TEXT-TO-IMAGE SYNTHESIS USING CONDITIONAL-VAE

Submitted By:
Jon Campbell Jr., Minh Anh To, Netra Mittal, Sayali Pingle

ABSTRACT

This research explores the use of Conditional Variational Autoencoders (CVAEs) for text-to-image generation, focusing on the integration of textual descriptions with visual outputs. We trained three models using the Fashion-MNIST and MS-COCO datasets, leveraging architectures such as CLIP and Seq2Seq with Gated Fusion to conditionally generate images based on both simple labels and detailed captions. While the models demonstrated the ability to capture high-level semantic features and generate contextually relevant images, limitations were observed in producing fine-grained details and sharpness, especially when conditioned solely on text. Additionally, the project was constrained by access to computational resources, limiting opportunities for more extensive experimentation and fine-tuning. This work provides a foundational exploration of text-to-image generation and highlights challenges and potential improvements for future research in this area.

1 INTRODUCTION

1.0.1 CONDITIONAL VARIATIONAL AUTOENCODER (CVAE)

This research project focuses on developing a CVAE capable of generating images from textual descriptions. The primary goal is to train a generative model that can transform both short-form labels and longer narrative inputs into corresponding visual representations.

The CVAE[1] extends the traditional VAE by incorporating conditioning information (e.g., textual prompts) into both the encoding and decoding processes, enabling targeted image generation aligned with specific inputs. This conditioning enhances the model's ability to produce outputs tailored to the provided context.

The encoder processes input data x and conditional information c (e.g., class labels) to parameterize a latent space distribution, typically a Gaussian with mean μ and variance σ^2. Using the reparameterization trick, the latent variable z is sampled as:

$$z = \mu + \sigma \odot \epsilon,$$

where ϵ is noise sampled from a standard normal distribution, enabling backpropagation through the stochastic process. The decoder combines z and c to reconstruct the output x', conditioned on c. The generative process is modeled as:

$$x' \sim p_\theta(x|z, c),$$

where p_θ denotes the probability of generating x' given z and c. The CVAE minimizes a combined loss function; the Kullback-Leibler divergence regularizes the latent space to align with a Gaussian prior, while the reconstruction loss ensures the output x' resembles x, conditioned on c.

We begin with training the conditional VAE on the Fashion-MNIST dataset, which provides a foundational understanding of generating images from basic textual labels. We then move to the more complex and diverse COCO dataset, leveraging advanced text feature extraction techniques to handle longer text descriptions.

1.0.2 A BRIEF OVERVIEW OF FUSION METHODS

This project explores feature fusion techniques (Pulapakura (2023)) to integrate textual and visual components effectively. Strategies such as additive, concatenative, and gated fusion mechanisms are evaluated to enhance the model's ability to translate linguistic nuances into coherent visual outputs. Fusion optimizes the interaction between textual and visual inputs, ensuring vivid and accurate image generation.

The gated fusion mechanism operates as follows:

[1] Sohn et al. (2015)

- **Input Features:**
  - **Latent Image Features ($z_{\textbf{image}}$):** Encoded by the image encoder and sampled from the VAE's latent distribution.
  - **Textual Features ($z_{\textbf{text}}$):** Encoded by a pretrained text model (e.g., CLIP) and projected into the same feature space as z_{image}.
- **Fusion Process:**
  - **Concatenation:** Combines z_{image} and z_{text} into a joint feature vector.
  - **Gating Layer:** A sigmoid function outputs weights (g) to balance the contributions of each modality:

$$g = \sigma(W_g \cdot [z_{\text{image}}, z_{\text{text}}] + b_g)$$

  - **Weighted Combination:** The fused feature is computed as:

$$z_{\text{fused}} = g \cdot z_{\text{image}} + (1 - g) \cdot z_{\text{text}}$$

- **Decoding:** The fused feature (z_{fused}) is passed to the decoder for final image generation.

2 Methodology and Model Architecture

Each model we employ, uses a slightly modified version of the CVAE structure outlined above. We lay out the specifics down below:

2.1 Fashion-MNIST Model

The Fashion-MNIST dataset from Xiao et al. (2017) comprises two segments: a training set with 60,000 images and a test set consisting of 10,000 images. Each image in the dataset is a grayscale bitmap of 28 by 28 pixels, categorized into one of ten fashion-related classes[2] , ranging from everyday apparel to footwear. Each image in Fashion-MNIST is encoded into a 784-pixel vector, where each pixel is represented by an integer ranging from 0 (white) to 255 (black), depicting the intensity of darkness.

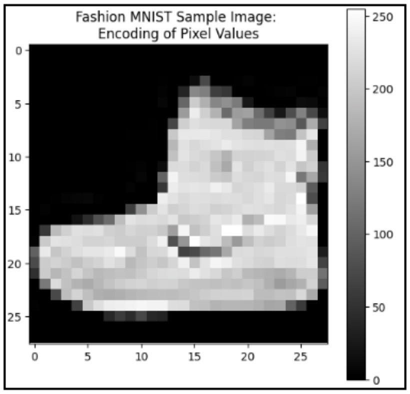

Figure 1: Fashion MNIST Sample Image (Encoding of Pixel Values)

In our implementation:

- **Encoder:** The encoder combines the image data (x) with a conditional input (c) using fully connected layers. A ReLU activation after the first layer fc1 introduces non-linearity, which captures complex interactions between the image features and the conditional input. Linear layers (fc2_mean and fc2_logvar) map features to the latent space, producing mean and log-variance parameters for the **reparameterization trick**.
- **Decoder:** The decoder reconstructs the input (x) by conditioning on the latent variable (z) and conditional input (c). The combined representation of z and c is processed through a fully connected layer with ReLU activation to model complex interactions. A final layer with a sigmoid activation ensures the output lies within $[0, 1]$, making it suitable for image reconstruction tasks.

[2] See Table A in the Appendix for a detailed list of classes.

- **Training process:** Training is performed over 40 epochs using the Adam optimizer. Each training batch involves preprocessing steps such as flattening and rescaling images, generating one-hot embeddings for the labels, and performing forward passes to compute the loss. The total loss is composed of a reconstruction error and the KL divergence, defined as:

$$\mathcal{L} = \mathbb{E}_{q(z|x,c)}[\log p(x|z,c)] - \text{KL}(q(z|x,c)\|p(z)).$$

Gradients are calculated via backpropagation, parameters are updated, and the average loss is logged per epoch. This ensures the CVAE captures data variability and generates outputs conditioned on specific attributes, making it suitable for our controlled image synthesis task.

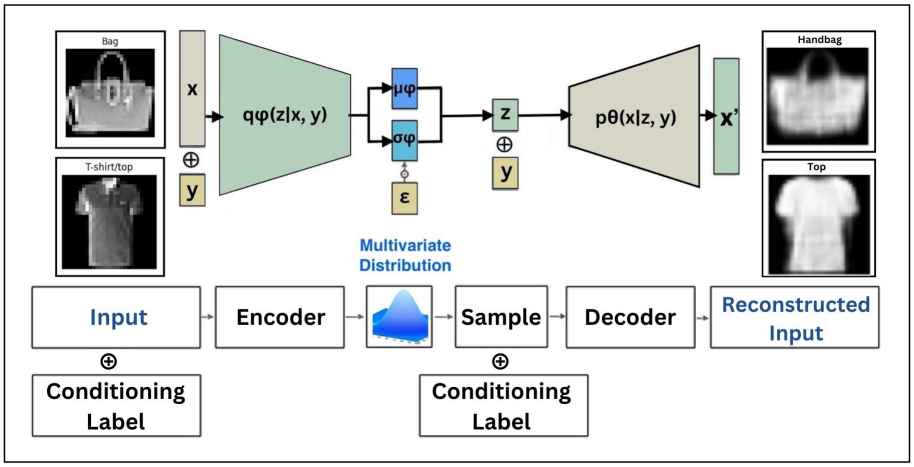

Figure 2: CVAE Model Architecture [modified from Ribeiro et al. (2024)]

2.2 COCO Dataset: Model 1 Using CLIP

The model architecture integrates a Conditional Variational Autoencoder (CVAE) with OpenAI's CLIP model (ViT-B/32) (Radford et al. (2021)) for text-to-image generation. It utilizes the COCO 2014 dataset from Lin et al. (2014), containing 83k images with approximately five captions per image, expanded into 415k unique image-caption pairs. Captions are encoded into 512-dimensional vector embeddings using CLIP, which act as the conditional inputs for the CVAE. Images are preprocessed to 224×224 resolution and converted to RGB format to ensure consistency.

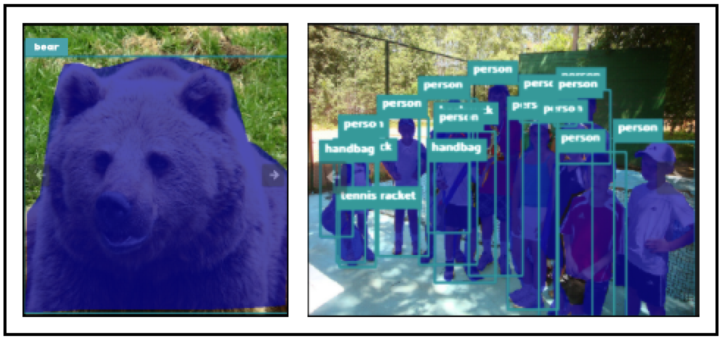

Figure 3: COCO-17 Dataset Sample Image (Dectection and Segmentation)

The model comprises three main components: an image encoder, a main encoder, and a decoder.

- **Image Encoder:** The image encoder uses convolutional layers to reduce the spatial dimensions of the input ($224 \times 224 \times 3$) while increasing the number of feature maps. Each convolutional layer is followed by batch normalization to stabilize training and a LeakyReLU activation function (slope = 0.1) to introduce non-linearity. The final feature map ($14 \times 14 \times 512$) is flattened and passed through a fully connected layer to produce a compact 512-dimensional representation.

- **Text Encoder:** The main encoder concatenates the 512-dimensional image embedding with the 512-dimensional caption embedding. This combined input is processed through a series of fully connected layers, each followed by LeakyReLU activations, to reduce the data to a 128-dimensional latent representation.
- **Decoder:** The decoder then combines this latent representation with the caption embedding, resulting in a 640-dimensional vector. This vector is passed through a fully connected layer, reshaped into a $512 \times 7 \times 7$ feature map, and upsampled to the original image dimensions ($224 \times 224 \times 3$) using transposed convolutional layers.

The model is trained using a reconstruction loss that combines binary cross-entropy and KL divergence. Training is performed with the Adam optimizer, using a learning rate of 5×10^{-3} and weight decay of 10^{-5}, over 200 epochs. This architecture effectively leverages caption embeddings to conditionally control the generated images while maintaining variability through the CVAE framework, making it suitable for high-quality text-to-image generation tasks.

2.3 COCO Dataset: Model 2 Seq2Seq with Gated Fusion

The model employs a Seq2Seq learning framework with gated fusion for multimodal integration, leveraging a subset of the MS-COCO dataset containing 7,000 images paired with five captions each, resulting in 35,000 training samples. At a high level: the Seq2Seq model uses an encoder-decoder structure to map input sequences to outputs. In this project, gated fusion integrates text and image features within the Seq2Seq framework, fine-tuning it to encode captions into 512-dimensional semantic feature vectors. These vectors condition the image generation process within the CVAE framework to produce visually coherent images from textual descriptions. The structure of our model is as follows:

- **Image Encoder:** The image encoder consists of two convolutional layers: the first with 64 filters and the second with 128 filters, both using a 4×4 kernel, stride of 2, and padding of 1. These layers capture hierarchical spatial features, and the output is flattened and passed through fully connected layers to parameterize the latent distribution with mean and log variance vectors. The encoder compresses the visual input into a compact yet descriptive 128-dimensional latent space representation.
- **Text Encoder:** The text encoder processes captions using a pre-trained T5-small transformer. Captions are transformed into 512-dimensional feature vectors by averaging the hidden states from the transformer's final layer. This allows the model to retain rich semantic information from the captions, ensuring the textual inputs effectively condition the image generation process.
- **Gated Fusion Module:** The multimodal integration is handled by a gated fusion module that combines the 512-dimensional text features with the 128-dimensional latent features from the image encoder. A gating mechanism with a sigmoid activation function dynamically adjusts the contributions of text and image features, ensuring a balanced and aligned integration of both modalities. This mechanism enhances the ability of the model to generate reconstructions that align well with both inputs.
- **Decoder:** The decoder reconstructs the final image by first reshaping the fused latent representation into a high-dimensional feature map through a fully connected layer. Transposed convolutional layers progressively upsample the feature map to recreate the spatial dimensions of the original image ($224 \times 224 \times 3$). A final $\tanh$ activation function scales the pixel values to the range $[-1, 1]$.

The model is trained using a weighted loss function combining MSE for reconstruction accuracy and KL divergence to regularize the latent space. Training is optimized using the Adam optimizer with a learning rate of 10^{-3}.

3 Results and Discussion

3.1 Fashion-MNIST: Generative Image Synthesis from Short-Form Textual Labels

3.1.1 Reconstruction Results

To evaluate the model's ability to learn meaningful latent representations, we visualized reconstructed images alongside their original counterparts. Using the trained CVAE, a batch of test images was flattened, combined with one-hot label embeddings, and passed through the Encoder and Decoder to generate reconstructions. The reconstructed images were then plotted against the original inputs for direct comparison (Figure 4) 4.

As shown in the figure, the CVAE successfully captures the overall structure and distinguishing features of the original inputs, such as the general shape and outline of the objects. However, the reconstructions exhibit some blurriness, which indicates potential room for improvement in capturing fine-grained details.

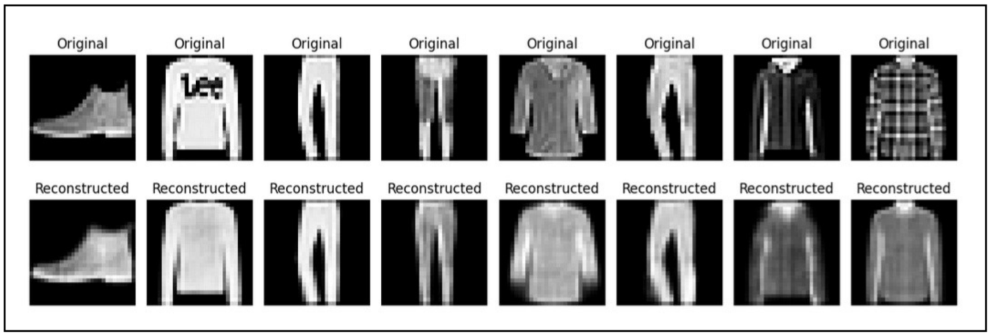

Figure 4: Reconstruction Quality Assessment

3.1.2 Generation Results

To evaluate the CVAE's ability to generate images conditioned on specific labels, we sampled latent variables and combined them with text embeddings corresponding to predefined categories such as "sneaker," "boot," "top," and "handbag." The model then generated images based on this input, which were visualized in rows for each label (Figure 5) 5.

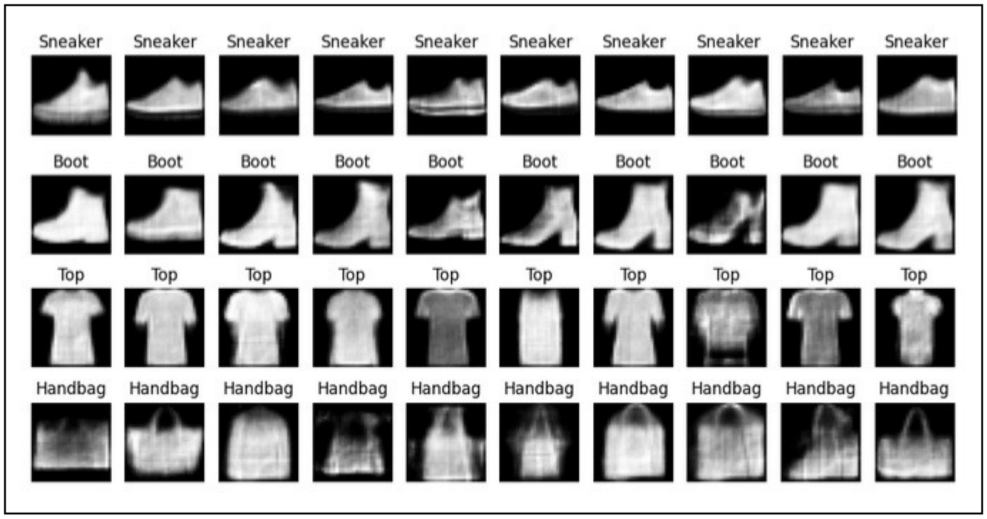

Figure 5: Generated images conditioned on specific labels ('sneaker,' 'boot,' 'top,' 'handbag')

As shown in the figure, the CVAE successfully generates images that align with the intended categories, capturing the overall structure and recognizable features of the specified objects. However, the generated images exhibit some variability in quality and detail, particularly in fine-grained features like texture. This highlights the model's potential for conditional image synthesis while suggesting areas for improvement in generating sharper, more detailed outputs.

3.2 COCO Dataset: Image Generation through Textual Narrative Summarization

3.2.1 Model 1 Results Using CLIP

After 200 epochs of training, the CVAE demonstrates its ability to generate images from noise and captions, though the outputs are still somewhat abstract. Below are examples of generated images with corresponding captions:

- Caption 1: a futuristic cityscape with flying cars and neon-lit buildings
- Caption 2: a serene beach with clear water and swaying palm trees

- Caption 3: a majestic tiger walking through a sunlit jungle
- Caption 4: a cozy cabin in a snowy forest with glowing lights
- Caption 5: a robot chef cooking a colorful gourmet dish in a high-tech kitchen

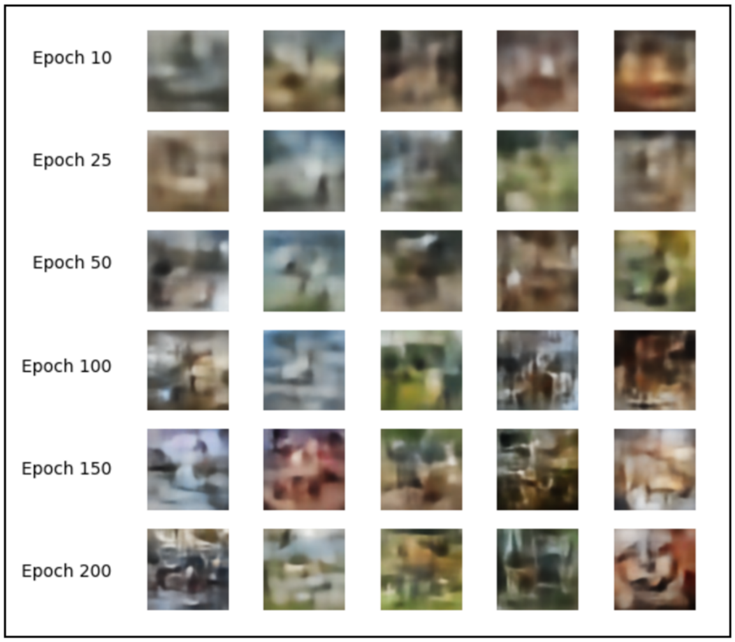

Figure 6: Generated images using the CVAE decoder

The figure below illustrates examples from the training process:

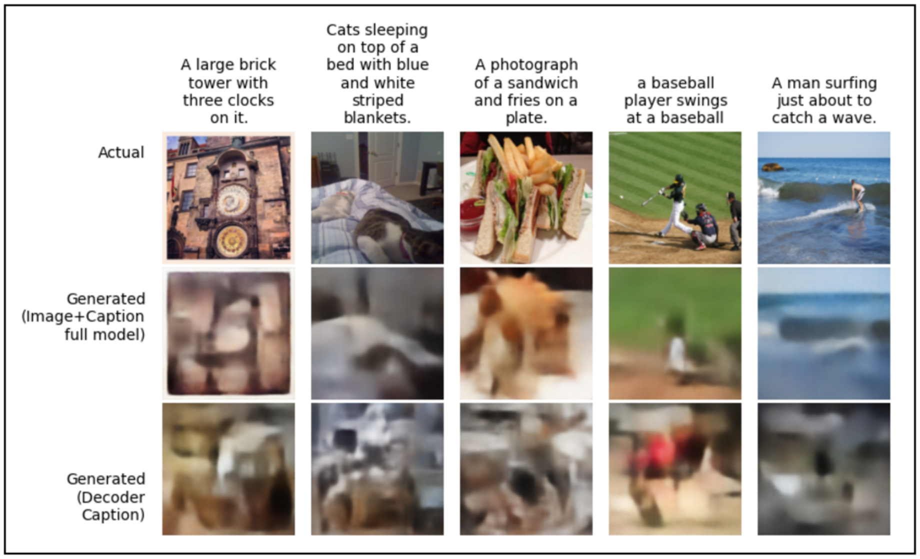
A large brick tower with three clocks on it.
Cats sleeping on top of a bed with blue and white striped blankets.
A photograph of a sandwich and fries on a plate.
a baseball player swings at a baseball
A man surfing just about to catch a wave.
Actual
Generated (Image+Caption full model)
Generated (Decoder Caption)

Figure 7: Images generated with the full model vs. the decoder using caption embedding and noise

The top row shows the original images from the dataset. The middle row contains images generated by passing the original images and caption embeddings through the full CVAE model. The bottom row depicts images generated solely from the caption embeddings. These examples highlight that the CVAE can approximate the original images when both image and caption embeddings are provided. However, the model struggles to generate coherent visuals when conditioned only on captions. This limitation likely stems from the pixel-wise reconstruction loss, which minimizes the error between individual pixel values but overlooks higher-level features critical for perceptual realism.

To address these limitations, incorporating a perceptual loss function could improve image quality. Unlike pixel-wise reconstruction loss, perceptual loss evaluates differences in high-level feature representations extracted from pre-trained models like VGG. This approach encourages the model to focus on structural and semantic similarities, resulting in more visually realistic and coherent reconstructions.

Next, we evaluate the CVAE's ability to generate images based on captions not encountered during training.

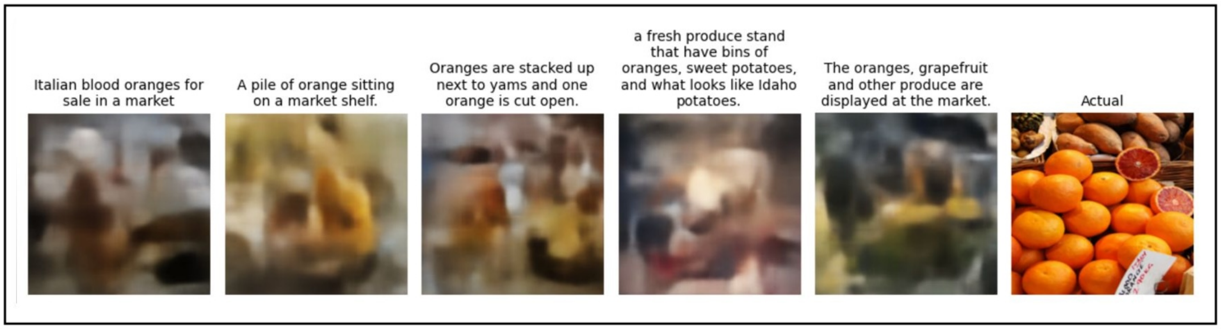
Italian blood oranges for sale in a market
A pile of orange sitting on a market shelf.
Oranges are stacked up next to yams and one orange is cut open.
a fresh produce stand that have bins of oranges, sweet potatoes, and what looks like Idaho potatoes.
The oranges, grapefruit and other produce are displayed at the market.
Actual

Figure 8: Generated images from unseen captions

The model demonstrates an ability to capture high-level semantic information, as seen in the following examples:

- The first two captions include only the word "orange," resulting in images that prominently feature orange hues.
- The third caption incorporates "orange" and "yam," leading to similar color themes in the generated images.
- The final caption contains the word "other," which could explain why the corresponding image exhibits less orange coloration.

These results suggest that the model is sensitive to specific keywords in the captions, influencing the color and thematic elements of the generated images. With additional training and refinement, the model holds potential for generating higher-quality and more diverse images.

3.2.2 Model 2 Results using Seq2seq with Gated Fusion

The results illustrate the progression of the Conditional-VAE model's ability to generate contextually relevant images from textual descriptions. After 40 epochs, the model's outputs for the caption "A car at a town centre" reveal its initial learning phase. While the generated samples capture some coarse features related to the input caption, such as vague urban-like shapes and muted color tones, the images are highly blurry and lack sharp details. This stage reflects the model's ongoing struggle to balance the integration of textual and visual modalities, with the latent space yet to fully capture meaningful representations of the input data.

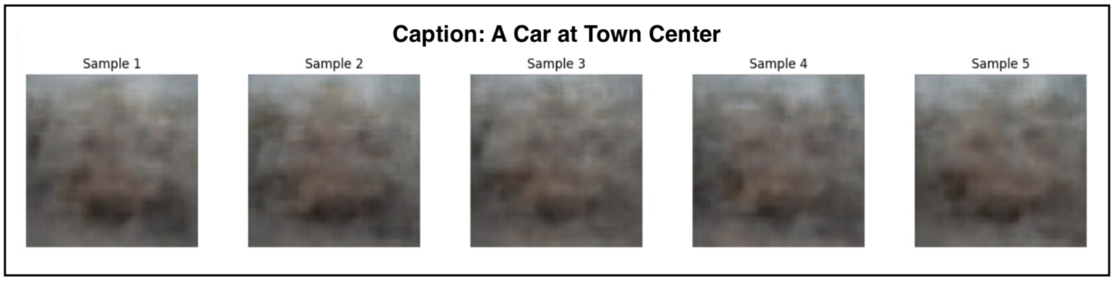

Figure 9: COCO M2, Generated images from captions (epoch= 40)

In contrast, the results after 100 epochs for captions such as “A beautiful landscape with mountains and a lake”, “A black car”, and “An orange tree” demonstrate significant improvements. The generated images align more closely with the descriptions, as evident in the landscape outputs, which feature discernible patterns of greenery and sky, and the orange tree outputs, where distinct orange hues emerge prominently. However, even after extended training, the images remain somewhat blurry, highlighting the model’s challenge in achieving pixel-level sharpness. These results suggest that while the Conditional-VAE effectively bridges text and image modalities, further enhancements, such as incorporating perceptual losses or fine-tuning the architecture, could improve the quality and detail of the generated outputs.

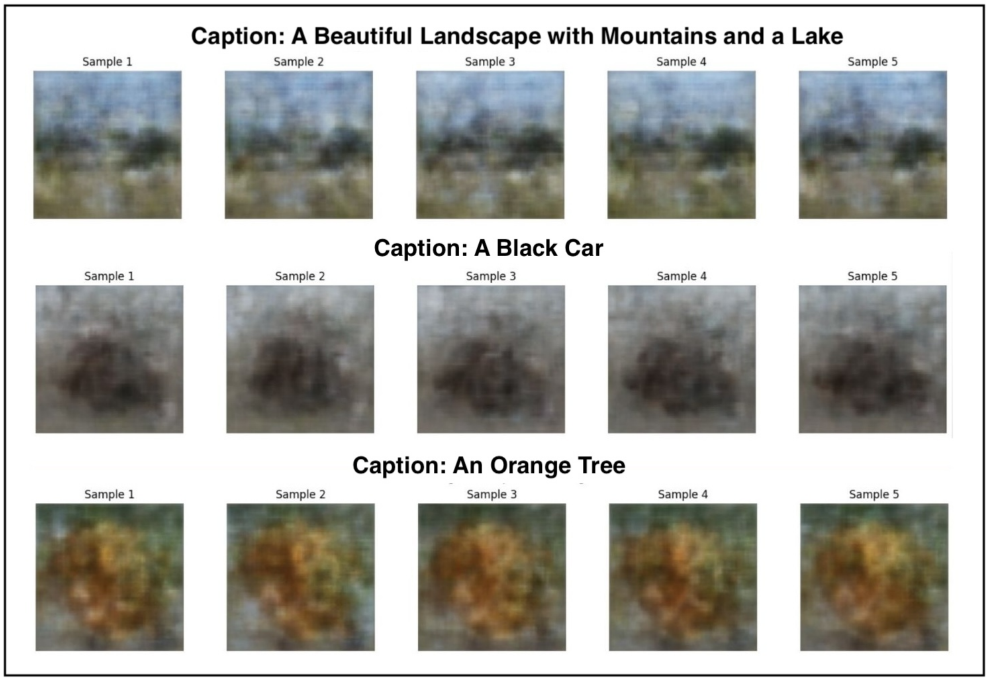

Figure 10: COCO M2, Generated images from captions (epoch= 100)

4 Conclusion

This research demonstrates the potential of Conditional Variational Autoencoders (CVAEs) for text-to-image synthesis, effectively bridging textual descriptions and visual outputs. Using the Fashion-MNIST and MS-COCO datasets, the models showcased an ability to capture high-level semantic features and generate contextually relevant images. Techniques like CLIP and Seq2Seq with Gated Fusion enhanced the integration of text and image modalities, allowing for richer and more precise image generation. However, challenges remain, including the blurriness and lack of fine-grained details in generated images, particularly for complex captions. Limited computational resources further constrained the scope of experimentation, restricting opportunities for extensive hyperparameter tuning and model refinement.

Future work could explore the use of perceptual loss functions to improve the realism and coherence of generated images by focusing on high-level structural features rather than pixel-wise accuracy. Incorporating larger, more diverse datasets and leveraging advanced architectures like diffusion models or transformer-based generators may also enhance the quality and diversity of outputs. Additionally, deploying distributed or cloud-based computational frameworks could help overcome resource limitations, enabling more extensive training and evaluation. This work provides a foundational exploration of text-to-image synthesis and highlights key areas for further research and development.

References

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Coco: Common objects in context. `https://cocodataset.org/#home`, 2014. Accessed: December 5, 2024.

Raj Pulapakura. Fusion method: Multimodal models and fusion – a complete guide. https://medium.com/@raj.pulapakura/multimodal-models-and-fusion-a-complete-guide-225ca91f6861, 2023. Accessed: December 5, 2024.

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. Learning transferable visual models from natural language supervision. https://github.com/openai/CLIP, 2021. Accessed: December 5, 2024.

Tiago Ribeiro, Fernando Silva, and Rogério Luís Costa. Modelling forest fire dynamics using conditional variational autoencoders. *Information Systems Frontiers*, 06 2024. doi: 10.1007/s10796-024-10507-9.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015. URL http://papers.neurips.cc/paper/5775-learning-structured-output-representation-using-deep-conditional-generative-models.pdf.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

A Appendix

Table A: Labels for Fashion-MNIST Dataset

Label	**Description**
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot