Table 1: Five recent research papers related to the area of variational autoencoders (VAEs)

Paper title, link	**Description**	**Relevance to course**
Learning Latent Subspaces in Variational Autoencoders, `https://arxiv.org/pdf/1812.06190`	This proposes a model that can learn low-dimensional latent representations which are directly correlated with binary labels of the data and recoverable after data modification. Thus, the model allows for more accurate attribute manipulation and identification of intra-class variation in images.	This connects the topics of VAEs, unsupervised learning, and information theory discussed in this course.
Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models, `https://arxiv.org/pdf/1911.03393`	This proposes four criteria that ideal multimodal generative models should follow and introduces a VAE that can perform transformations between images and languages on complex datasets. By combining VAEs with mixtures of experts, the model leads to new possible frameworks for multimodal learning.	This connects the topic of VAEs with the concepts of image latent variables and text embeddings discussed in the course. In addition to the main VAE model, the paper uses separate encoder and decoder neural networks for the experiments.
D-VAE: A Variational Autoencoder for Directed Acyclic Graphs, `https://arxiv.org/pdf/1904.11088`	This introduces a generative model that can optimize directed acyclic graphs by learning a latent space to encode graph performance rather than structure. The model paves the way for future work that can efficiently find optimal neural network architectures and Bayesian network representations of data.	This connects the concepts of VAEs and deep learning with directed acyclic graphs and Bayesian networks as discussed in the course. It also refers to recurrent neural networks in its methodology.
A Contrastive Learning Approach for Training Variational Autoencoder Priors, `https://arxiv.org/pdf/2010.02917`	This proposes a new prior for VAEs that can be trained using contrastive learning to overcome issues with posterior approximation. The training method is scalable and widely applicable, and it improves the generative quality of VAEs by allowing for sharper and more diverse images to be created.	This connects VAEs with the contrastive learning paradigm discussed in the course. It uses importance sampling and Langevin dynamics during testing and compares against MCMC sampling.
Alleviating Adversarial Attacks on Variational Autoencoders with MCMC, `https://arxiv.org/pdf/2203.09940`	This introduces a method that improves the robustness of latent representations against modified input data using MCMC sampling during inference time. This method maintains the quality of the outputs and allows for future VAEs to be more resistant against adversarial attacks.	This connects the topic of VAEs with MCMC sampling, specifically the Hamiltonian Monte Carlo algorithm, from the course.

Table 2: Five hypotheses on potential improvements to the results of the papers

Paper	**Hypothesis**	**Justification**
Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models	Introducing products of posteriors into the joint posterior approximation would allow the model to transform between multiple modalities at once and not just between pairs of modalities.	The current joint posterior approximation is a sum of the posteriors of each modality, but it does not have any term modeling the joint relationships across modalities. This means that the model can only translate from one modality to another modality at a given time, which may be inefficient for certain tasks.
Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models	The model would be more accurate if, instead of relying on an explicitly defined joint posterior, it can implicitly learn how to combine information from different modalities.	The existing regime of training the model using an explicit posterior can be difficult to apply to settings where inputs may have missing modalities. Furthermore, the posterior weights every modality equally, which could be an overly simplistic assumption at times.
D-VAE: A Variational Autoencoder for Directed Acyclic Graphs	Instead of using one-hot encoding, learning the optimal representations of vertices with a separate neural network before using the encoder network would improve the model's accuracy.	Vertex type is currently represented by a one-hot vector, which is inflexible since it assumes all vertex types are equally different. This can be false for some graphs (e.g., a linear layer is more similar to a softmax layer than a self attention layer in the Transformer architecture as a graph).
A Contrastive Learning Approach for Training Variational Autoencoder Priors	The model would improve the stability of the outputs by ensuring that the cross-entropy loss and the reweighting factor do not become extremely large.	The current loss and reweighting factor approach infinity for certain ranges of input values, which could cause the computations to be unstable during sampling. Defining an upper bound or adding a regularization term to the proposed equations could help mitigate this.
Alleviating Adversarial Attacks on Variational Autoencoders with MCMC	The method, which is applied during inference time, would be more efficient and also retain a similar level of accuracy if it used Langevin Monte Carlo (LMC) instead of Hamiltonian Monte Carlo (HMC).	Due to requiring many leapfrog steps for numerically integrating systems of differential equations, each iteration of the HMC algorithm can be computationally expensive. LMC is a similar sampling method which has a simpler algorithm and thus potentially quicker for use.

Introduction:

The paper “Learning Multimodal VAEs through Mutual Supervision”, communicated by Joy et al. at ICLR 2022, introduces the Mutually supErvised Multimodal VAE (MEME), a novel variational autoencoder model for multimodal data that can generate samples across disparate modalities. The paper presents a new way of formulating the problem of learning information from multiple modalities and demonstrates that the model outperforms prior approaches on various benchmarks.

Hypothesis:

To combine and learn information from multiple modalities, previous multimodal VAE models typically rely on custom posteriors which are explicitly defined in terms of the distributions of modalities. For instance, the paper “Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models” (Shi et al. 2019) discusses two approaches to do this: defining the posterior as a product (product-of-experts) or as a sum (mixture-of-experts) of the distributions of each modality. However, the method of using an explicit posterior assumes that all modalities are present in every observation (fully observed), and it does not generalize well to instances where some observations might have one or more modalities missing (partially observed). To resolve this issue, I hypothesize that a viable alternative method is to somehow make the model *implicitly* learn how to combine information from different modalities. This might work because the model would now be able to capture more complex latent relationships between modalities which would be difficult to identify manually.

Results:

The authors of the MEME paper employ this hypothesis by incorporating the self-supervision paradigm into the model's training procedure and avoiding the use of an explicit posterior. For the case with two modalities, they derive a self-supervised evidence lower bound for a general VAE with partially observed labels. Then, they construct the overall objective function as a weighted sum of the self-supervised ELBOs for each unique ordering of the modalities. This objective ensures that the modalities are mutually supervised - that is, each modality's encoder can be used as a conditional prior for the other modality, which allows information from both modalities to flow in both directions. The authors then modify this objective to the case with one missing modality using so-called pseudo-samples motivated by prior literature. Furthermore, they extend this objective to the case with more than two modalities, and they also provide suggestions for the code implementation, such as for sampling numerically stable gradient estimates.

After describing these new mathematical formulations, the authors conduct several experiments and analyses to ascertain the performance of MEME. Using full and partial observations, they evaluate the model on the MNIST-SVHN dataset to perform transformations between two types of images and on the CUB dataset to perform transformations between images and text. They find that MEME outperforms existing models for both full and partial observations on the MNIST-SVHN dataset as indicated by the cross coherence score, which measures the model's ability to reconstruct one modality given another modality as input. Similarly, they find that MEME maintains competitive accuracy on the CUB dataset as indicated by canonical correlation analysis, a statistical technique to measure coherence between the learned information from separate modalities. Moreover, the authors measure the accuracy of the latent samples by fitting linear classifiers to predict the input, and they observe that MEME is also more accurate in this regard compared to existing models, with this accuracy being rather uniform across different types of transformations. Finally, the authors use the Wasserstein probability metric to evaluate the dissimilarity between paired and unpaired data and the semantic similarity between the encodings within each image class, and they unsurprisingly again find superior performance from MEME.

Impact and next steps:

Overall, MEME is a state-of-the-art VAE model for multimodal data which outperforms other models according to a variety of metrics. Importantly, the model is able to generalize to observations with missing modalities, which solves an open problem in the application of VAEs to multimodal settings. Consequently, MEME creates the foundation for future domain-specific VAEs that can perform multimodal data generation with greater accuracy. Moving forward, I would go further into the paper by directly experimenting with the provided code. Since the authors have only used a few datasets to evaluate MEME so far, I would test the model on a wider range of datasets, tasks, and modalities to fully understand any improvement in its performance. Furthermore, I would systematically modify or remove parts of the model before testing to understand their contributions to the model performance.