Final Report: Language-conditioned Masked VAE

Dongze Ye* dongzeye@usc.edu
Jullian Yapeter* yapeter@usc.edu
Yuncheng Qiu* qiuyunch@usc.edu
Lei Cao* caolei@usc.edu

Abstract

Humans are naturally multimodal learners by virtue of our five senses. Multimodal data undoubtedly provides rich information at varying levels of abstraction and, as such, should help models gain a better understanding of its domain. However, the direction of deep learning research has been heavily focused on modality-specific models. While learning from multimodal data should be beneficial for many applications, multimodal deep learning is difficult; it often requires complex rigid architectures which cannot be efficiently fine-tuned for downstream tasks, and paired data from multiple modalities is much less accessible than single-modality data. Our Language-conditioned Masked Variational Autoencoder (LCMVAE) is a simple solution for multimodal representation learning, which takes advantage of modality-specific pre-trained encoders and a flexible, lightweight, latent-mixing network to effectively generate meaningful multimodal representations. While this paper focuses on bimodal image-caption pairs, LCMVAE is extensible to any number of modalities.

1 Introduction

Multimodal learning is an active, but far from mature, field of deep learning. In each of the popular modalities of deep learning research (e.g. computer vision, natural language, and audio), nuanced network architectures, loss functions, and auxiliary learning objectives are designed with certain inductive biases in mind (e.g. locality of pixels in images, time-series aspect of natural language and audio). For this reason, designing a singular architecture to be compatible with multimodal input is difficult, as the flexibility and generality of the network has to be traded off with the rigid but beneficial modality-specific inductive biases. Can we design a deep learning network to both be flexible and take advantage of the state-of-the-art in modality-specific processing? That is the question we aim to answer in this paper.

Early methods of multimodal representation learning utilized adaptations of Variational Autoencoders (VAE) to learn joint image-text representations (Sohn et al., 2015; Pandey and Dukkipati, 2016; Pu et al., 2016; Suzuki et al., 2016; Wu and Goodman, 2018). More recently, various models have been developed to solve multimodal tasks such as visual question answering (Agrawal et al., 2015), visual reasoning (Suhr et al., 2019), and text grounding (Yeh et al., 2018). These solutions focus on either modality-agnostic architectures (Jaegle et al., 2021; Baevski et al., 2022), complex, and often hand-engineered, multimodal fusion modules (Wu et al., 2019; Pramanik et al., 2020), or the conversion of one input modality into another (Radford et al., 2021; Li et al., 2021). These models also often rely on auxiliary labels such as object tags and object regions (Li et al., 2020, 2019), and are designed for specific tasks, making them ill-suited to be adapted for downstream tasks.

Our method, LCMVAE, aims to provide a flexible solution to harness multimodal data for use in any downstream setting. The key insight of LCMVAE is that we are aiming not to directly learn a multimodal representation, but rather to meaningfully mix pre-trained modality-specific latents. LCMVAE has four useful properties: (1) the ability to leverage modality-specific SOTA models that have been pre-trained on large datasets, (2) capable of fusing multimodal information in a data-driven, self-supervised, manner without architectural hand-engineering nor auxiliary labels, (3) follows a framework that readily scales to an arbitrary number of data modalities, and (4) can be easily fine-tuned for downstream tasks using self-supervised objectives.

*Please refer to our final presentation slides for distribution of work.

2 Related Work

Masked image encoding methods learn representations from images corrupted by masking. Motivated by the success of BERT (Devlin et al., 2018a) in NLP, whose training involves predicting masked text tokens, researchers have proposed similar vision models such as iGPT (Chen et al., 2020) and ViT (Dosovitskiy et al., 2020) that predict masked pixels or patches in an image. More recently, (He et al., 2021) proposed MAE (Masked Autoencoders), scalable masked image autoencoders with notable performance in self-supervised learning.

However, the benefits of masked encoding in multi-modal settings is not a well researched domain. Our project will be among the first to explore masked autoencoder's potential in multi-modal learning with a focus on language-image fusion.

Variational Autoencoders (VAEs) are a family of generative models that has shown promising performance on multi-modal learning in recent years (Khattar et al., 2019; Ivanovic et al., 2021; Huang et al., 2021). VAEs were first introduced by (Kingma and Welling, 2014) and (Rezende et al., 2014), which adapted variational inference methods and already foreshadowed VAE's ability on missing data imputation. Later multimodal extensions of VAEs, such as conditional VAEs (Sohn et al., 2015; Pandey and Dukkipati, 2016; Pu et al., 2016), JMVAE (Suzuki et al., 2016) and MVAE (Wu and Goodman, 2018), focused on learning unidirectional or *bi*-directional conditional probabilities. More recently, VAEs with further enhanced abilities to handle missing data in complex, multimodal datasets have been proposed (Nazábal et al., 2018; Ainsworth et al., 2018; Fortuin et al., 2020; Collier et al., 2021).

Our language-aided masked image reconstruction task is clearly analogous to multimodal missing data imputation. However, previous works treated VAEs as *standalone* models; whereas, we explored applying VAE as a modality fusion subnetwork for our bi-modal problem and our experiments produced insights on how to better utilize VAEs for multi-modal deep learning in general.

Multimodal learning approaches bridge different modalities; often either supplying missing modalities based on observed ones, or harnessing multiple observed modalities to perform complex tasks. Some previous works fused modalities with extra engineering. For example, UniVSE (Wu et al., 2019) heavily hand-engineers its visual-language alignment strategy, decomposing it into object, attribute, and relation-level alignments; OmniNet (Pramanik et al., 2020) fused the latents of different modalities with a spatio-temporal cache mechanism. However, LCMVAE does not require any extra engineering; our data-driven fusion module is a VAE that will hopefully result in a smoother, and more generalizable, latent space.

Most recently, (Zhang et al., 2021) outlined the common composition of vision-language models today and emphasized the importance of improving the image encoder, which inspired us to utilize a state of the art vision model for image feature extraction, such as MAE (He et al., 2021). Other recent works, like CLIP (Li et al., 2020) and GLIP (Li et al., 2021), leverage large, unfiltered, diversity of image caption pairs from the web, and trains transformer-based models to take as input a set of images and captions and predict the most likely pairings. Motivated by CLIP and GLIP, our model was trained on large datasets of image-caption pairs, e.g., COCO Captions (Chen et al., 2015). However, LCMVAE only requires one contextual prompt instead of one for each prediction class as in CLIP's case. This reduces the amount of prompt engineering required.

3 Problem Statement

Our goal is to investigate whether language supervision (i.e. image captions) and masked image encoding could be fused using a VAE in a meaningful way, thus generating a semantically rich embedding which can be used on downstream tasks, such as object detection or semantic segmentation.

While LCMVAE is ultimately meant to generate joint representations of paired data from different modalities of any kind, we are focusing on image-caption pairs as its first application. Our goal for this project is to find a set of design decisions for LCMVAE such that it can best extract complementary information from each of the data modalities. As a metric of LCMVAE's performance, we expect to outperform image-only computer vision models on vision tasks; thus proving the benefit of multimodal data and LCMVAE's capacity to exploit it. We also want to validate the benefit of masking images (which originated from the success of BERT (Devlin et al., 2018b), whose training involves predicting masked text tokens). Recently, the idea of masked encodings is gaining popularity in com-

puter vision; with researchers proposing models such as MAE (He et al., 2021). However, the benefits of masked encoding in multimodal settings is not a well-researched domain. Our project will be among the first to explore masked autoencoder's potential in multimodal learning with a focus on image-language fusion.

4 Description of Solution

We kept the design of LCMVAE simple. The method can be described in two parts: **phase 1)** modality-specific encoding, followed by, **phase 2)** multimodal latent-mixing. In the first phase, each modality of the paired multimodal dataset gets fed into their respective pre-trained, frozen, modality-specific encoder. In the second phase, we concatenate the resulting latents from phase 1 and feed it through a Variational Autoencoder (VAE), pre-training it with a self-supervised reconstruction objective. The output of the VAE's encoder is hence a single latent representation of the given multimodal data. In the downstream phase, with the modality-specific encoders still kept frozen, the encoder of the latent-mixing VAE is attached to a task-specific head and further fine-tuned to optimize the latent-mixing process for the given downstream task. While the described two phase framework can be applied to an arbitrary number of modalities, in this project, we focus on a bimodal image-caption application. The proposed architecture diagram is illustrated in Figure 1.

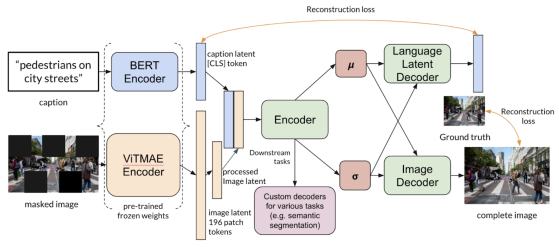


Figure 1: LCMVAE architecture diagram: our latent-mixing VAE generates smooth and continuous multimodal latent space. Input images are masked to produces semantically rich image embeddings and encourages the VAE to rely on useful language information. A pre-concatenation convolution layer learns to best utilize 196 patch-wise token embeddings. We employ an auxiliary loss to encourage high mutual information between the language input and the output multimodal embedding.

Modality-specific Transformers encoders. Our method proposes to utilize Transformers pre-trained on large datasets to encode each component of paired image-caption. Transformers have been shown to outperform traditional methods in terms of encoding data of different modalities; such as using LSTMs for natural language and Convolutional Neural Networks for images. While originally proposed for natural language processing, Transformers have been shown to be effective image encoders as well (Dosovitskiy et al., 2020). In this project, BERT and ViTMAE are used to encode captions and images respectively. The weights of these Transformers are kept frozen in order to improve training efficiency, relying on the multimodal latent-mixing VAE to learn how to best combine the pre-trained latents generated by the Transformers.

Multimodal Latent-mixing VAE. The embeddings obtained from the modality-specific encoders are then mixed using a VAE. First, the embeddings are concatenated into a single latent vector, a multilayer perceptron (MLP) is then used to compress the latent, outputting two equal-sized vectors representing a posterior distribution's mean and variance. This posterior represents a distribution of multimodal mixed latents. During the pre-training phase, we sample a multimodal latent from the posterior distribution and employ a convolutional decoder to generate image reconstructions. The VAE is pre-trained to minimize a weighted average of mean squared error loss between the ground truth images and the predicted image reconstructions, as well as the Kullback-Leibler (KL) divergence between the multimodal latent posterior distribution and a normal Gaussian prior. The balance between these two objectives is controlled using a hyperparameter, β. We chose a VAE over a vanilla Autoencoder (AE) as it creates a regularized, more continuous, latent space. As a result, VAE-generated representations often generalize better than that of AEs, thus producing improved results on downstream tasks. This VAE is the key to LCMVAE's efficiency. Other SOTA multimodal architectures require large amounts of compute to train. For example, Perceiver (Radford et al., 2021) was trained for 5 million steps on 32 TPUs. Since we are freezing the weights of the pre-trained transformers and only training our lightweight VAE, we were able complete this project, both the pre-training and downstream training phases, by training for a grand total of 16000 steps (40 epochs of COCO) on a single V100/A40 GPU. We believe with increased

compute, LCMVAE would achieve better results.

In addition to the modality-specific encoders and latent-mixing VAE, we propose a set of processing techniques as well as an auxiliary objective to promote more effective latent mixing, and produce richer multimodal latents. An architecture of our framework without any of these add-ons can be seen in Figure 5 in Appendix A.

Masking. LCMVAE employs a technique introduced by (He et al., 2021), which is to randomly mask out a high portion of image patches (75% to 95% of the image) and perform encoding only on unmasked patches. This augmentation technique forces the encoder model to produce semantically rich embeddings with which transfer learning yielded superior performance on downstream tasks. Moreover, LCMVAE utilizes image masking to encourage the model to rely on useful natural language information to compensate for incomplete visual signals.

Pre-concatentation convolutional token processing. While using the class [CLS] token of BERT's encoding carries sufficient semantic information of the image caption, our experiments show that the [CLS] token of the ViTMAE is insufficient to perform image reconstruction well. As such, we include a convolutional processing layer to pool the 196 image patch sequence tokens (derived from converting 224×224 images into a sequence of 16×16 patches) that are outputted by the ViTMAE to produce a single latent that respects the original positioning of the individual patches. This is especially important when masking is performed. Pre-processing is performed such that encodings of unmasked tokens are first rearranged and weaved together with zero vectors to achieve the correct relative positioning of unmasked patches. The pre-processed latent tensor then gets fed through a convolutional neural network to obtain a a rich image latent representation that is of the same dimension as the caption's latent representation (768-vector).

Latent reconstruction auxiliary objective. In this project, we are also investigating the effects of introducing a language latent reconstruction objective, which is optimized jointly with the image reconstruction objective. This auxiliary objective encourages the mixed multimodal latent to retain high mutual information with the caption latent. With image reconstruction as LCMVAE's principal self-supervised objective, the network might be biased towards relying solely on visual signals. Co-optimizing image reconstruction and language latent reconstruction would in theory force the network to pick out language information that is advantageous for understanding (and predicting) image pixels. Bimodal image-caption representations generated by LCMVAE should thus contain both high-level semantics and low-level pixel-wise image understanding, which would be beneficial for downstream visual semantic understanding tasks, such as our downstream task of choice for this project: semantic segmentation. The latent reconstruction objective is weighed relative to the image reconstruction using a hyperparameter, δ.

Downstream training. After pre-training on the joint image-caption reconstruction objective, we perform transfer learning on downstream tasks by swapping out the VAE decoder for a task-specific head. For our semantic segmentation downstream task, we attach a decoder, that outputs pixel-wise probabilities of each object class, to the end of our pre-trained VAE encoder. During this downstream training phase, our BERT and ViTMAE encoders are still kept frozen, but our latent-mixing VAE encoder is un-frozen and allowed to fine-tune to the new task. Therefore, the newly attached semantic segmentation head is co-optimized with the multimodal latent-mixer. This is necessary as the optimal latent mixture for each downstream task would differ from that of the pre-training image reconstruction task.

Our technical contribution is four-fold: (1) LCMVAE is modularly designed and readily able to utilize the state-of-the-art representation learning method for each data modality. (2) Our method is able to make use of pre-training on large single-modality datasets (e.g. able to leverage publicly available pre-trained weights). (3) Instead of hand-engineering the fusion process of the modality-specific latents, we employ a VAE to effectively fuse latents by using a self-supervised image reconstruction objective, regularized on auxiliary language latent reconstruction loss (i.e. LCMVAE is easily trainable end-to-end). (4) LCMVAE is capable of scaling up to take as input tuples of any number of modalities by simply adding more modality-specific encoders and auxiliary reconstruction objectives; all the precedence of which is laid out by our bi-modal framework. Our code is released at https://github.com/jullian-yapeter/lcmvae.

5 Experiment Setting

5.1 Dataset

BERT and ViTMAE are pretrained on their respective modality-specific datasets: our BERT encoder is pre-trained on the datasets BookCorpus (Zhu et al., 2015) and English Wikipedia (Wikimedia Foundation), and our ViTMAE were pre-trained using the ImageNet-1K dataset (Ridnik et al., 2021). Note that this expensive pre-training step only has to be done once, and that LCMVAE can be fine-tuned on a smaller, task-specific, dataset.

MS COCO (Lin et al., 2015) is a high quality crowd-labeled dataset. It covers a wide range of classes but is small by modern standards. Additionally, COCO is a large-scale object detection, segmentation, and captioning dataset. Due to these factors, we use COCO for our multi-modal model both for pre-training (reconstruction) and downstream tasks. Specifically, we are using COCO Caption 2017 (Chen et al., 2015), which contains 118, 000 training images, 5, 000 validation images, and 6 annotation JSON files containing captions, segmentation labels, and etc. For simplicity, we also resized the images to a fixed size: $(3 \times 224 \times 224)$. Due to resource limitation and to obtain an easier semantic segmentation task, we only use images from COCO (for both training and validation) with the following nine categories: person, bicycle, car, motorcycle, airplane, bus, train, truck and boat. As a result, we have 102k training images and 4,350 validation images, along with the corresponding captions.

5.2 Evaluation Procedure

We evaluate our models on their performance in image reconstruction and semantic segmentation (as a downstream task) using COCO's **validation** set. Given the nature of the chosen tasks, we measure our models' performance both quantitatively (using the loss functions and validation metrics) and qualitatively (through visualization).

Image reconstruction. LCMVAE and its variants receive validation images with different mask ratios (0%, 25%, 50%, and 75%) as inputs, along with full or empty captions. However, VAE and AE baselines always ignore language inputs.

Semantic segmentation. Our models are only evaluated on *un-masked* images along with full or empty captions.

The **reconstruction and segmentation targets** are always based on original, un-masked images.

We use simple, **image-wise**[1] loss functions – **MSE** for reconstruction and **cross-entropy loss** for segmentation– as performance metrics. As is standard in VAEs, we also use a KL divergence term (scaled by β) as a regularizer for the posterior distribution. Additionally, we examine reconstruction and segmentation qualities by visualizing model outputs along with the targets.

In addition, we use the **mIoU** (mean Intersection-over-Union) as a performance metric for semantic segmentation on the COCO dataset. However, to the best of our knowledge, semantic segmentation (and thus the mIoU metric) is rarely performed on COCO, and thus we only compare the semantic segmentation results for LCMVAEs and baselines within our project. Due to limited resource, we do not test the transfer learning performance of our models on different datasets.

5.3 Baselines

The vanilla **VAE** and the deterministic **AE** serve as simple baselines for masked image reconstruction. Our implementation of the baseline VAE and AE has a symmetric encoder-decoder design with convolutional and deconvolutional layers and uses batch-normalization and LeakyReLU activation in between layers. However, these baselines only operate on masked images *without* using caption information.

LCMVAE-Baseline is a pilot version of our proposed architecture and serves as a baseline for both image reconstruction and semantic segmentation. Its encoder consists of pre-trained ViT-MAE and BERT (with frozen weights) followed by an MLP. Consistent without our design philosophy, LCMVAE-baseline also uses light-weight task-specific decoders such as an MLP for image reconstruction and a deconvolutional network for semantic segmentation.

As compared to the LCMVAE-baseline, our full model is lighter weight (15M trainable parameters, almost $6\times$ lighter than the baseline model) as we swapped out the baseline's reconstruction MLP decoder for a new deep convolutional design (that leverages added skip connections for improved gradient flow across layers and the advantage of shared kernel weights). Our full LCMVAE model also includes previously described features such as input

[1] We take the sum of losses over all pixels for a (fixed-size) image, since using the average loss per pixel led to worse results during initial trials.

image masking, pre-concatenation convolution and language-latent reconstruction, which are not included in our baseline model.

5.4 Implementation Details

Masking. For VAE and AE baselines, we create masks by modifying input images directly turning random patches (size 16×16) or individual pixels of an image black (pixel value of 0). For LCMVAE, images are automatically patchified, tokenized, and masked by MAE's encoder (He et al., 2021).

As suggested by (He et al., 2021), we use 75% **mask ratio** for LCMVAEs. For a fair comparison, we use the *same* mask ratio for VAE-baseline. We acknowledge that the naive masking scheme for VAE-baseline would impose unwanted artifacts (black pixels) to input images. Nevertheless, this issue highlights the advantage of MAE's (and thus LCMVAE's) masking strategy, which involves a creative use of the Transformer architecture. We compare results of training using 75% mask ratio versus using no mask at all, with different mask ratios at test time.

Ablation study. We create variants of LCMVAE by dropping individual modules (masking, caption inputs, and etc.) and all reasonable combinations of its modules. In total, we experiment on 24 variants of LCMVAE to thoroughly examine the importance of each module as well as potential interaction effects. Detailed model configurations are included in Table 4.

Pre-training (reconstruction). LCMVAEs are pre-trained with image-caption pairs, while VAE-baseline only uses images. Loss curves and training parameters are shown in Appendix B. Due to limited resource, we fix $\beta = 0$ for the baselines, in order to explore an upper-bound of reconstruction quality; for the final LCMVAE (and its ablated variants), we fix $\beta = 50$, which is a value that leads to reasonable image reconstruction results during sample runs. Additionally for our language latent-reconstruction objective, we compare the effects of setting δ to 0 and 5e4. While our image inputs are 224×224 images, we are outputting 56×56 reconstructions in order to reduce the number of trainable parameters. The predicted reconstruction is then up-sampled bilinearly to 224×224 in order to compute mean squared error loss.

Fine-tuning (segmentation). LCMVAEs are then fine-tuned for semantic segmentation with newly initialized deconvolutional heads. The deconvolution head is parameterized almost identically to the pre-training reconstruction deconvolution decoder, except that it outputs 10 channels (9 object classes and 1 background class) instead of 3 RGB channels. The network, including the semantic segmentation head, is trained using pixel-wise cross entropy loss (the reported losses are pixel-wise losses summed over whole images).

Training schedule. All models follow the same training schedule: 25 epochs for image reconstruction and then 15 epochs for semantic segmentation. More details can be found in Appendix B.

6 Results

6.1 Overfitting Experiment

To test our implementation, we first overfit LCMVAE on a tiny subset of COCO (10 images) without masking and visually examine its reconstructions. As shown in Figure 8, we observe that our model's reconstructions are always semantically similar to the targets. This behavior suggests that given sufficient training, LCMVAEs may make use of captions for better image reconstruction.

6.2 (Masked) Image Reconstruction

Overall reconstruction quality is, unfortunately, subpar for all models considered in our project. The validation image with the lowest reconstruction loss is shown in Figure 2. Although the reconstruction loss for this image is already 70–145$\times$ lower than the mean reconstruction losses recorded in Table 4, the reconstructed image is still far from recognizable. Such result is likely because masked image pre-training and image reconstruction for real life images are highly resource intensive. For reference, (He et al., 2021) pre-trained the original MAE for 1500 epochs on ImageNet-1K using 128

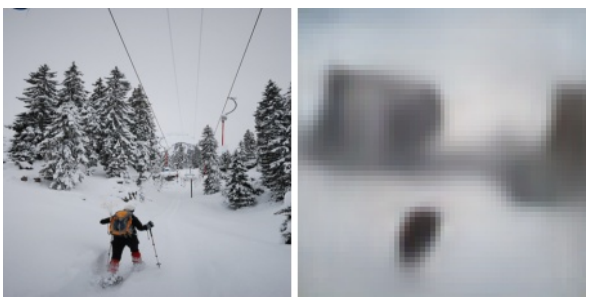


Figure 2: Visualization of LCMVAE's image reconstruction result. This sample has a reconstruction loss of 812.6 (lowest among all validation images) with no masking at test time.

TPUv3 cores. Whereas, we only pre-train each of our models for 25 epoches on a small subset[2] of the COCO dataset using a single V100/A40 GPU on CARC. We also cannot fine-tune modality-specific Transformer encoders due to resource limitations. Therefore, we believe our reconstruction results are reasonable under our current compute constraints. More visualization examples can be found in Appendix D.

Mask ratio during validation. Recall that although we have fixed masked ratios (0% or 75%) for pre-training our models, we vary mask ratios at test time in order to examine our model's performance in reconstructing images of different mask ratios.

Effect of masked image pre-training. As shown in Figure 3, the mean reconstruction loss for LCMVAE-Full (with masked pre-training) is slightly decreased when reconstructing masked images. However, the mean reconstruction loss for LCMVAE-noMask (without masked pre-training) increases rapidly with higher mask ratio. This pattern is also verified by visualizing particular examples (Figure 11 & 12). Moreover, looking at the expected losses of masked image reconstruction for all LCMVAE variants (Table 5), we find a strong evidence that masked pre-training can significantly improve LCMVAE's performance in reconstructing masked images.

However, LCMVAEs with masked pre-training tend to perform worse on reconstructing unmasked images (as shown in Table 4). This tendency is expected since we use a short, fixed training schedule (25 epochs) for all models. In fact, since LCMVAEs with masked pre-training only receive 25% of image signals, it highly likely requires more training than its unmasked counterparts.

[2]Our training set contains 100k images, which is about 137× smaller than ImageNet-1K.

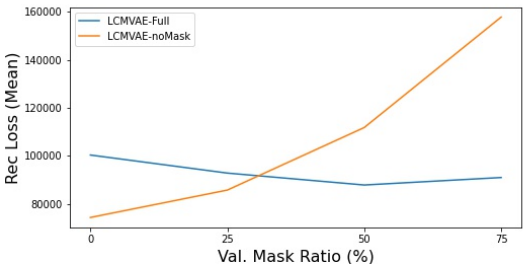


Figure 3: **Reconstruction loss for for LCMVAEs with and without masked pre-training.** LCMVAE-Full is trained with masked images while LCMVAE-noMask with *un-masked* images. Both models are then evaluated using validation images from COCO with different mask ratios.

Model	Rec Loss (mean)	Std Dev
AE	77890	34028
VAE	89749	39350
LCMVAE-Full	**90942**	**40202**
LCMVAE-Baseline	343100	-

Table 1: **Full model vs. Baselines.** We report the mean and standard deviation of reconstruction losses on COCO validation images. Mask ratio is 75% for both training and validation.

Effect of captions. We fail to observe any significant effect of captions with our experiments on image reconstruction.

Comparison with baselines. As shown in Table 1, LCMVAE has similar performance on masked image reconstruction when compared to VAE and AE baselines. However, the full LCMVAE is significantly better than the LCMVAE-baseline due to our new design after the midterm. Nevertheless, since LCMVAEs rely on the embeddings from pre-trained Transformer encoders with frozen weights, but the VAE and AE baselines are trained end-to-end using masked images, we believe LCMVAE with trainable Transformer weights and sufficient training may easily surpass the performance of the baseline models on masked image reconstruction.

6.3 Semantic Segmentation

LCMVAE's overall performance on semantic segmentation is significantly better than its perfor-

Figure 4: Comparison of semantic segmentation results for the full LCMVAE using caption vs. *not* using caption at test time.

mance on masked image reconstruction.

Effect of Captions. As shown in Table 6, we find strong evidence that LCMVAE is capable of leveraging captions during the downstream semantic segmentation task, leading to consitiently higher mIoU scores. This result is consistent with our hypothesis that captions may provide useful contextual information for downstream tasks.

Effect of Masked pre-training. We fail to observe any significant effect of masked pre-training on LCMVAE's semantic segmentation performance.

Effect of Other modules. Out of LCMVAE's features, the pre-concatenation convolutional layer seems to be the second most significant in improving performance (behind the use of captions). However, models trained without captions and no pre-concatenation convolution layer obtains the worst mIoU measures.

7 Discussion

Multimodal data provides rich information, but research has been heavily focused on single modality models and current architectures are still difficult to train and fine-tune for downstream tasks. In this study, we designed a simple data-driven multi-modal architecture, LCMVAE, and explored the use of self-supervised learning methods, like masked image encoding, for improving our model's efficiency in learning downstream tasks.

We thoroughly evaluate our theoretical claims with multiple experiments and empirically verify that LCMVAE is able to harness contextual information provided via captions to improve its performance in downstream tasks such as semantic segmentation, even with limited pre-training. Although our model exhibits no particular advantage in reconstructing masked image, LCMVAE still appears to improve proportionally to additional training data and increased training iterations.

Nevertheless, the comparison of semantic segmentation results is limited to LCMVAE and its ablated variants for reasons aforementioned. Therefore, additional experiments are needed to evaluate our segmentation results objectively. Moreover, we hypothesize that by using captions, LCMVAE may obtain some unfair advantage by knowing the target classes and the location of certain objects without even looking at the input images. Yet, it still seems impressive for LCMVAEs to "understand" and capitalize on the contextual information from captions to achieve seemingly better visual understanding.

Unfortunately, we do not have comparable prior works that consider both masked image pre-training, and vision-language fusion with VAEs against which to make further observations.

The full LCMVAE-baseline had better performance than all reduced LCMVAE-baseline during our initial ablation study, leading to the belief that all the features of LCMVAE, or at least a certain combination of its features, are advantageous. However, some of LCMVAE's features that initially showed promise did not exhibit strong evidence of improving the model in our final experiments. Due to the lack of training, we are unable to conclusively determine whether all components of LCMVAE are truly helpful.

Our experiments show that LCMVAE can interpret vision-language inputs and generate multimodal latent embeddings, though not as effectively as we had hoped. Although we iterated on LCMVAE's architecture design as informed by our intermediate results, the final performance is still below our expectation but better than the original naive design. Hence, we hypothesize that additional design modifications may lead to a better fusion of modalities. Unfortunately, we have not verified this thought yet due to limited computational resources, but we hope this prospect will inspire future work.

Lastly, we realize that transfer learning is a complicated art. LCMVAE relies on the hypothesis that the outputs of Transformers can be easily fused and transferred to different tasks, which underlines our decision for pre-training on masked image reconstruction and then fine-tuning on semantic segmentation. However, though VAEs are lightweight solutions for mixing multimodal latents, it may have a rather low ceiling for performance. Our current VAE architecture might be too naive to realize the full potential of mixing Transformer outputs. In future work, we would like to explore different designs for a multimodal latent mixer, such as recurrent networks like LSTMs and potentially a comparatively lightweight Transformer decoder specially designed for token-mixing. Ultimately, we believe that since our resource-constrained study shows that a lightweight network could meaningfully leverage pre-trained modality-specific encoders, it is worth pursuing further research to design better latent mixers to make multimodal learning more efficient and deployable for real life applications.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. Vqa: Visual question answering.

Samuel K. Ainsworth, Nicholas J. Foti, and Emily B. Fox. 2018. Disentangled VAE representations for multi-aspect and missing data. *CoRR*, abs/1806.09060.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Mark Collier, Alfredo Nazabal, and Christopher K. I. Williams. 2021. VAEs in the presence of missing data.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Raetsch, and Stephan Mandt. 2020. GP-VAE: Deep probabilistic time series imputation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1651–1661. PMLR.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2021. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377.

Feiran Huang, Xiaoming Zhang, Jie Xu, Zhonghua Zhao, and Zhoujun Li. 2021. Multimodal learning of social image representation by exploiting social relations. *IEEE Transactions on Cybernetics*, 51(3):1506–1518.

Boris Ivanovic, Karen Leung, Edward Schmerling, and Marco Pavone. 2021. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robotics and Automation Letters*, 6(2):295–302.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General perception with iterative attention.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA. Association for Computing Machinery.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2021. Grounded language-image pre-training. *CoRR*, abs/2112.03857.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common objects in context.

Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. 2018. Handling incomplete heterogeneous data using VAEs. *CoRR*, abs/1807.03653.

Gaurav Pandey and Ambedkar Dukkipati. 2016. Variational methods for conditional multimodal deep learning.

Subhojeet Pramanik, Priyanka Agrawal, and Aman Hussain. 2020. OmniNet: A unified architecture for multi-modal multi-task learning.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images,

labels and captions. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Joint multimodal learning with deep generative models.

Wikimedia Foundation. Wikimedia downloads.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. UniVSE: Robust visual semantic embeddings via structured semantic representations.

Mike Wu and Noah D. Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. *CoRR*, abs/1802.05335.

Raymond A. Yeh, Minh N. Do, and Alexander G. Schwing. 2018. Unsupervised textual grounding: Linking words to image concepts.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Appendix A Architecture Diagrams

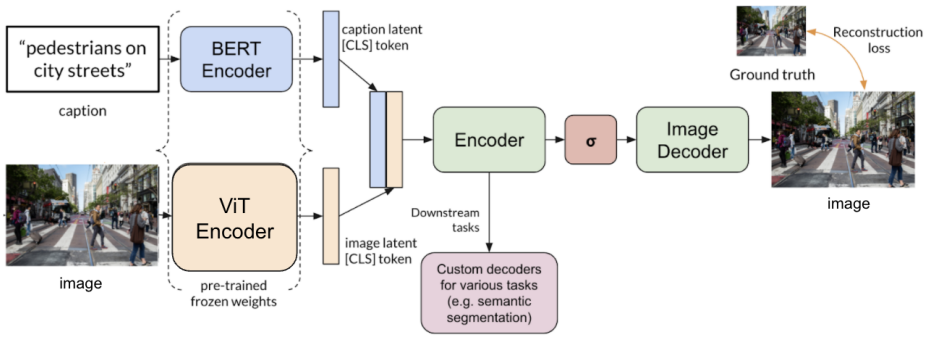


Figure 5: LCMVAE's initial architecture diagram: Our method encodes multimodal data with their respective Transformer encoders, and fuses latents in a fully data-driven manner by training an Autoencoder. LCMVAE is trained with self-supervised reconstruction objectives.

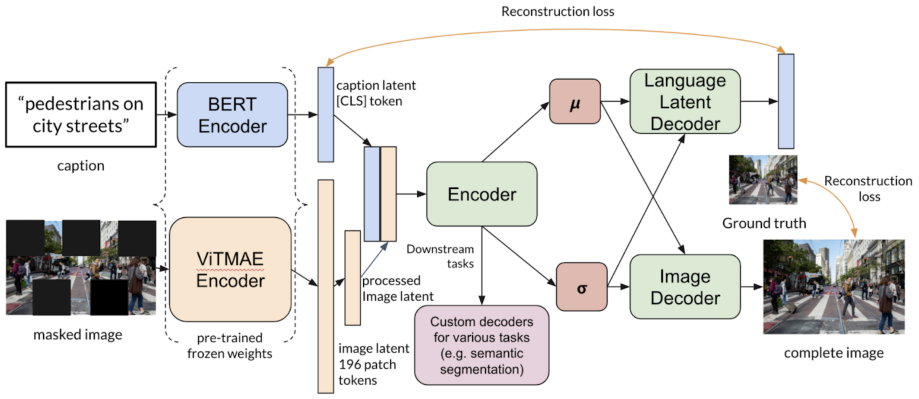


Figure 6: LCMVAE's new architecture diagram. We replaced the Autoencoder with a VAE to generate a smooth and continuous multimodal latent space. We mask the image input to produce semantically rich image embeddings. A pre-concatenation convolution layer is added in to learn how to best utilize the 196 patch-wise token embeddings outputted by ViTMAE.

Appendix B Training Details

Configuration	Value
Optimizer	Adam
Learning rate	3e-4
Learning rate scheduler	*None*
Batch size	256
Epochs	25
Initialization	`xavier_unif`
β	50
γ	5e4

Table 2: Hyper-parameters for pre-training (image reconstruction).

Configuration	Value
Optimizer	Adam
Learning rate	3e-4
Learning rate scheduler	*None*
Batch size	256
Epochs	15
Initialization	`xavier_unif`
β	50
γ	5e4

Table 3: Hyper-parameters for training (semantic segmentation).

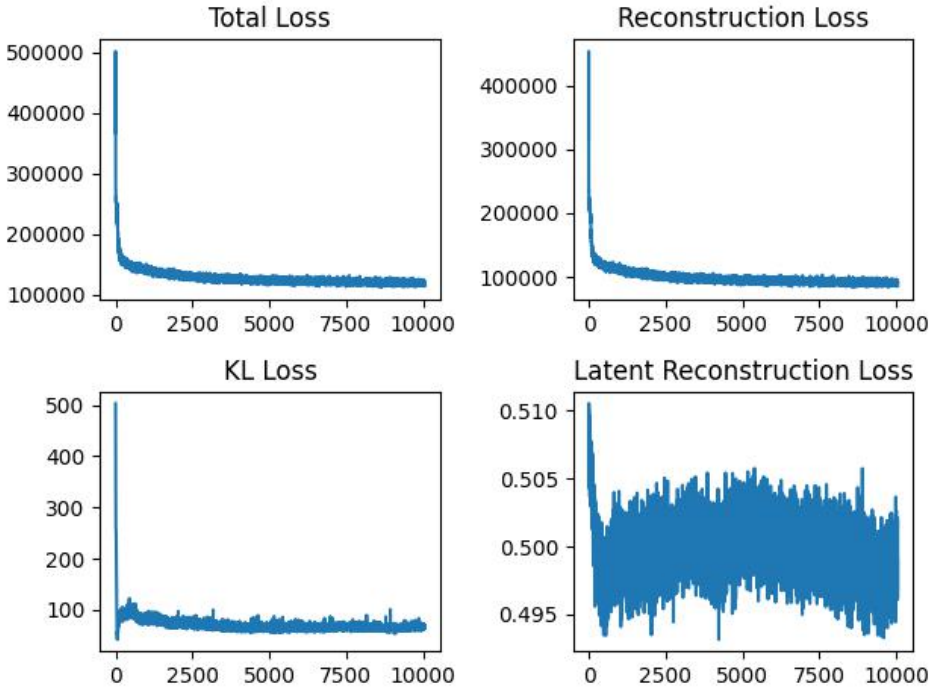


Figure 7: Training loss curves for LCMVAE-Full. The reconstruction loss is the sum of pixel-wise MSE for each image (with fixed-size). For simplicity, we set $\beta = 50$.

Appendix C Quantitative Results

Mask	Caption	Variational	Latent-Reg	Pre-Conv	Rec Loss (Mean)	Std Dev
No	No	No	No	Yes	56464	26796
No	Yes	No	No	Yes	57339	27046
No	Yes	No	Yes	Yes	59802	31303
No	Yes	Yes	Yes	Yes	74299	33735
No	Yes	Yes	No	Yes	74860	33980
No	No	Yes	No	Yes	74886	34002
Yes	Yes	No	No	Yes	92470	35553
No	No	No	No	No	95875	44583
Yes	No	No	No	Yes	96927	37538
No	Yes	No	Yes	No	100193	46946
Yes	Yes	Yes	Yes	Yes	100310	41122
No	Yes	No	No	No	100321	47790
No	No	Yes	No	No	100346	44079
Yes	No	No	No	No	101268	44742
No	Yes	Yes	Yes	No	101320	45567
No	Yes	Yes	No	No	102266	46815
Yes	Yes	No	Yes	No	105284	47167
Yes	Yes	No	No	No	105694	47391
Yes	No	Yes	No	No	106959	46696
Yes	Yes	Yes	Yes	No	107931	47722
Yes	Yes	Yes	No	No	109120	48406
Yes	Yes	No	Yes	Yes	112490	47263
Yes	No	Yes	No	Yes	114798	46577
Yes	Yes	Yes	No	Yes	116220	62717

Table 4: List of **LCMVAE configurations** for ablation study and **reconstruction losses** based on **unmasked validation images**. β is fixed to 50 in our training process. The list is sorted by mean of reconstruction loss in ascending order. The losses are computed by taking the sum of pixel-wise MSE or Cross-Entropy loss for each image (with fixed size).

Mask	Caption	Variational	Latent-Reg	Pre-Conv	Rec Loss (Mean)	Std Dev
Yes	Yes	No	No	Yes	79207	36102
Yes	No	No	No	Yes	81033	36692
Yes	Yes	No	Yes	Yes	82886	37607
Yes	Yes	Yes	No	Yes	87918	39403
Yes	Yes	Yes	Yes	Yes	90942	40202
Yes	No	Yes	No	Yes	91516	40996
Yes	No	No	No	No	109360	49054
Yes	Yes	No	Yes	No	111689	50675
Yes	Yes	No	No	No	112026	51239
Yes	No	Yes	No	No	113444	50311
Yes	Yes	Yes	No	No	114562	51105
Yes	Yes	Yes	Yes	No	114620	51003
No	Yes	Yes	Yes	No	138879	61586
No	Yes	Yes	No	No	140576	62680
No	No	Yes	No	No	144698	60718
No	Yes	No	Yes	No	151395	64729
No	Yes	No	No	No	153783	66807
No	No	No	No	No	154863	65218
No	No	Yes	No	Yes	156962	56411
No	No	No	No	Yes	157094	55389
No	Yes	Yes	Yes	Yes	157797	54841
No	Yes	Yes	No	Yes	159092	57019
No	Yes	No	No	Yes	165828	53700
No	Yes	No	Yes	Yes	675534	174330

Table 5: List of **model configurations** and **reconstruction losses** for preliminary experiments on **75% masked image**. β is fixed to 50 in our training process. The list is sorted by mean of reconstruction loss in ascending order. The losses are computed by taking the sum of pixel-wise MSE or Cross-Entropy loss for each image (with fixed size).

Mask	Caption	Variational	Latent-Reg	Pre-Conv	mIoU	Std Dev^{IoU}	Seg Loss	Std Dev^{loss}
No	Yes	No	Yes	Yes	0.370	0.202	25349	26962
No	Yes	Yes	No	Yes	0.362	0.205	26877	29276
Yes	Yes	No	No	Yes	0.360	0.204	26901	30286
Yes	Yes	Yes	No	Yes	0.358	0.202	26793	30359
No	Yes	Yes	Yes	Yes	0.354	0.208	27116	30064
Yes	Yes	Yes	Yes	Yes	0.352	0.209	26812	28659
Yes	Yes	No	Yes	Yes	0.350	0.206	27457	30978
No	Yes	No	No	Yes	0.348	0.209	25880	28000
Yes	Yes	Yes	No	No	0.330	0.203	26755	24138
No	Yes	Yes	No	No	0.316	0.205	27478	25574
Yes	Yes	No	Yes	No	0.313	0.207	27689	25473
No	Yes	No	No	No	0.312	0.206	29260	29938
Yes	Yes	No	No	No	0.310	0.203	27804	26642
No	Yes	Yes	Yes	No	0.304	0.203	26306	23760
Yes	Yes	Yes	Yes	No	0.297	0.208	27903	26081
No	Yes	No	Yes	No	0.291	0.217	26543	25011
No	No	Yes	Yes	Yes	0.189	0.229	40501	51805
Yes	No	No	Yes	Yes	0.186	0.233	40195	52710
No	No	No	Yes	Yes	0.186	0.227	41747	54623
Yes	No	Yes	Yes	Yes	0.183	0.232	40144	51589
Yes	No	No	Yes	No	0.134	0.228	36859	34787
No	No	No	Yes	No	0.129	0.230	38295	41441
No	No	Yes	Yes	No	0.128	0.230	35593	33428
Yes	No	Yes	Yes	No	0.122	0.232	36123	34834

Table 6: List of **model configurations**, **mIoU** and **segmentation losses** for experiments on semantic segmentation. The list is sorted by mIoU in ascending order. Loss and mIoUs are computed using the COCO validation set without masking during evaluation.

Model	Val. Mask Ratio	Rec Loss (mean)	Std Dev
VAE-Baseline	0%	89826	39666
AE-Baseline	0%	77979	34208
VAE-Baseline	75%	89749	39350
AE-Baseline	75%	77890	34028

Table 7: VAE-Baseline Vs. AE-Baseline

Model	Val. Mask Ratio	Rec Loss (mean)	Std Dev
Full	0%	100310	41122
No Mask	0%	74299	33735
Full	75%	90942	40202
No Mask	75%	157797	54841

Table 8: Full Vs. No Mask

Model	Val. Mask Ratio	Rec Loss (Mean)	Std Dev
Full	0%	100310	41122
No Caption	0%	114798	46577
Full	75%	90942	40202
No Caption	75%	91516	40996

Table 9: Full Vs. No Caption

Model	Val. Mask Ratio	Rec Loss (Mean)	Std Dev
Full	0%	100310	41122
No Variation	0%	112490	47263
Full	75%	90942	40202
No Variation	75%	82886	37607

Table 10: Full Vs. No Variation

Model	Val. Mask Ratio	Rec Loss (Mean)	Std Dev
Full	0%	100310	41122
No Latent-Reg	0%	116220	62717
Full	75%	90942	40202
No Latent-Reg	75%	87918	39403

Table 11: Full Vs. No Latent-Reg

Model	Val. Mask Ratio	Rec Loss (Mean)	Std Dev
Full	0%	100310	41122
No Prec-Conv	0%	107931	47722
Full	75%	90942	40202
No Prec-Conv	75%	114620	51003

Table 12: Full Vs. No Prec-Conv

Appendix D Visualizations for Image Reconstruction

Figure 8: Overfitted image reconstruction: LCMVAE trained to overfit a small amount of seed data (10 images) showed signs of capturing semantics; when it was prompted with a validation set image, it reconstructed an image from the seed dataset that was semantically similar.

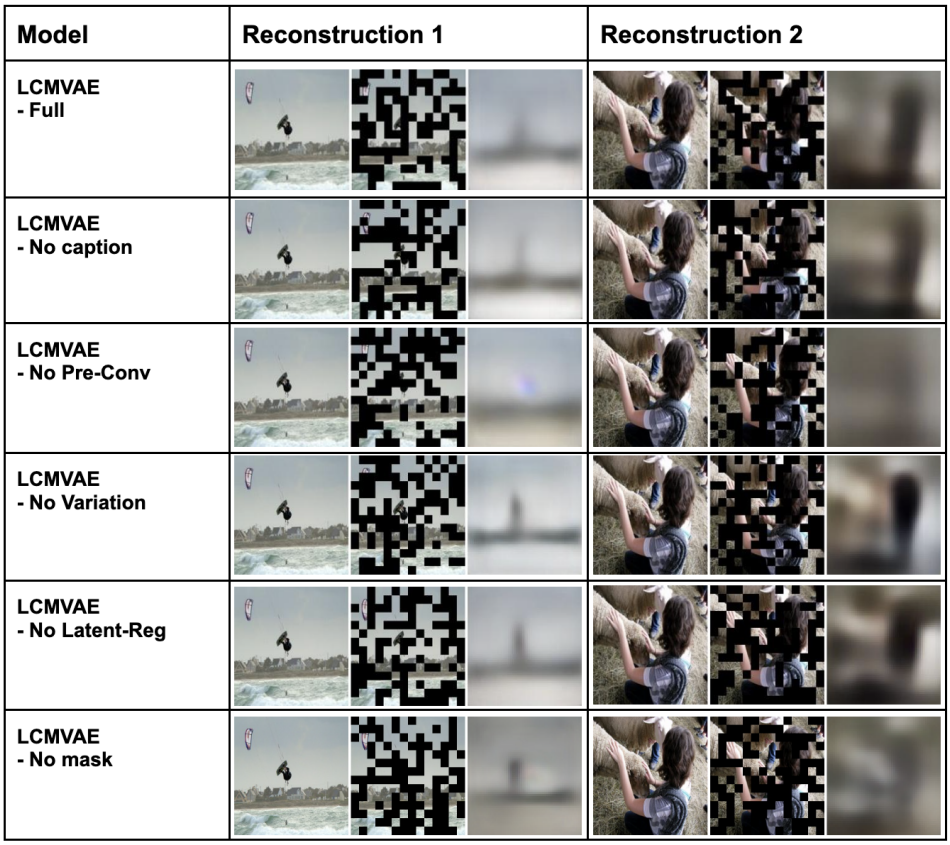


Figure 9: Image reconstruction: Image reconstruction performance of LCMVAE-Baseline and its variants. Changes in both training parameters and architectural adjustments improve the reconstruction results compared with Midterm report.

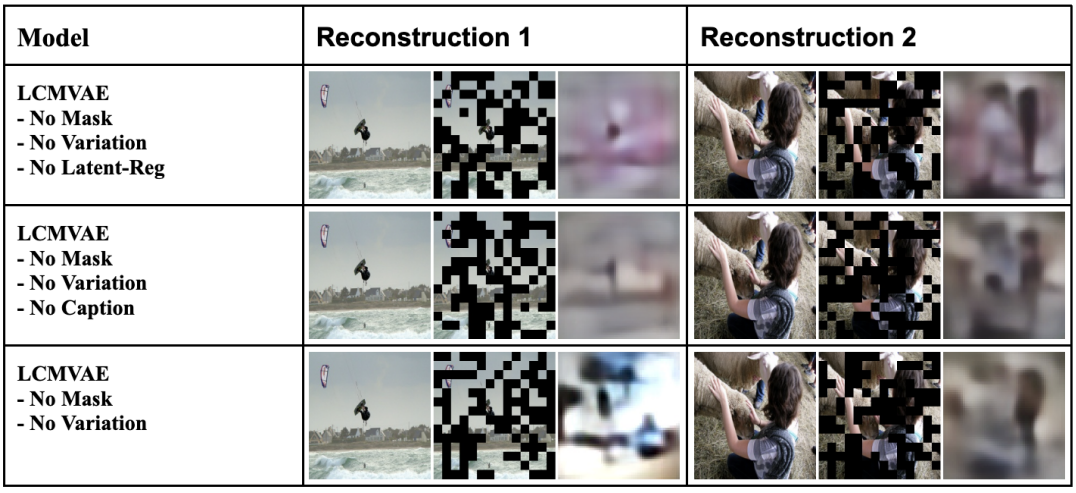


Figure 10: Worst reconstruction results: Most of the worst performance comes from models trained without variational module and masked pre-training.

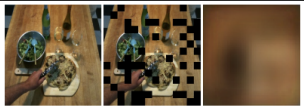
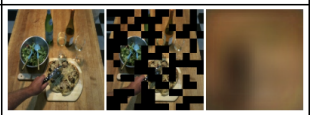
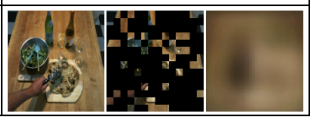
Model	Reconstrcution 1	Loss 1	Reconstruction 2	Loss 2
LCMVAE **- Full** **- Val. Mask Ratio = 0%**		75780		63756
LCMVAE **- Full** **- Val. Mask Ratio = 25%**		19687		110495
LCMVAE **- Full** **- Val. Mask Ratio = 50%**		121040		127155
LCMVAE **- Full** **- Val. Mask Ratio = 75%**		70063		55435

Figure 11: Reconstruction result: LCMVAE trained with 75% mask ratio on images with different mask ratio for validation.

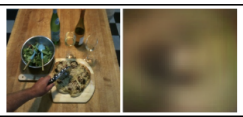
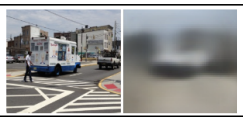
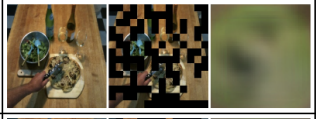
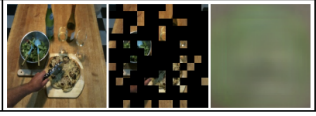
Model	Reconstruction 1	Loss 1	Reconstruction 2	Loss 2
LCMVAE **- Trained without mask** **- Val. Mask Ratio = 0%**		11899		121308
LCMVAE **- Trained without mask** **- Val. Mask Ratio = 25%**		74312		19133
LCMVAE **- Trained without mask** **- Val. Mask Ratio = 50%**		105982		95383
LCMVAE **- Trained without mask** **- Val. Mask ratio = 75%**		235720		172745

Figure 12: Reconstruction result: LCMVAE trained without masking on images with different mask ratio for validation.

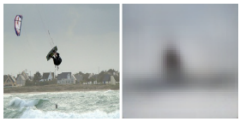
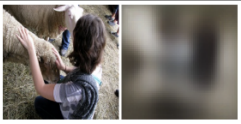
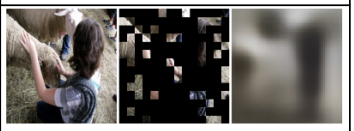
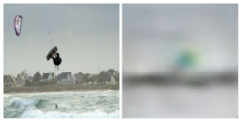
Model	Val. Mask Ratio	Reconstruction 1	Reconstruction 2
VAE-Baseline	0%		
VAE-Baseline	75%		
AE-Baseline	0%		
AE-Baseline	75%		

Figure 13: Reconstruction result: VAE and AE baselines on images with different mask ratio for validation.

Appendix E Visualizations for Semantic Segmentation

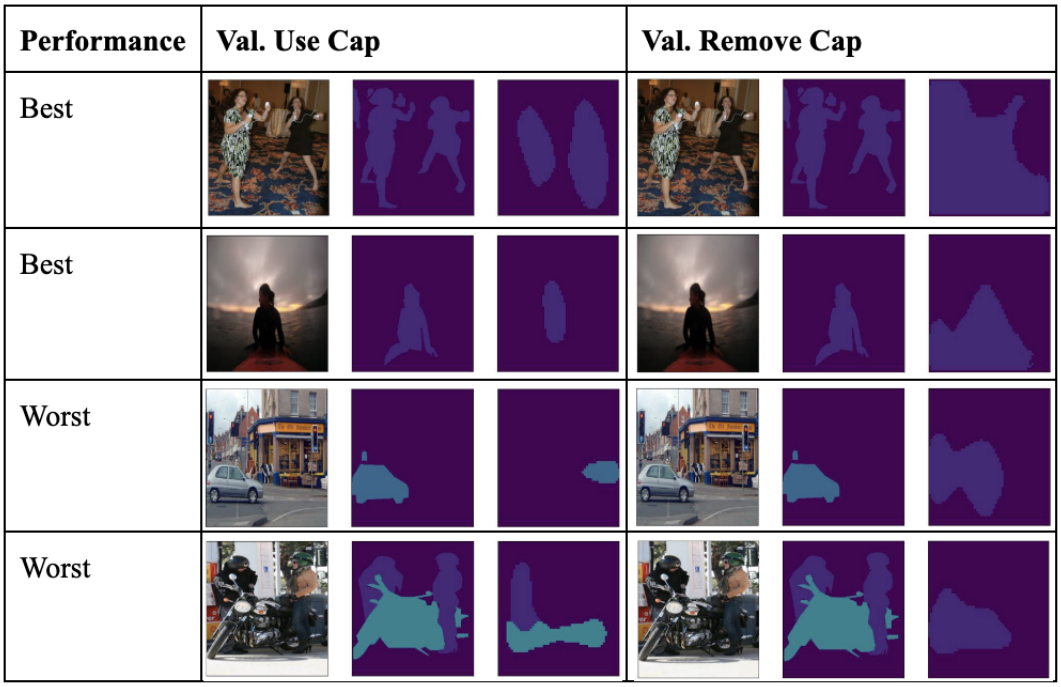


Figure 14: Visualization for semantic segmentation: Semantic segmentation results of LCMVAE on validation. We observe that Caption improves the performance of segmentation results. Our model currently tends towards predicting most/all of the image as background pixels. Poor performance in this downstream task is most likely due to subpar pre-training.

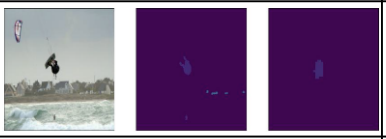
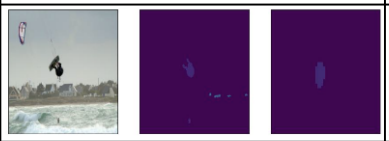
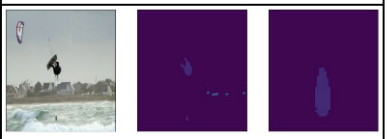
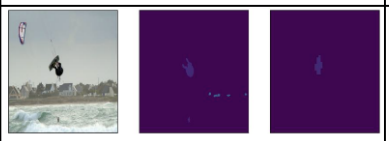
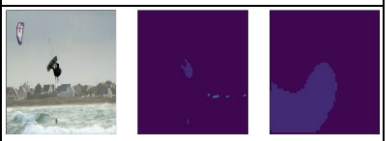
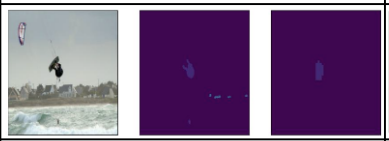
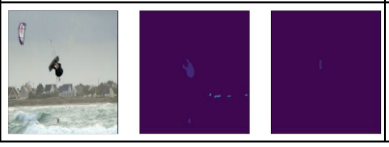
Model	Val. Use Cap	Val. Remove Cap
LCMVAE - Full		
LCMVAE - No Caption		
LCMVAE - No Pre-Conv		
LCMVAE - No Variation		
LCMVAE - No Latent-Reg		
LCMVAE - No Mask		

Figure 15: Visualization for semantic segmentation: Semantic segmentation results of LCMVAE with its variants on validation.