CSEN 342 Project Report

Neural Image Compression with Text-guided Encoding for both Pixel-level and Perceptual Fidelity

Team Name: **Python Coders**
Robin Lee W1597796
Lucas Van Der Heijden W1619420
Chan Nam Tieu W1629784

Introduction:

In the era of digital communication and multimedia content, the efficient storage and transmission of visual data has become increasingly critical. Traditional compression methods like JPEG and PNG, while widely adopted, often struggle to balance compression efficiency with image quality preservation. This limitation has sparked considerable interest in leveraging deep learning approaches for image compression, often leading to sophisticated autoencoder architectures or entropy models (VAEs) to solve this problem.

One non-intuitive discovery is that image compression techniques can be improved with the additional context of text captions for the image. The text captions allow a great deal of information about the image to be stored in a small number of bits. When this channel is combined with the image embedding and/or latent space, greater performance is achieved in downstream encoding and decoding. However, models utilizing these captions end up with great variability (either with unwanted artifacts or loss of detail) in their outputs, making them unsuitable for accurately reconstructing the original image.

The authors propose TACO, a model that utilizes ELIC encoder-decoder model as a backbone and only uses the text captions for only the encoding step of the image. not decoding it. The theory behind this is that the model will be forced to learn only high level information from the caption, reducing variability as the caption cannot be used to construct individual pixel regions. Thus reducing unwanted artifacts while preserving small details and textures.

As student researchers, we aimed to improve upon this paper's results.

Approaches:

Before seeking how to improve the TACO compression model, it is important to first understand it. As described beforehand, TACO consists of encoder-decoder architecture (provided by ELIC) and utilizes both the image and its associated caption as an input.

Model Used:

ELIC (Efficient Learned Image Compression with scalable residual nonlinearity) serves as the backbone for TACO. From a high level, it is an encoder-decoder model that utilizes residual connections and attention layers in order to preserve important and specific attributes from a given image embedding. It is with this logic that an accurate reconstruction of the image is created via the decoder from its latent representation.

TACO utilizes this architecture and adds an associated image caption as a component to the input. Using a CLIP encoder, the input to the network would be the concatenated result of an image embedding and the corresponding word embedding of the caption.

Dataset Usage:

For training, the training split of the MS-COCO dataset was used, consisting of 82,783 images with five human-annotated captions for each image. All 5 captions were used for training. Additionally, each image was randomly cropped to a 256x256 resolution.

For evaluation and validation, a mixture of MS-COCO validation set, CLIC, and Kodak were used. A "restval" subset of the MS-COCO validation dataset (~30,504 images) were center cropped to a 256x256 resolution in order to preserve semantic meaning in the associated image captions. The CLIC and Kodak datasets were left uncropped (as they were either fitted to 256x256 resolution or could be losslessly converted) and used as it. However, the CLIC and Kodak datasets did not have accompanying image captions. These were generated by the Once-for-All (OFA) Network that is publically available.

With this, the authors point out several limitations of their approach. Firstly, the additional encoding cost for processing text increases quadratically with the sequence length due to the design of their text injection mechanism, which computes cross-attention between image and textual features. This can lead to prohibitively long encoding times in cases where the text sequence is particularly lengthy, such as video-like data with voice transcripts. Secondly, they highlight the issue of training scalability, noting that their model, TACO, is currently trained using an image-text dataset, which is significantly less abundant compared to image-only or text-only datasets. Given this as a baseline, we additionally thought of other ways the architecture could be improved upon. For example, improvements upon the ELIC architecture or diversifying the dataset.

All of the approaches were created independently and can be thought of as multiple ablation studies.

Increased Contextual Information:

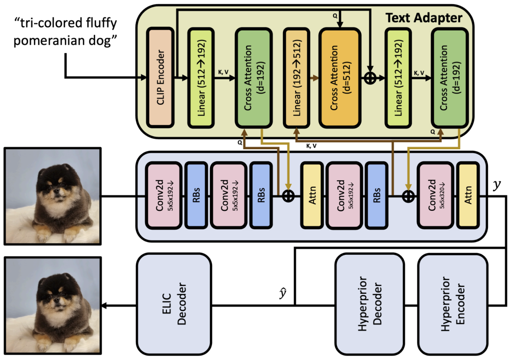
Robin made improvements to the TACO's compression pipeline by increasing the contextual information in captions for each image. This comes in the form of increasing the caption lengths to describe the scene and providing context that reasons about the scene.

In the original TACO architecture, the caption lengths are relatively short at 38 tokens, which is around 15 words. We assumed that this was done to mitigate the O(n^2) attention mechanism time cost. However we believe that 15 words is not enough to fully describe the picture and fully create a latent representation that can reconstruct fine details. There could be an argument that having longer captions would contribute noise, however this is highly dependent on the content of the words in the caption. With our implementation, we believe that there is a limited noise in our caption.

In addition to increasing caption lengths, an additional context of reasoning about the image was also provided. The main purpose of this was in order to create a more complex and therefore more information packed latent representation. We believe these additional relationships contribute to greater differentiation and higher quality detail reconstruction.

Improved Attention Layer Performance:

Lucas made improvements to the model mainly at the architectural attention layer. The authors describe this as being the main bottleneck to their model because traditional attention is an O(N squared)operation. However, Lucas introduces a linear form of attention where it makes substantial, faster, conversational efficiency, especially on inference and training.
The figure below shows the structure of the original TACO model:



In particular, the text adapter utilizes cross attention layers to insert encoded text information into the encoder. These cross attention layers have an O(n^2) complexity. Currently this is not an issue as the CLIP Encoder only allows for 77 token max length inputs and the tested captions were relatively short, about a sentence each. However, if the CLIP encoder were swapped for a different one, and the model architecture was similarly changed to handle long captions, the inputs into these cross attention layers would increase in size. This could in turn cause the O(n^2) complexity layers to become a significant source of computational overhead.

Improved Caption/Word Encoding:

Chan enhances CLIP-generated embeddings with LLM-based keyword extraction to overcome CLIP's deficiency in capturing word importance. Unlike vision-language models that learn meaning through inter-image relations, text is independently processed by LLMs, identifying significant words more accurately. With the application of TF-IDF and Named Entity Recognition, significant terms are emphasized within the embeddings, boosting robustness and accuracy, especially for lengthy captions.

Methodology:

Architecture / Dataset Improvements:

For *Increased Contextual Information*, entirely new captions were one-shot generated by OpenAI's GPT-4o. This was done on all of the images in the training and validation datasets (MS-COCO 2014, CLIC, Kodak). Examples of these query/result pairs can be found in the appendix section A1. For the training dataset, 5 captions were generated per image. This was done in a similar way as the original dataset and curbed the variability of grammar and noise and fully extracted the sentiment of the captions.

For *Improved Attention Layer Performance*, new attention mechanisms were used in place of the standard attention mechanisms present in the original TACO implementation. The linear attention mechanism described in the “Linformer: Self-Attention with Linear Complexity” (Wang, et al.) was used. This version of the attention mechanism promises linear complexity scaling or O(n) while roughly maintaining the performance of the original attention mechanism.

For *Improved Caption/Word Encoding*, CLIP-generated text embeddings are enhanced with TF-IDF weighting and Named Entity Recognition (NER) to emphasize important words that have an impact on compression. TF-IDF vectorizer assigns higher weights to informative words and eliminates stopwords, and NER (with spaCy) detects proper nouns and objects (e.g., "Eiffel Tower," "Golden Retriever") to emphasize important words even more. These extracted keywords form a token-weight matrix, which increases the weight (1.5) of important tokens in CLIP's embeddings, preventing uniform token treatment and increasing the effect of important words.
During training, weighted sum pooling refines text embeddings from CLIP's encoder such that salient words contribute more significantly to the final representation. L2-normalized last hidden state embeddings are weighted by the normalized token-weight matrix, which proportionally scales each token's contribution. The weighted embeddings are then element-wise multiplied and summed along the token dimension, boosting the contribution of semantically meaningful words while preserving contextual coherence.

Training and Evaluation:
With all of our varying approaches, our training methodology is relatively similar to the TACO’s training procedure. After our architectural or data changes are made, we follow a straightforward pipeline.

With our augmented network (Improved Attention Layer Performance, Improved Caption/Word Encoding) we train on the corresponding training split of the MS-COCO dataset with five human-annotated captions for each image. All 5 captions were used for training. Furthermore, each image was randomly cropped to a 256x256 resolution. Then, for evaluation and validation, a mixture of MS-COCO validation set, CLIC, and Kodak were used and evaluation was done exactly as the original papers.

For our augmented dataset (Increased Contextual Information) a similar process was done. Using the original TACO network, it was trained on the caption-augmented training dataset. Furthermore, it was evaluated the same way for the validation dataset.

These processes were mainly done on the WAVE HPC’s NVIDIA Tesla V100 or NVIDIA GTX 1080 ti GPU partitions. Additionally, more computation was done on Google Cloud via GCP.

Result Validation:
To ensure we fairly evaluate results, we validated the authors’ findings. For this, we used the codebase and model weights provided by the authors. With this, we evaluate their network

performance by recording metrics (LPIPS, PSNR, FID) on the same test data splits (MS-COCO, CLIC, and Kodak) that the authors used.

From our experiments, their results match our recomputed results with some variability.

Experiments and Analysis:

Experimental Setup:

The experimental set up was the same for all of our approaches and experiments, with regards to testing for accuracy. After we all made our augmentations, we trained our respective models at different λ values (which determines the compression rate of the final model) noted in the original paper and evaluated them using our evaluation metrics.

For the linear attention mechanism improvement, the change in inference and training time was also measured. The details of how this was done are listed below, in the relevant subsection.

Increased Contextual Information:

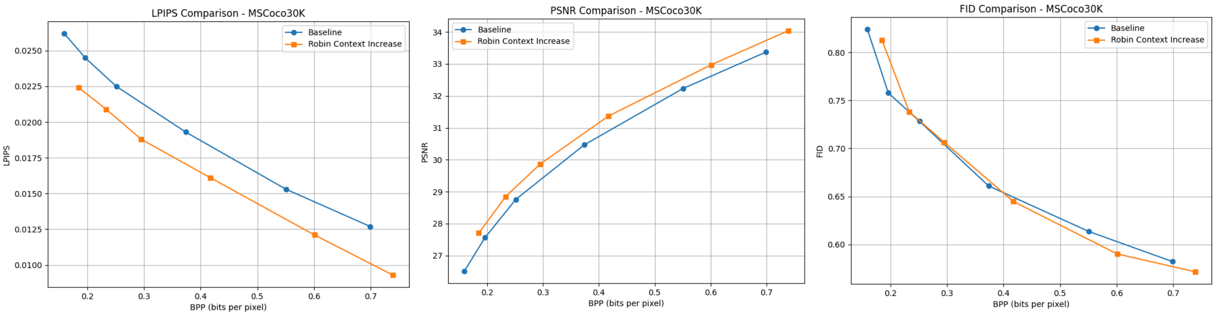


Fig. 1 Compression results: MS-COCO. Increasing Context of captions outperforms the TACO's baseline in LPIPS and PSNR. It also achieves competitive results in TACO's baseline FID.

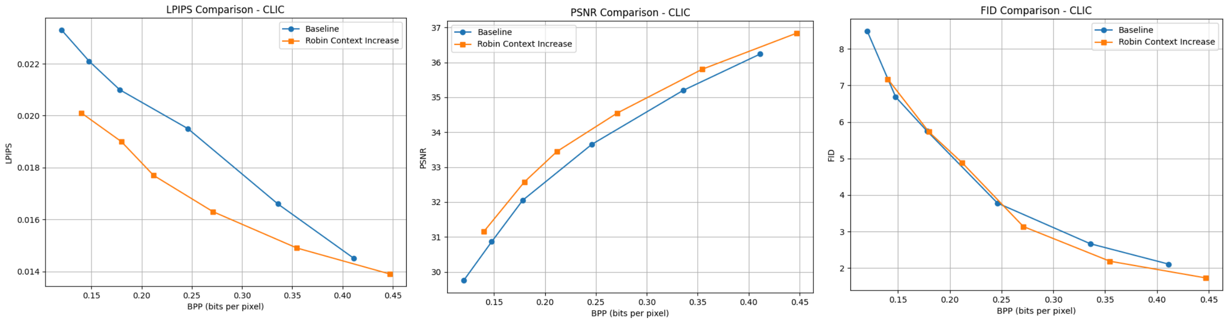


Fig. 2 Compression results: CLIC. Increasing Context of captions outperforms the TACO's baseline in LPIPS and PSNR. It also achieves very similar results in TACO's baseline FID.

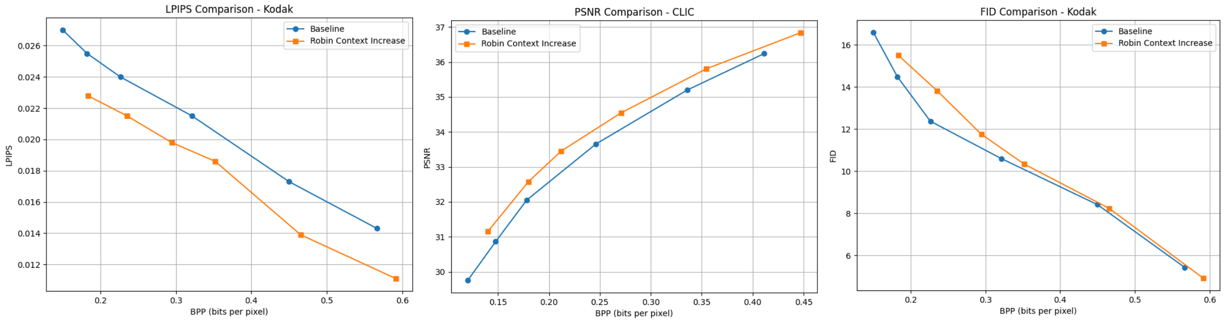


Figure 3. Compression results: Kodak. Similar to MS-COCO and CLIC, Context Increase outperforms or matches very closely with TACO baselines.

In our evaluation, this increase in contextual information positively impacts the TACO architecture. This augmentation outperforms the baseline TACO model across various datasets with respect to LPIPS. When it comes to pixel fidelity, measured by PSNR, our augmentation is slightly better than the TACO baselines. In terms of FID and realism, our augmentation retains a similar score to the original TACO baseline in all datasets.

These results are consistent with our predicted outcomes. The improvements to the LPIPS score can be attributed to the more complex latent space representation. Thus creating a more accurate reconstruction of the image. The improvements to the PSNR score can be attributed to the reasoning element of the captions, thus capturing a higher level of detail in its representation. The competitive values of the FID score is expected as the general realism from the original architecture (lack of artifacts and blurriness) is retained.

It is important to note that the benefits of this method will eventually converge. This is due to the fact that captions can only describe an image so well, before it contributes to noise instead. In order to mitigate this, the content within captions should be scrutinized.

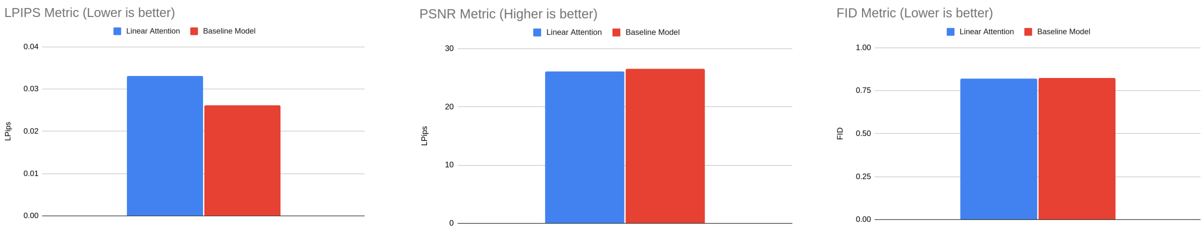
Linear Attention Mechanism:

The change in inference and training time was measured over 5000 image-caption pairs. The average time taken per pair is listed in milliseconds. The experiments were done on a RTX 4090 GPU with 24 gB of vram.

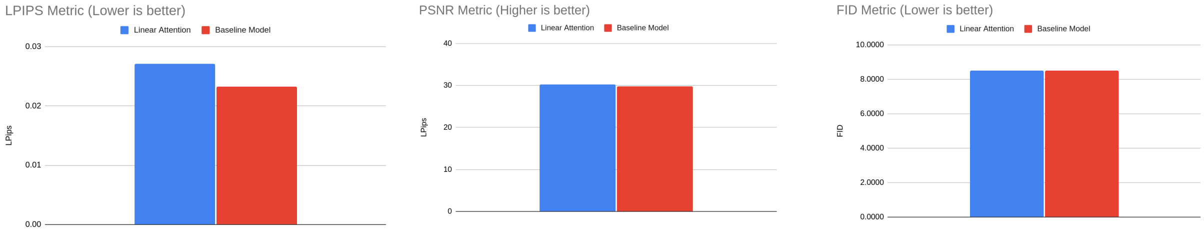
The time taken was measured against the size of the inputs into the cross attention layers. However, changing the size of these inputs was not as trivial as changing the input length. The architecture of the model ensures that the input size will only ever be 192, regardless of the input text length. Because of this, the model architecture must be changed to change the size of the cross-attention inputs. For each cross-attention layer there are two sources of inputs, the input from the encoded text, and the features from the image. The text input is already scaled by a linear layer, so changing the output of that layer is simple. But, the image input size is determined by the number of filters in the convolutional layers. This was changed by simply concatenating the outputs to create larger ones for the purposes of experimentation. The output of the cross-attention layers was then scaled to original size by simply dropping the unnecessary dimensions.

The results are as follows:

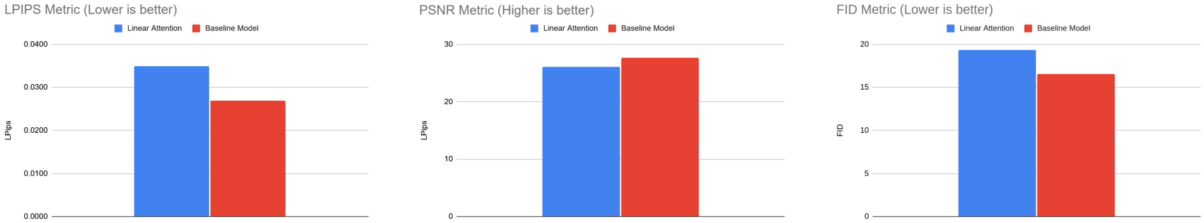
MS-COCO



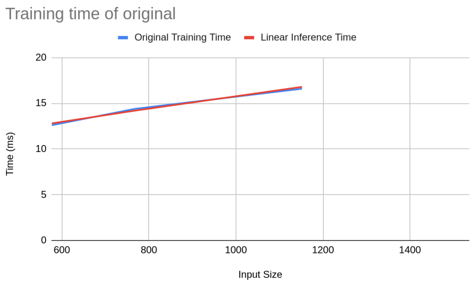
CLIC

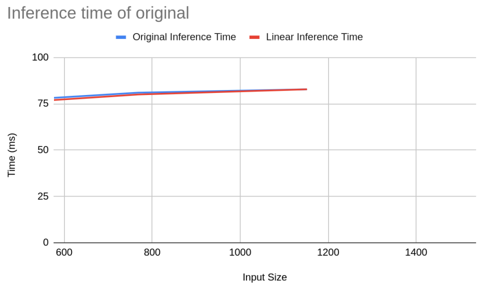


Kodak



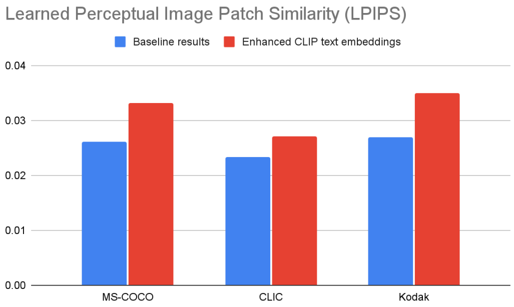
As seen above, the accuracy drops below that of the original. The performance drop is not significant, but it is noticeable. It is worth noting that this model was only trained for 30 epochs as opposed to the 50 that the authors trained theirs for. This was due to time constraints as the model would take 3-4 days to train for 50 epochs. Further training would likely improve the model's performance marginally, as it seemed like it might be underfitting when training was stopped. However, based on the trend during training, the performance improvement would have been minor, and not enough to put the linear model on par with the original.

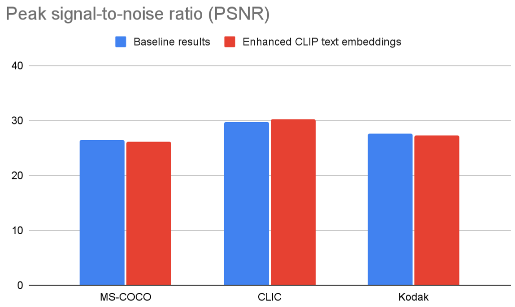


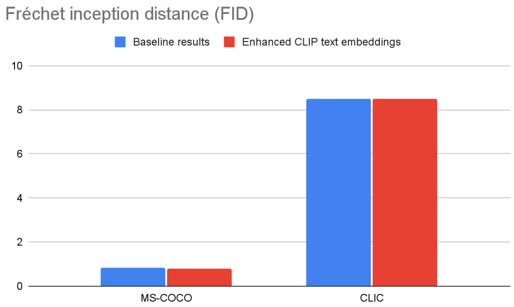


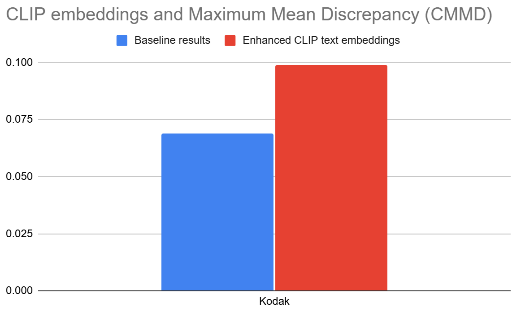
As the results above show, there is no real difference between the linear and the original attention mechanism. This indicates the cross-attention layers, despite the O(n^2) complexity, are not the primary source of computational complexity. As a result, performance improvements must be found elsewhere. Additionally, this means that the performance cost of using the linear attention mechanism does not work for this model.

<u>Improved Caption/Word Encoding:</u>









Perceptual similarity (LPIPS) trade-off: LPIPS measures perceptual differences, where lower is better. Our enhanced CLIP text embeddings resulted in slightly higher LPIPS scores across all datasets (MS-COCO: 0.0262 → 0.0332, CLIC: 0.0233 → 0.0271), indicating minor perceptual differences in reconstructed images. However, the increase is small, suggesting a minimal trade-off while maintaining competitive perceptual quality.

Pixel-wise fidelity (PSNR) impact: PSNR, which measures pixel-level similarity, showed a slight decrease for MS-COCO and Kodak but an increase for CLIC. This suggests that text-guided enhancements introduce small pixel variations, especially in diverse datasets. However, the ≤ 0.5 dB change remains within an acceptable range, ensuring practical fidelity.

Fréchet inception distance (FID) improved for MS-COCO (0.8241 → 0.8157), indicating better alignment with the original distribution, while minor increase in CLIC suggests some distributional shifts. *CMMD* (CLIP embeddings & MMD), a more robust alternative to FID, slightly increased for Kodak (0.069 → 0.099), suggesting subtle high-level feature changes. Given CMMD's greater sample efficiency and unbiased nature, further investigation is needed to interpret these shifts in relation to human perceptual quality.

Since the original captions used for weighting are relatively short and concise, the impact of our approach remains subtle. With longer captions providing more context, we expect greater improvements, as the enhanced embeddings would have more meaningful information to influence the compression process.

It is also worth noting that due to hardware constraints on the HPC, the maximum job duration is limited to 48 hours, with access restricted to only 2 GPUs, whereas the original authors likely used at least 5. Since training one epoch takes approximately 100 minutes, completing 50 epochs as in the original paper would require over 83 hours, exceeding the allowed runtime. Therefore, training was only conducted for 27 epochs with λ=0.0004 within the 48-hour limit, and these constraints should be considered when interpreting the results.

Conclusions:

Our research presents several innovative approaches to enhance the TACO image compression model. Through our three distinct improvement strategies - increased contextual information, improved attention layer performance, and enhanced caption/word encoding - we have demonstrated that there is significant potential for advancing text-assisted image compression techniques.

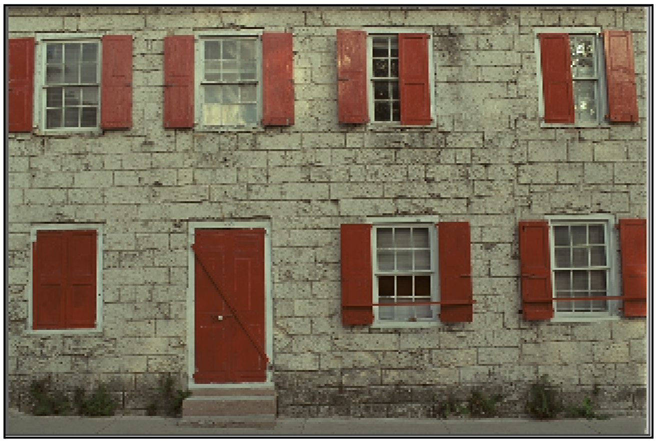
Our investigation into increased contextual information through longer, more detailed captions suggests that richer textual descriptions can lead to more nuanced latent representations, potentially improving the reconstruction of fine details. The implementation of linear attention mechanisms has shown that the computational complexity of the cross-attention layers is minimal compared to the rest of the model. Additionally, the enhancement of CLIP-generated embeddings through LLM-based keyword extraction represents a novel approach to improving the quality of textual feature representation.

While these improvements show promise, they also highlight areas for future research. Future work might focus on developing more efficient attention mechanisms, exploring alternative text embedding techniques, or investigating the potential for multimodal compression strategies that incorporate additional types of contextual information. The insights gained from this research could also inform the development of more sophisticated image compression systems that better serve the needs of modern digital communication.

Appendix:

A1. Using GPT-4o (ver. 2024-11-20) with temperature: 1.

Example Query:
Describe the image in detail. Reason about the image. Return the response in paragraph form.



Example Response:
The image depicts the front facade of an old, weathered stone building with uneven textures and an aged appearance. The wall features a mix of light beige and gray tones, with a rustic, slightly distressed finish that suggests its historical character. The building has two stories, each with symmetrical red shutters framing the windows, except for a few that are closed. The windows, painted white, are simple in design and accentuate the vintage charm. A matching red door is situated at the bottom center, accessed by a small stone step. The overall harmony of red and white contrasts beautifully with the coarse stone exterior, creating a visually striking aesthetic. Stray greenery is sparse, and the sidewalk in front of the building appears clean and understated, emphasizing the structure’s timeless appeal.

Citations:

Lee, H., Kim, M., Kim, J. H., Kim, S., Oh, D., & Lee, J. (2023). Neural Image Compression with Text-guided Encoding for both Pixel-level and Perceptual Fidelity.

He, D., Yang, Z., Peng, W., Ma, R., Qin, H., & Wang, Y. (2022). ELIC: Efficient Learned Image Compression with Unevenly Grouped Space-Channel Contextual Adaptive Coding.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision.

Wang, S., Li, B., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-Attention with Linear Complexity.