table 2:** The experimental results on the eICU dataset, with 10, 20, and 50 hospitals participating, respectively. *Treatment*, *Medication*, and *Diagnosis* represent the three missing modality imputation tasks of treatment recommendation, drug recommendation, and disease diagnosis, respectively. *Mortality* represents the downstream task of mortality prediction.

Hospitals	Method	GQ		RQ	
		Treatment	Medication	Diagnosis	Mortality
10	MVAE	21.5 $\pm$ 2.1	25.2 $\pm$ 2.2	25.6 $\pm$ 12.0	52.0 $\pm$ 2.1
	MMVAE	16.3 $\pm$ 0.8	23.5 $\pm$ 1.5	18.8 $\pm$ 6.6	44.6 $\pm$ 3.2
	MoPoE	26.4 $\pm$ 3.9	29.8 $\pm$ 3.0	30.0 $\pm$ 12.0	47.1 $\pm$ 4.3
	CLAP	24.9 $\pm$ 3.7	29.4 $\pm$ 2.3	29.6 $\pm$ 11.9	48.8 $\pm$ 4.3
	GraphPL	<b>32.9</b> $\pm$ 8.1	<b>35.8</b> $\pm$ 2.5	<b>41.2</b> $\pm$ 14.3	<b>52.2</b> $\pm$ 3.5
20	MVAE	19.9 $\pm$ 2.3	23.8 $\pm$ 1.7	20.2 $\pm$ 0.7	49.7 $\pm$ 3.3
	MMVAE	15.4 $\pm$ 0.9	22.6 $\pm$ 1.5	16.9 $\pm$ 0.6	43.6 $\pm$ 3.6
	MoPoE	22.8 $\pm$ 2.9	26.8 $\pm$ 2.3	24.0 $\pm$ 0.9	46.4 $\pm$ 3.6
	CLAP	22.5 $\pm$ 2.6	26.6 $\pm$ 2.1	24.8 $\pm$ 0.9	47.4 $\pm$ 3.3
	GraphPL	<b>29.2</b> $\pm$ 3.9	<b>32.6</b> $\pm$ 2.4	<b>32.1</b> $\pm$ 1.3	<b>49.9</b> $\pm$ 4.0
50	MVAE	17.2 $\pm$ 0.5	22.7 $\pm$ 0.6	15.6 $\pm$ 0.7	<b>45.9</b> $\pm$ 1.9
	MMVAE	14.4 $\pm$ 1.5	22.1 $\pm$ 0.5	14.7 $\pm$ 0.8	39.9 $\pm$ 2.1
	MoPoE	19.0 $\pm$ 1.1	24.3 $\pm$ 0.6	18.7 $\pm$ 1.2	42.5 $\pm$ 1.5
	CLAP	21.9 $\pm$ 1.6	25.2 $\pm$ 0.6	20.6 $\pm$ 1.1	43.9 $\pm$ 1.3
	GraphPL	<b>25.4</b> $\pm$ 2.6	<b>29.9</b> $\pm$ 0.9	<b>26.6</b> $\pm$ 1.9	45.8 $\pm$ 2.1

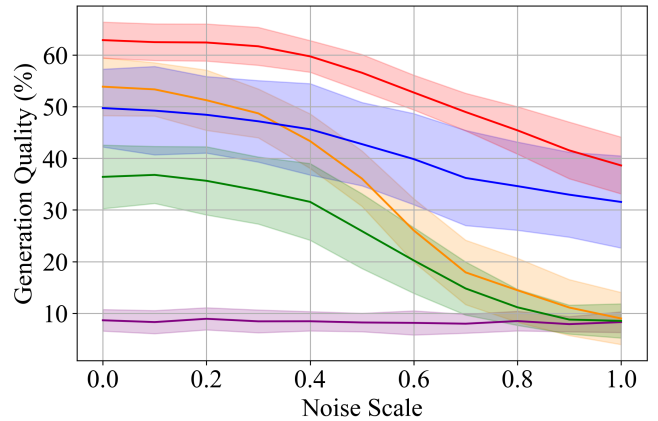
ited in data, we select top categories by frequency until cumulative probability exceeds 90%, encoding their multi-label data as multi-hot vectors for input.

Given varying real-world modality missing probabilities and imputation value (e.g., demographics are mostly complete, making imputation unmeaningful), we construct patchworks by randomly dropping one modality from diagnosis, treatment, or medication per client. This lets the model handle three key imputation tasks: disease diagnosis, drug recommendation, and treatment recommendation. Mortality prediction serves as the downstream task to evaluate representation quality. Both tasks use AUPRC due to class imbalance for assessment.

Table 2 presents the experimental results. GraphPL outperforms other baselines in both generation quality and representation quality across different numbers of participating hospitals, further confirming its advantages in real-world distributed scenarios.

### 3.4. Addressing the Modality Collapse Issue

To verify that GraphPL alleviates the modality collapse issue, we conduct experiments on PolyMNIST by adding various noise (via interpolation between real and noisy images) to each observed modality during inference. Fig. 4 shows generation quality changes with increasing noise scale when imputing a specific missing modality across methods. Metrics reflect the lowest generation quality under noise in different observed modalities. Baselines using POE for multi-modal feature fusion tend to rely on partial modalities, thus lacking robustness. In contrast, GraphPL uses GNNs to dynamically



**Fig. 4:** Generation quality of different methods (MVAE, MMVAE, MoPoE, CLAP, GraphPL) under varying noise scales.

fuse all observed modalities, maintaining better generation quality at high noise and thus mitigating modality collapse.

## 4. CONCLUSIONS

In this paper, we explore the use of GNNs for multi-modal feature fusion in patchwork learning, achieving an effective and efficient patchwork learning method across various benchmark datasets. Furthermore, compared to existing methods, GraphPL is better at balancing the use of information from different modalities and alleviates the modality collapse issue. We hope this work will contribute to the development of the emerging field of patchwork learning.

## 5. ACKNOWLEDGEMENT

The work of Xingjian Hu, Jianhua Zhu and Liangcai Gao is supported by the projects of Beijing Nova Interdisciplinary Program (20240484647) and National Natural Science Foundation of China (No. 62376012), which is also a research achievement of State Key Laboratory of Multimedia Information Processing, National Engineering Research Center of New Electronic Publishing Technologies and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

## 6. REFERENCES

- [1] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi, “The eicu collaborative research database, a freely available multi-center database for critical care research,” *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.
- [2] Sen Cui, Abudukelimu Wuerkaixi, Weishen Pan, Jian Liang, Lei Fang, Changshui Zhang, and Fei Wang, “CLAP: Collaborative adaptation for patchwork learning,” in *ICLR*, 2024.
- [3] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al., “Multimodal deep learning,” in *ICML*, 2011, vol. 11, pp. 689–696.
- [4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, “Multimodal machine learning: A survey and taxonomy,” *TPAMI*, vol. 41, no. 2, pp. 423–443, 2018.
- [5] Jiayi Chen and Aidong Zhang, “Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks,” in *KDD*, 2022, pp. 87–96.
- [6] Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu, “Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning,” in *AAAI*, 2022, vol. 36, pp. 3063–3071.
- [7] Suraj Rajendran, Weishen Pan, Mert R Sabuncu, Yong Chen, Jiayu Zhou, and Fei Wang, “Patchwork learning: A paradigm towards integrative analysis across diverse biomedical data sources,” *arXiv preprint arXiv:2305.06217*, 2023.
- [8] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao, “Kame: Knowledge-based attention model for diagnosis prediction in healthcare,” in *CIKM*, 2018, pp. 743–752.
- [9] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao, “Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks,” in *KDD*, 2017, pp. 1903–1911.
- [10] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera, “Mitigating modality collapse in multimodal vaes via impartial optimization,” in *ICML*. PMLR, 2022, pp. 9938–9964.
- [11] Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri, “Multimodal patient representation learning with missing modalities and labels,” in *ICLR*, 2024.
- [12] Geoffrey E Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [13] Yuge Shi, Brooks Paige, Philip Torr, et al., “Variational mixture-of-experts autoencoders for multi-modal deep generative models,” *NeurIPS*, vol. 32, 2019.
- [14] Diederik P Kingma, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [15] Mike Wu and Noah Goodman, “Multimodal generative models for scalable weakly-supervised learning,” *NeurIPS*, vol. 31, 2018.
- [16] Thomas M Sutter, Imant Daunhawer, and Julia E Vogt, “Generalized multimodal elbo,” *arXiv preprint arXiv:2105.02470*, 2021.
- [17] Thomas N Kipf and Max Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [18] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in *CVPR*, 2018, pp. 6848–6856.
- [19] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” *ICLR*, vol. 3, 2017.
- [20] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [21] Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel, “Unite and conquer: Plug & play multi-modal synthesis using diffusion models,” in *CVPR*, 2023, pp. 6070–6079.
- [22] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen, “General facial representation learning in a visual-linguistic manner,” *arXiv preprint arXiv:2112.03109*, 2021.