National College of Ireland

# Patch-Based Generative Reconstruction of Task-Oriented Images for Bandwidth-Efficient Communication

MSc Research Project
MSc in Artificial Intelligence

Sameera Sudath Tennakoon Wanninayaka Tennakoon Mudiyanselage
Student ID: 23399031

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Sameera Sudath Tennakoon Wanninayaka Tennakoon Mudiyanselage |
| **Student ID:** | 23399031 |
| **Programme:** | MSc in Artificial Intelligence |
| **Year:** | 2025 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Vikas Sahni |
| **Submission Due Date:** | 11/12/2025 |
| **Project Title:** | Patch-Based Generative Reconstruction of Task-Oriented Images for Bandwidth-Efficient Communication |
| **Word Count:** | 7168 |
| **Page Count:** | 23 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Sudath Tennakoon |
| **Date:** | 11th December 2025 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Patch-Based Generative Reconstruction of Task-Oriented Images for Bandwidth-Efficient Communication

Sameera Sudath Tennakoon Wanninayaka Tennakoon Mudiyanselage
23399031

**Abstract**

In bandwidth limited environments such as space and underwater explorations, search and rescue missions, etc., sending full images is challenging. Limited data rates, power and signal issues make it hard to transmit full resolution images. This research addresses the gap in current techniques by presenting a novel, bandwidth usage efficient pipeline that balances data reduction with image quality. Identifying and sending only the important parts of an image can significantly reduce the amount of data required for transmission. These parts are chosen using Saliency maps from task specific models. The selected parts are combined into a stitched grid and compressed by sending only the codebook indices of Vector-Quantized Generative Adversarial Network (VQ-GAN). On the receiver side, VQ-GAN decoder and pretrained Masked Autoencoder (MAE) fill the missing regions. Experimental results show that the reconstructs images with higher visual quality and improved pixel-wise metrics than the baseline. With VQ-GAN compression, images still closely match baseline performance while achieving compression ranging 20× to 160×, demonstrating the effectiveness of this method under severe bandwidth constraints.

## 1 Introduction

Task-oriented communication is very important in real-time, remote, and resource constrained settings. In these kinds of situations, communication bandwidth is limited and critical. As an example, most of NASA's Mars rovers use relay communication as their primary and direct-to-Earth (DTE) communication as their secondary method. These technologies have strict limitations in bandwidth and also they are constrained by short communication windows and limited energy availability.[1] These factors make large scale image communication very challenging. Similarly, underwater exploration robots struggle with low bandwidth. Communication systems of these depend on acoustic signals that support very low data rates over long distances. Even the most advanced acoustic modems face lot of limitations when transmitting large volumes of data efficiently.[2] This especially comes in to play when dealing with high resolution imagery. These constraints

[1]Space Communications: `https://www.nasa.gov/missions/tech-demonstration/space-communications-7-things-you-need-to-know`

[2]Underwater Communication Methods: `https://www.waterlinked.com/applications/underwater-communication`

further highlight the importance of transmitting only task relevant data in such environments.

Also, task-oriented communication is important for delay sensitive applications. Because quick decision making is crucial for these situations. Timely communication of critical visual data in disaster rescuing operation can be lifesaving. Also, robots that are used for search and rescue missions often operate in hazardous areas. So normally under these kinds of situations, most times network connections are unstable or congested and have very limited bandwidth. But despite this, robots must quickly transmit critical information about the scene. Delays in communication can be very costly.[3] This is another instance that highlights the importance of fast and efficient transmission of only task relevant images.

So, instead of always sending the whole image, new approaches in task driven communication focus on sharing the most important parts of an image. By this way, only the essential image details get transmitted. This will speed up the process and make it more efficient. Without sending every pixel in the image, these methods focus on sharing information in the most meaningful and efficient way depending on the task. It helps to reduce the amount of data that needs to be transmitted.

Even with the recent progression, existing methods still transmit a significant amount of redundant information. This is a substantial issue under strict bandwidth limited environments. This research introduces a novel communication pipeline that identifies and transmits only task relevant patches of an image and then compresses them using a Vector-Quantized Generative Adversarial Network (VQ-GAN)(Esser et al.; 2021), also known as an Adversarial Vector-Quantized Autoencoder (Adversarial VQ-VAE). Then using only the transmitted patches, the full image is reconstructed at the receiver by a Masked Autoencoder (MAE)(He et al.; 2022). The objective is to reduce the amount of data to be transferred without compromising the performance of the intended task.

This research has both theoretical and practical significance in various scenarios. As examples mentioned above, most times detecting important object or area is critical than viewing the whole scene. Therefore, transmitting only most important parts with minimal delay and energy use is sometimes more beneficial than sending the complete images. The contributions of this study include a pipeline that uses saliency-guided patch selection combined with generative reconstruction with extreme compression. Even though generative models such as VQ-GAN, MAE, vision transformer (ViT), Generative Adversarial Networks (GAN) and Diffusion have been used before for image communication, a pipeline consists of saliency maps, VQ-GAN and MAE still novel in this domain. This study also included a comprehensive evaluation across different models and configurations. Then the presented method is compared with baselines to assess the performance.

## 1.1 Research Question

The central research question guiding this work is: **How can task-oriented selective patch transmission, combined with Adversarial VQ-VAE (VQ-GAN) and MAE reconstruction, reduce bandwidth usage for image communication while maintaining performance, when compared to MAE and full image VQ-GAN compression baselines, in terms of PSNR, SSIM, compression ratio, and task accuracy?**

[3]Mission-critical communication: `https://global-sys.com/news/mission-critical-communication-noise-issues`

## 1.2 Contributions

This research introduces a novel task-oriented communication pipeline that combines saliency map driven semantic patch selection, hybrid and saliency guided masking strategy, VQ-GAN compression and MAE generative reconstruction. The implemented saliency-based patch selection ensures that only the most semantically important regions of an image are transmitted. By this way it reduces the redundant data transmission under bandwidth constraints. The hybrid masking strategy integrates both random and saliency-guided masking which further strengthens reconstruction robustness by balancing semantic focus with generative flexibility. Together, these components form a communication method that has not previously been explored in this domain. So, this research offers a bandwidth-efficient approach capable of preserving essential structural and task-relevant information.

Another significant contribution is the comprehensive evaluation of system configurations to decide which models are the most effective depending on the application. Different model variants including one incorporating VQ-GAN compression and one without it are tested across different masking setups. This includes saliency masking and hybrid masking with different saliency weights. Each configuration is further evaluated across various masking ratios. This extensive analysis provides a detailed insights of of rate–distortion behavior, semantic preservation, and bandwidth efficiency. Ultimately this helps to identify optimal configurations for deployment in real-time, resource-limited environments such as space exploration, underwater robotics, search-and-rescue operations.

## 1.3 Report Structure Overview

This report will cover literature review section, methodology section, design specification section, implementation section, evaluation section and conclusion and future work section. Under Literature Review, existing work in semantic communication and generative compression will be discussed. The Methodology section explains the approach used and describes the design principles behind the pipeline introduced in this research. The Design Specification section gives details of each system component covering both architecture and functionality. To show their practical use, the Implementation section details how the methodology and design were applied. Also, an evaluation strategy which outlines the approach used to assess performance, including datasets, metrics, and baselines included in this section. In addition to that Conclusion and Future Work section summarizes the findings, discusses the implications, and highlights the next possible steps for research.

# 2 Related Work

In bandwidth limited environments, transmitting full resolution images is not effective or sometimes even possible due to various factors such as data rate, energy, and communication reliability. Because of these kinds of challenges conventional image communication tends to move towards task-oriented communication. In this field, transmit task relevant data is more important rather than focusing on entire image(Yang et al.; 2023). This shift has created a new field of research exploring generative and semantic aware compression techniques. The primary objective of this domain is to reduce data usage while trying to preserve the task performance as much as possible(Liu et al.; 2024).

## 2.1 Masked Autoencoders for Reconstruction

Recent research in this field of task-oriented communication has considered using generative models for efficient image communication under bandwidth limited sceneries. Masked Autoencoders (MAEs) are one of those models that have gained attention in the field of computer vision. MAEs work by masking a substantial part of image patches (e.g: 75%) randomly and then it reconstructs the missing parts or masked part using an asymmetric encoder-decoder. This shows the potential of MAE-based compression and reconstruction in bandwidth constrained environments(He et al.; 2022).

## 2.2 Deep Neural Networks, Vision Transformers and Generative Adversarial Networks Based Techniques

Advancements in semantic-aware communication have significantly improved image communication under limited bandwidth conditions. A key example of this is Deep Joint Source-Channel Coding (Deep-JSCC). It is a technique that improves data compression in transmission. Deep-JSCC does this by using end-to-end deep neural networks. Deep-JSCC does not separate source coding and channel coding. Instead, Deep-JSCC combines them into a single model. It maps input images to transmittable representations. At the receiver end, Deep-JSCC reconstructs the images using deep learning. This enables more efficient communication.

One study introduces two new Deep-JSCC variants, named Inverse-JSCC and Generative-JSCC, which leverage deep generative models such as StyleGAN (Style-Based Generator Architecture for Generative Adversarial Networks) to improve perceptual quality(Erdemir et al.; 2023). These methods optimize a weighted combination of MSE and LPIPS loss which can help to retain better semantic information. As a result, both of these two approaches outperform conventional Deep-JSCC, specifically in low SNR and low-bandwidth scenarios.

Building on this progress, the JSCCformer-f model proposes a feedback-enabled transmission scheme in which the transmitter iteratively refines image reconstruction using receiver feedback(Wu et al.; 2024). The architecture of this framework is based on the Vision Transformer (ViT)(Kolesnikov et al.; 2021). Further advancing semantic communication, another study uses a pre-trained GAN at the receiver to reconstruct images from semantic segmentation maps rather than full-resolution images and this greatly reduces the bandwidth while maintaining perceptual quality(Lokumarambage et al.; 2023). These works highlight the effectiveness of integrating generative models, semantic representations, and adaptive feedback for image semantic communication.

One method extracts latent representations using GAN inversion. It compresses data efficiently and protects sensitive information with a privacy filter and knowledge base(Han et al.; 2023). This approach reduces communication costs while keeping good image quality. Another method, Enhanced Low-Bit-Rate Generative Image Compression (EGIC), uses a GAN-based compressor with a semantic-segmentation-guided discriminator. It improves perceptual details at very low bit-per-pixel rates(Körber et al.; 2024). These methods still have some limitations such as producing artifacts (unwanted distortions or irregularities) under complex channel conditions. Also, fine details may not always be preserved, and large labeled datasets are often required for effective training. Even in the presence of such limitations, GAN-based compression techniques remain effective for semantic image communication.

## 2.3 Variational Autoencoder and Diffusion Based Techniques

Variational Autoencoders (VAEs) are another useful model for learned image compression, mainly because they can represent complex images in a compact way(van den Oord et al.; 2017). Based on this, adaptive VQ-VAE improves traditional VQ-based methods by adding extra information to better represent the latent space. This helps achieve better performance in vision tasks (Chen and Lin; 2023). ResNet VAEs have also been used for compression by introducing quantization aware layers. This modification enables efficient entropy coding and supports parallel execution on GPUs. This result in a powerful class of lossy image coders that outperform previous approaches in natural image compression(Duan et al.; 2023). Another VAE-based method, QARV (Quantization-Aware ResNet VAE), uses quantization-aware training and a layered design to support variable compression rates with fast decoding. QARV performs better than basic models by balancing quality and compression. Because of this reason, QARV can be useful for situations where bandwidth changes(Duan et al.; 2024). Diffusion models have recently gained traction because of their powerful generative capabilities(Rombach et al.; 2022). However, their application in communication has been limited due to bandwidth constraints. A novel diffusion-driven semantic communication framework addressed this gap by combining a VAE-based down-sampling mechanism and treating the channel as a diffusion forward process. Experimental results showed good performance in both semantic similarity and pixel-level fidelity compared to traditional Deep-JSCC approaches. Superior results were particularly evident at high compression ratios and low signal to noise settings(Guo et al.; 2025).

In another research, a two-stage extreme compression framework used VAE-based compression with pretrained stable diffusion. It introduced space alignment loss and a control module for content injection. This method achieves realistic reconstructions even at extremely low bitrates by outperforming some state of the art methods in visual quality(Li et al.; 2025). A related approach focuses on semantic preserving image transmission, where the sender and receiver share a semantic knowledge base. The transmitter uses classifiers and dictionary learning to extract features and the receiver employs a diffusion-based generator to reconstruct images with high semantic fidelity. This system is efficient in bandwidth-limited environments. It prioritizes semantics over pixel-level details(Liang et al.; 2024).

A stable diffusion-based goal-oriented semantic communication framework (SD-GSC) has been proposed to improve image transmission over noisy wireless channels(Li and Deng; 2025). It uses a semantic autoencoder to extract key image information that reduces transmit data size while preserving reconstruction quality. To mitigate channel noise, an SD-based denoiser is applied to the received semantic information under known channel conditions, and a parallel SD denoiser (PSD-GSC) is used when the channel is unknown. This approach improves peak signal to noise ratio (PSNR) and Fréchet inception distance (FID), compared to existing methods and demonstrates that combining semantic extraction with diffusion-based denoising can enhance bandwidth efficiency.

VAE-based methods often blur fine details and fail to match the sharpness achievable by adversarial models such as GANs. Their performance is also sensitive to codebook design and quantization settings. This create a trade-offs between compression ratio and reconstruction fidelity. Moreover, VQ-VAE latents are not robust to noisy channels. This means discrete code corruption can severely degrade performance. The patch-aware VQ-VAE(Cheng et al.; 2024), introduces computational overhead and makes deployment

more complex, whereas quantization-aware ResNet VAEs(Duan et al.; 2023) depend on GPU acceleration and may not generalize well to domain-specific data. Diffusion models are computationally expensive because of iterative denoising. This makes them impractical for real-time or low-power applications. Extreme compression frameworks(Li et al.; 2025) can cause semantic drift. They produce images that look realistic but may be semantically incorrect. Semantic-preserving methods(Liang et al.; 2024) need shared knowledge between sender and receiver. This is difficult in dynamic or diverse settings. SD-GSC and PSD-GSC(Li and Deng; 2025) handle channel noise better. However, they increase computational complexity and also need knowledge of channel statistics or extra learning modules which make real-time use difficult. Overall, these methods advance semantic communication and compression. Still, they face limits in computation, channel robustness, and the balance between realism and semantic accuracy.

## 2.4 Saliency Maps

Saliency maps show which parts of an image influence a model's decision the most. They are widely used in explainable deep learning. These maps highlight the important regions for predictions(Szczepankiewicz et al.; 2023). In bandwidth aware image transmission, saliency maps are very useful. These maps help the system identify the most informative or important parts of an image, which can lead to reduced data transmission.

Gradient weighted Class Activation Mapping (Grad-CAM) uses gradients of a target class flowing into the last convolutional layer to create coarse localization maps. Grad-CAM works with many convolutional neural network (CNN) models and does not require changes to the architecture. It can be applied to classification, captioning, and visual question answering. Studies show that it makes models more transparent, helps detect errors, and allows users to compare how strong or weak their models are even if predictions look similar(Selvaraju et al.; 2017). Another method for saliency maps is Grad-CAM++, which is built on Grad-CAM. This method provides better localization and works well when several objects of the same class appear in an image. Unlike Grad-CAM, which may fail in such cases, Grad-CAM++ uses a weighted mix of higher-order derivatives. This leads to more accurate saliency maps. Experiments show it consistently performs better than Grad-CAM in explanation and localization quality(Chattopadhay et al.; 2018). But Grad-CAM++ is computationally more expensive due to the calculation of higher order derivatives. This can slow down real time applications. In addition, even if it improves localization, it may still produce coarse heatmaps and sometimes struggle with very fine grained object boundaries. The method's effectiveness also depends on the quality of gradients, making it sensitive to noisy or saturated activations.

Score-CAM method removes the need for gradients. Instead, it assigns weights to activation maps using forward-passing class scores. This makes saliency maps more stable and fair. Score-CAM performs strongly in recognition and localization tasks and passes strict sanity checks(Wang et al.; 2020). Score-CAM also has some drawbacks. It requires multiple forward passes for each activation map, which increases computational cost and makes it slower compared to gradient-based methods and its performance can vary depending on the network depth and complexity.

With the growth of Vision Transformers, explainability methods have expanded to attention-based saliency. One approach uses Deep Taylor Decomposition to calculate local relevance scores. These scores are then passed through attention layers and skip connections. This keeps total relevance consistent across layers. Benchmarks show it

outperforms heuristic attention visualizations, making it effective for transformer-based models(Chefer et al.; 2021). However, in this saliency method attention maps can appear fragmented and may fail to capture complete object regions. As a result, visual explanations are less coherent than those produced by CNN-based models. Additionally, the complex self-attention structure can make it difficult to accurately localize important image areas, particularly in scenes with multiple interacting objects.

Another recent method, Opti-CAM, combines CAM-based and masking-based techniques. It creates saliency maps by maximizing the logit of a masked image for a chosen class. It also fixes flaws in common evaluation metrics. Opti-CAM achieves state of the art performance and beats older CAM-based methods. Importantly, results show that better localization does not always mean better interpretability. This gives useful insights for future saliency research(Zhang et al.; 2024). However, Opti-CAM optimization process is complex and requires careful regularization to prevent overfitting. Other than that, even if it improves localization, Opti-CAM may still produce saliency maps that are more spread out compared to other methods.

In summary, saliency methods range from gradient-based to optimization-driven approaches. They are essential for interpretability and efficient image communication. For bandwidth-efficient transmission, saliency maps can guide patch selection. This ensures the system prioritizes semantically meaningful regions for sending.This enable adaptive compression which keep important regions at higher quality. By this way it improves both transmission efficiency and the usefulness of the received image for downstream tasks.

## 2.5 Identified Gaps

Even with recent advancements, most methods in generative and semantic-aware image transmission field still treat compression, semantic reconstruction, and adaptive transmission in isolation instead of combining them into a unified approach. Techniques such as DeepJSCC, VAE-based compression, and MAE-driven reconstruction show strong results in improving rate–distortion performance or enhancing perceptual quality. But they often struggle to achieve both extreme bandwidth reduction and high semantic fidelity at the same time. In addition, although attention mechanisms and saliency maps are widely used for feature extraction in vision tasks, their ability to selectively transmit only the most semantically important regions under tight bandwidth constraints remains largely underexplored.

To address these gaps, this study introduces a pipeline combining saliency-guided patch selection, VQ-GAN-based compression, and MAE-based reconstruction. This model is compatible with both wired and wireless channels, though the primary focus is on the architectural part rather than the transmission medium. On the transmitter side, salient patches are encoded using the VQ-GAN encoder and converted into compact codebook indices. At the receiver side of this pipeline, VQ-GAN decoder and a pretrained MAE reconstruct the missing patches, ensuring both semantic integrity and spatial coherence. This setup greatly reduces the transmitted data and at the same time preservs high perceptual quality through learned generative priors.

## 2.6 Summary

Table 1 summarizes key studies discussed in the above sections. This table highlights recent progress and remaining challenges in semantic communication and related generative

frameworks.

Table 1: Summary of key papers.

| **Paper Reference** | **Method** | **Main Contribution** | **Key Limitations** | **What's Missing** |
|---|---|---|---|---|
| (Erdemir et al.; 2023) | Generative-JSCC / Inverse-JSCC | Combines GANs with Deep-JSCC for better perceptual/semantic quality in low bandwidth/SNR. | High training complexity; potential reconstruction artifacts. | No saliency/adaptive patch selection; no MAE integration. |
| (Wu et al.; 2024) | JSCCformer-f | Uses feedback-aware Vision Transformer for refined image reconstruction. | Needs extra feedback channel; computationally heavy. | No semantic-driven transmission; no generative compression. |
| (Lokumarambage et al.; 2023) | Semantic Map-based GAN | Transmits semantic maps instead of full images, saving bandwidth. | Dependent on segmentation and GAN quality. | No adaptive compression or autoencoder integration; fine details may be lost. |
| (Chen and Lin; 2023) | Adaptive VQ-VAE | Improves VQ compression via adaptive latent representations. | Sensitive to codebook/quantization; not robust to noise. | No saliency or generative reconstruction; only compression-focused. |
| (Duan et al.; 2024) | QARV | Quantization-aware VAE enables variable-rate compression and fast decoding. | Sensitive to channel noise; requires GPU. | No semantic-aware or selective transmission. |
| (He et al.; 2022) | Masked Autoencoder | Reconstructs masked patches using self-supervised asymmetric encoder-decoder. | Needs large-scale pretraining; may lose texture. | No compression or saliency-guided patch selection. |

# 3 Methodology

## 3.1 Dataset Preparation and Image Preprocessing

The dataset used in this research consists of 1300 images belonging to a single semantic class extracted from ImageNetDeng et al. (2009). Using a single class makes sure consistent structural patterns across images which makes it easier to isolate the impact of saliency-based masking and compression. All images were center-cropped and resized to 512×512 and then those images were normalized using ImageNet mean and standard deviation. Images were converted into 3-channel PyTorch tensors and processed in batches of size 1, preserving full spatial detail for visualization and analysis. This standardized preprocessing was done to ensure compatibility with both the Grad-CAM++ module, VQ-GAN and MAE.

## 3.2 Semantic Region Identification and Patch-Level Importance Estimation

Grad-CAM++ is applied on a pretrained ResNet50He et al. (2016) backbone to identify the most important parts of each image. Forward and backward hooks are added to the final convolutional layer to collect the feature maps and their gradients. After a forward pass identifies the predicted class, a targeted backward pass retrieves the class-specific gradients. These gradients are then used in the Grad-CAM++ equations. This combines first and higher-order gradient terms to produce a class-focused heatmap. Then this heatmap is normalized, resized to 512×512, and used for patch-level relevance scoring.

To match the MAE patch layout, the heatmap is split into a 14×14 grid (196 patches). Each of these patches are given an importance score by adding the normalized Grad-CAM++ values inside it. Patches with important semantic regions get higher scores, while background patches get lower ones. This transfers pixel-level saliency into the MAE's patch format, allowing masking to be guided by semantic relevance instead of being random or uniform. The resulting 196-dimensional importance vector is then used for saliency-based and hybrid masking in the later stages of the method.In Figure 1 you can see a example for Grad-CAM heatmap and how patch rank work.


(a) Original Image (b) GradCAM++ Heat map (c) Patch rank

Figure 1: Original image,Saliency map and Patch rank of image

## 3.3 Masking Strategy

Instead of depending on traditional random masking, this framework implemented task-driven masking using Grad-CAM++. In addition to the fully saliency-driven approach, the study proposes an alternative masking strategy that combines both random and saliency-based masking. Hereafter, it will be referred to as hybrid masking. These approaches enable more controlled and interpretable experimentation. Because of this approach, the model focuses on semantically meaningful regions, and the use of random masking ensures that less important areas are also preserved to some extent. Integrating these complementary masking techniques enhances both the interpretability and effectiveness of the representation learning process.

### 3.3.1 Saliency Masking

In saliency masking, patches are selectively removed based on their importance as indicated by a model's interpretability map. Here specifically Grad-CAM++ was used

to identify the importance. The Grad-CAM++ heatmap identifies the spatial regions that are most influential to the model's predictions and this information is then used for masking. First, the Grad-CAM++ heatmap is resized to match the dimensions of the input image and it is partitioned into patches similar to the MAE's patch size. Then each patch is assigned an importance score by adding the Grad-CAM values within its patch area. This importance score is then used to determine the patch rank which refers to the ordering of patches based on their importance scores. Here, higher-ranked patches being more semantically important and lower-ranked ones being less important. Once it is ordered by the patch rank, a predefined mask ratio is applied. After that the lowest importance patches are selected for masking by preserving semantically meaningful regions.

### 3.3.2 Hybrid Masking

Hybrid masking combines saliency-based masking with random masking. After obtaining the saliency scores from the preceding saliency masking process, a fixed portion of the total mask ratio is dedicated to removing the least important patches, ensuring that regions with low contribution to the model's prediction are masked first. By this way it helps to preserve the intended semantic focus of the masking process. The remaining portion of the mask ratio is then applied using random selection from the unmasked patches by introducing stochasticity into the masking pattern. Overall, the hybrid approach preserves the meaningful structure provided by saliency information and at the same time it introduces sufficient diversity in the masking patterns. This makes it well suited for experiments that require both control and flexibility in the masking design. In Figure 2 you can see the difference between saliency and hybrid masking.

(a) Original Image (b) Saliency masking (c) Hybrid masking

Figure 2: Saliency masking vs Hybrid masking

## 3.4 Patch Stitching and VQ-GAN based Compression

After masking, all visible (unmasked) patches are extracted from the original image and reassembled into a compact "stitched" layout. The visible patches are arranged into a tight square grid, with the grid side length computed as the ceiling of the square root of the number of visible patches, forming a dense image containing only retained information. This stitched image is saved and resized for the compression stage. After the compression module produces a decoded output, the reconstructed patches are split

back into 14×1 tiles and placed into their original spatial locations on a 512×512 canvas, restoring the global layout while leaving masked regions empty. This two-step procedure simplifies compression and ensures correct spatial alignment for the downstream MAE reconstruction.

The stitched visible-patch image is then processed using a VQ-GAN, which serves as the core compression mechanism in the semantic communication pipeline. On the transmitter side, the encoder compresses the stitched content into a latent representation, which is quantized to the nearest entries in a learned codebook, producing discrete code indices that form a highly bandwidth-efficient representation. On the receiver side of the transmitter, the VQ-GAN decoder reconstructs the visible regions from these tokens (indices) by expanding the quantized latent features back into a full-resolution image. The adversarial training of VQ-GAN encourages sharper and more realistic outputs. Since, transmitter–receiver structure transmits only discrete code indices instead of full pixel data, it significantly reduces information overhead while preserving essential semantic and structural information.

## 3.5 Image Reconstruction Using Masked Autoencoder

The reconstruction stage evaluates the model's ability to recover missing visual information under different masking strategies. Two reconstruction pathways are considered. One where the visible patches are first compressed using a VQ-GAN before being processed by the MAE, and another where the MAE operates directly on pixel-space patches.

### 3.5.1 MAE Reconstruction With VQ-GAN Preprocessing

In this pathway, the visible patches are first compressed using a VQ-GAN to produce discrete latent indices representing the visible content. This compressed representation is then decoded and fed into the MAE. In the next step MAE reconstructs the masked regions using the decoded image. VQ-GAN reduces feature dimensionality while preserving semantic structure. It enables a more bandwidth-efficient input to the MAE. Combining VQ-GAN preprocessing with MAE's transformer-based reconstruction improves both semantic fidelity and perceptual sharpness in the reconstructed images.

### 3.5.2 MAE Reconstruction Without VQ-GAN

Here, instead of sending through VQGAN, after saliency or hybrid mask is generated, the masked image is directly fed into a MAE reconstruction. The MAE reconstructs the masked regions from raw pixel-space patches. The hybrid and saliency masking guides the MAE to focus on informative areas while still predicting content in less critical regions. This pathway highlights the benefits of hybrid masking and salience masking compared to MAE traditional random masking process.

## 3.6 Evaluation Methodology

The evaluation framework systematically compares the proposed reconstruction approaches against established baselines across a range of masking settings. To test the reconstruction quality, official MAE implementation from Facebook (Meta) research teamHe et al. (2022) serves as the baseline. And it's evaluated under masking ratios from 25% to 90% in 5% increments to capture how reconstruction performance changes with increasing

masking severity. Also, to compare the compression ratios, two baselines were used. One of them is VQ-GAN(Esser et al.; 2021), but without stitching the image, it sends the whole image though it. And the other compression baseline is JPEG XL standard[4] , serving as a traditional compression benchmark.

The proposed models are assessed in two categories. First, saliency-guided masking is tested both with and without VQ-GAN preprocessing. Second, hybrid masking combines saliency-based masking at ratios of 40%, 50%, and 60% with random masking for the remaining portion. Each hybrid configuration is also evaluated with and without VQ-GAN integration to assess the effects of compression on reconstruction quality.

All models were evaluated using multiple quantitative metrics. Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Fréchet Inception Distance (FID), and Mean Absolute Error (MAE), which capture aspects such as semantic consistency, pixel-level fidelity, perceptual quality, and overall reconstruction error. Apart from those metrics VQ-GAN–based models are assessed for compression ratio. This enables analyze the effect of latent compression on information preservation. Also, beyond reconstruction and compression, all models are evaluated for task accuracy to ensure functional effectiveness.

# 4 Design Specification

## 4.1 Overview

This pipeline implements an image reconstruction framework that integrates saliency-guided masking, VQ-GAN compression, and Masked Autoencoder (MAE) reconstruction. The objective is to reduce the amount of data to be transferred without compromising the performance of the intended task and at the same time enabling evaluation of compression efficiency when VQ-GAN is used. The system automates preprocessing, mask generation, compression, reconstruction, and visualization, following a standardized folder hierarchy and naming convention for reproducibility. In Figure 3, you can see the higher-level overall architecture of the pipeline.

## 4.2 Masking Strategies

Masking in the system is controlled by a single Grad-CAM fraction parameter. If Grad-CAM fraction set to 1 result in fully saliency-based masking, 0 produces purely random masking. And if you set it to any value between 0 and 1, it generates a hybrid mask that mixes saliency-guided and random patch removal. This unified approach allows seamless selection between the three masking approaches

## 4.3 VQ-GAN Feature Compression

In this setup, visible patches produced after masking are stitched into a grid and passed through a VQ-GAN–style model for compression. The VQModel first uses an encoder composed of a pixel-unshuffle layer, convolutional down-sampling, and several residual blocks to produce a low-dimensional latent feature map. This latent map is then quantized using a VectorQuantize module with a codebook of 8192 entries, replacing each latent

[4]JPEG XL White Paper: `https://ds.jpeg.org/whitepapers/jpeg-xl-whitepaper.pdf`

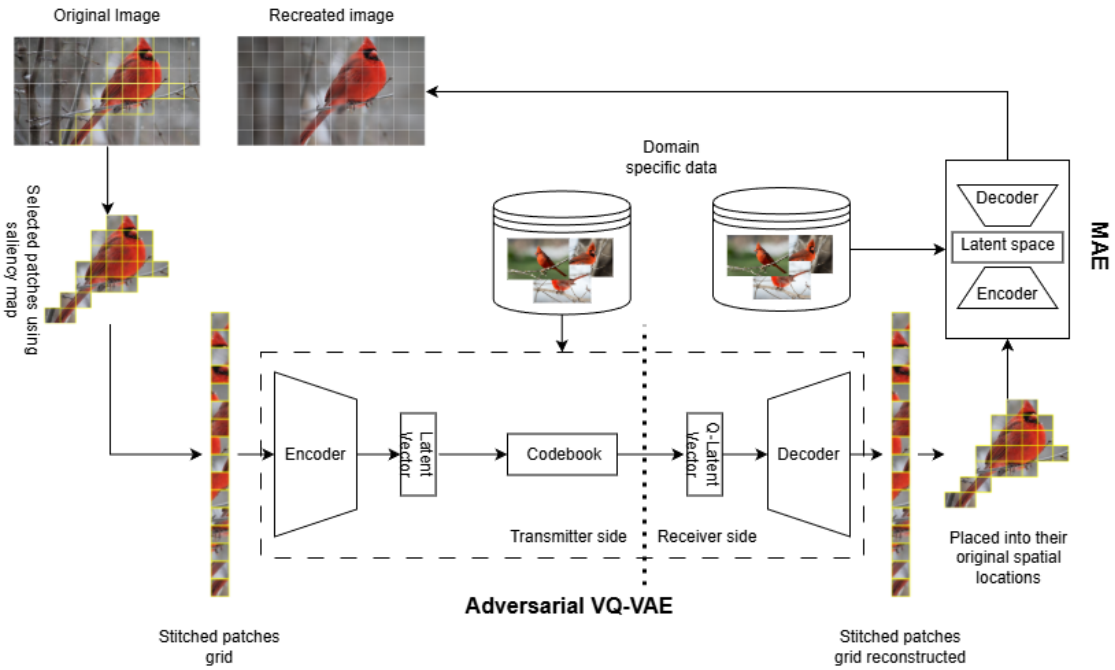



Figure 3: Higher level overall architecture of the pipeline

vector with the nearest codebook embedding and producing discrete indices. These compact indices simulate a semantic communication pipeline, since they are far smaller than the original image data. The decoder then takes the quantized latent features and begins reconstructing the image. For this it first processes them through convolutional up-sampling layers to gradually increase the spatial resolution. Next, bottleneck residual blocks refine the feature representation. Finally, a pixel-shuffle operation restores the image to its original resolution.

## 4.4 Masked Autoencoder (MAE) Reconstruction

In this study, the reconstruction was done using the MAE-ViT Large (ViT L) model(He et al.; 2022) with adversarial (GAN) loss. Here, the GAN component enhances the standard reconstruction objective to produce sharper and more perceptually realistic details. The Vision Transformer–based Masked Autoencoder reconstructs the full image from the masked input by predicting missing patches based on the visible ones, generating both the reconstructed regions and a combined image that blends visible and predicted patches. All outputs including originals, masked images, reconstruction-only results, and combined reconstructions are saved in a structured folder layout organized by masking ratio and masking strategy. And visualization is done for each case by showing the original, masked, reconstructed-only, and reconstructed-visible images. By this way it enables clear qualitative and quantitative comparison across different masking configurations.

## 4.5 Data Management and Outputs

All input and output images are resized to 512×512 pixels to ensure consistency. The folders are structured in a well-organized way with separate directories for the original images, saliency maps, masked images, reconstructed images, evaluation results, and stitched images. Compression statistics are recorded in a CSV file if VQ-GAN is ap-

plied in the pipeline. It contains summarizing data of image sizes and transmitting data sizes for each image. This structured management helps to maintain the reproducibility, modularity, and traceability of results across all masking strategies.

### 4.6 Architectural Requirements

The system was designed focusing on key operational and architectural requirements. The system is modular, and this enables each component such as masking, VQ-GAN, and MAE to be configured independently. The ViT-L backbone ensures scalability because it can support high-resolution images and also it has a large model capacity. Consistency of the study is maintained through identical training settings across all variants. This enables controlled comparisons with different configurations and models. The architecture is also support extendibility, if you want to add additional compression modules, masking strategies, loss functions, or other components as needed in future.

### 4.7 Algorithm Description

The pipeline follows these steps for each image:

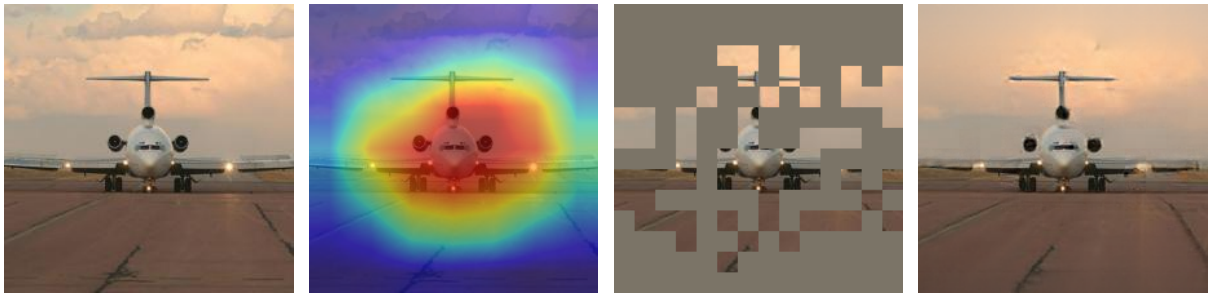
1. Input image is normalized and resized for consistency.
2. Grad-CAM++ is computed for the predicted class to generate a saliency map.
3. Masking is applied according to the selected strategy. This includes saliency, hybrid, or random masking.
4. Visible unmasked patches are stitched together in a grid.
5. The stitched image is processed through VQ-GAN for compression. The generated codebook indices of VQ-GAN are sent through the transmitter.
6. On receiver side, codebook indices decode though the VQ-VAE decoder. Then the stitched image is passed to the MAE for final reconstruction.
7. Reconstructed images are saved using a standardized naming convention.
8. Compression statistics are recorded if VQ-GAN is used.
9. Compute evaluation metrics.

## 5 Implementation

To implement this research Python (3.10) was used as the primary programming language and PyTorch was used for model development, training, and inference. The implementation utilized a framework integrating Grad-CAM++, masked autoencoders and VQ-VAE. The implementation consists of eight main model configurations which can divide into two categories as with and without VQ-GAN. Each of these two categories was tested under different settings, including saliency masking and hybrid masking with saliency-based ratios of 60%, 50%, and 40%.

For all configurations, multiple masking ratios ranging from 25% to 90% in increments of 5% were applied to evaluate reconstruction performance. For models with VQ-GAN, compression ratio calculations were also done to assess bandwidth efficiency of the pipeline. Each reconstructed image was stored with its corresponding original image, the applied mask, and the Grad-CAM heatmap. In Figure 3 you can see the implementation outputs for a single image. Together, these outputs enabled smooth and comprehensive qualitative and quantitative evaluation.

The implementation used different types of tools and frameworks to support data management, processing, and visualization. Torchvision was used for image transformations and ViT integration. NumPy and scikit-image were used for numerical operations and image analysis. Because of the image preprocessing capabilities OpenCV was used and Matplotlib and Seaborn were used for plotting. CUDA-enabled GPUs provided the computational acceleration required for large-model training and inference. The development setup included both PyCharm and Jupyter Notebook covering both traditional script development as well as interactive experimentation. The primary outputs of this final stage consist of reconstructed images, applied masks, Grad-CAM heatmaps, compression ratios, task accuracy and calculated performance metrics.



(a) Original Image (b) Saliency map (c) Masking map (d) Generated image

Figure 4: Implementation outputs

# 6 Evaluation

The main purpose of this evaluation process is to systematically analyze the performance of the implemented reconstruction approaches across different masking conditions and also how it performs against baselines. The official Masked Autoencoder (MAE) implementation from Facebook (Meta) research teamHe et al. (2022) serves as the baseline for reconstruction quality and it's assessed across masking ratios from 25% to 90% in increments of 5%. This range captures both mild and severe information loss, allowing us to evaluate model robustness under increasing reconstruction difficulty. For compression two baselines were used, one being sending the full image through VQ-GAN(Esser et al.; 2021) without stitching and the other one is JPEG-XL[5], an international standard (ISO/IEC 18181) developed by the JPEG committee.

[5]JPEG XL White Paper: `https://ds.jpeg.org/whitepapers/jpeg-xl-whitepaper.pdf`

## 6.1 Evaluation without VQ-GAN

Hybrid masking consistently outperforms the baseline across all masking percentages. It shows significant improvements in pixel-wise metrics, perceptual quality metric such FID, and task accuracy. Saliency masking also surpasses the baseline at every masking level, achieving strong perceptual performance and higher task accuracy, but its MAE values indicate some reduction in pixel-level precision.

As the masking percentage increases especially beyond 70% pixel-wise metrics decrease slightly for all the models. However, for task accuracy and FID, both Hybrid and Saliency masking manage perform over the baseline. The performance gap also becomes increasingly noticeable at higher masking levels. Saliency masking continues to achieve the best perceptual scores (FID) in most cases showing its ability to preserve semantically meaningful content even under heavy masking. In Figure 5 and Figure 9 (a), you can see the performance metric comparisons and task accuracy respectively.

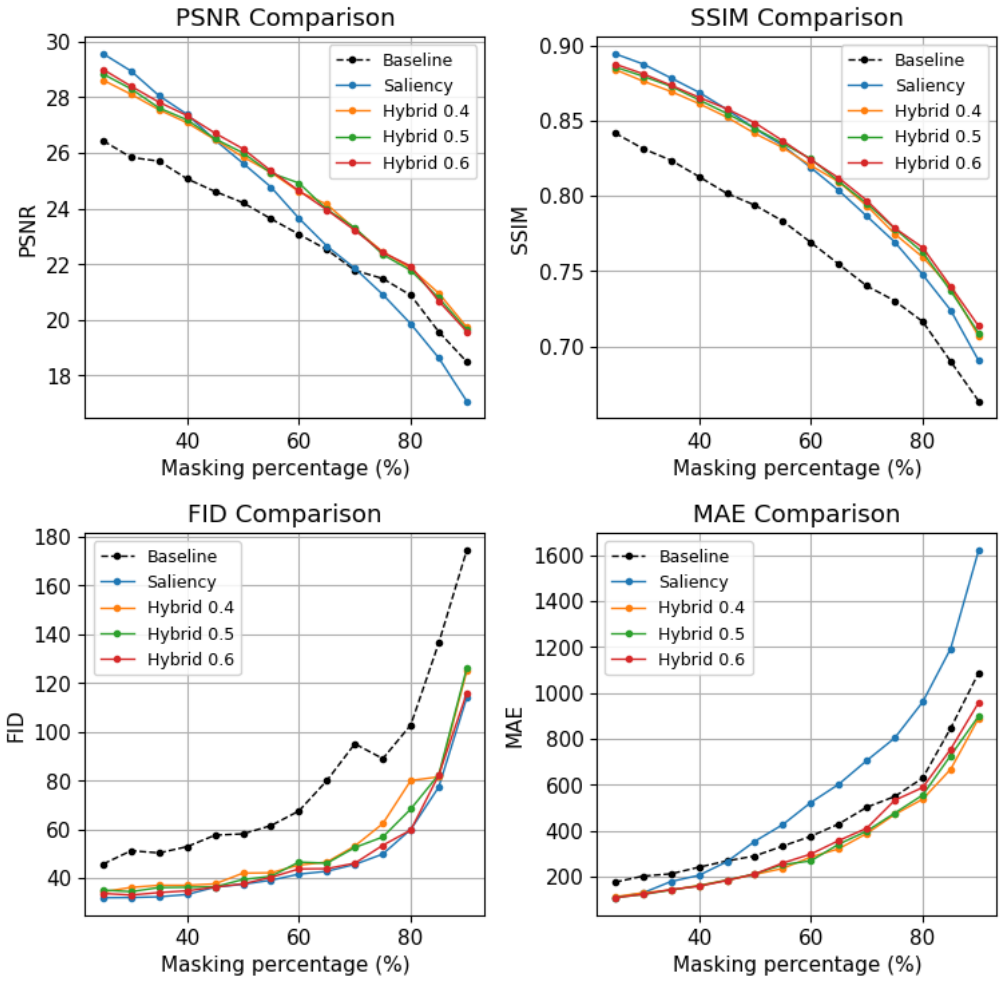



Figure 5: PSNR, SSIM, FID, MAE across different masking percentages

### 6.1.1 Visual Comparison

Other than quantitative evaluation, visual inspection of the reconstructed images also shows that the implemented masking strategies consistently perform better over the baseline. Both hybrid and saliency masking approaches produce reconstructions that

are perceptually closer to the ground truth, preserving structural details and semantic content more effectively than the baseline which use random masking.

This visual advantage is clearly visible at higher masking ratios, where baseline reconstructions tend to lose significant detail and appear blurred or distorted. But in saliency and hybrid masking, even if the unimportant parts are blurred or distorted up to some extent the still manages preserve the important regions. In contrast, hybrid masking maintains a balance between salient and non-salient regions, while saliency-guided masking effectively prioritizes critical image regions. These observations strengthen the conclusion that the implemented approaches achieve better reconstruction quality across all masking scenarios and it also complements with the quantitative metrics. In figure 6 you can clearly see this.

Figure 6: Combined visualization of Baseline (Top), Saliency masking (Middle), and Hybrid masking (Bottom) for different masking percentages (50%, 60%, 70%, 80%, 90%).

## 6.2 Evaluation with VQ-GAN compression

The second set of experiments evaluates the proposed masking strategies with VQ-GAN preprocessing. Figure 7 presents PSNR, SSIM, FID, MAE, and Figure 9 (b) you can see the task accuracy for hybrid and saliency-guided masking across different masking ratios, along with the baseline results. In this scenario hybrid masking closely matches or exceeds the baseline in pixel-wise metrics (PSNR and MAE) at all masking levels. Saliency masking performs similarly to the baseline in SSIM and FID at low masking levels but falls behind in PSNR and MAE as masking percentage increases. Other than visual quality matching visual quality with baseline, VQ-GAN integration provides up to 20× to 160× compression This compression ratio improves as masking percentage increases. In Figure 8 you can clearly see that it perform substantially better than the two different baselines.

For task accuracy, both Hybrid and Saliency masking outperform the baseline at every masking level. At low masking ratios, their performance is very similar and at higher

masking ratios, both methods show a clear advantage over baseline. Saliency masking managed to achieve the highest task accuracy in most cases consistently.

For perceptual quality (FID), both Hybrid and Saliency masking also outperform the baseline. Saliency masking achieves the best FID scores overall, showing its strength in preserving high-level semantic content even under strong compression.

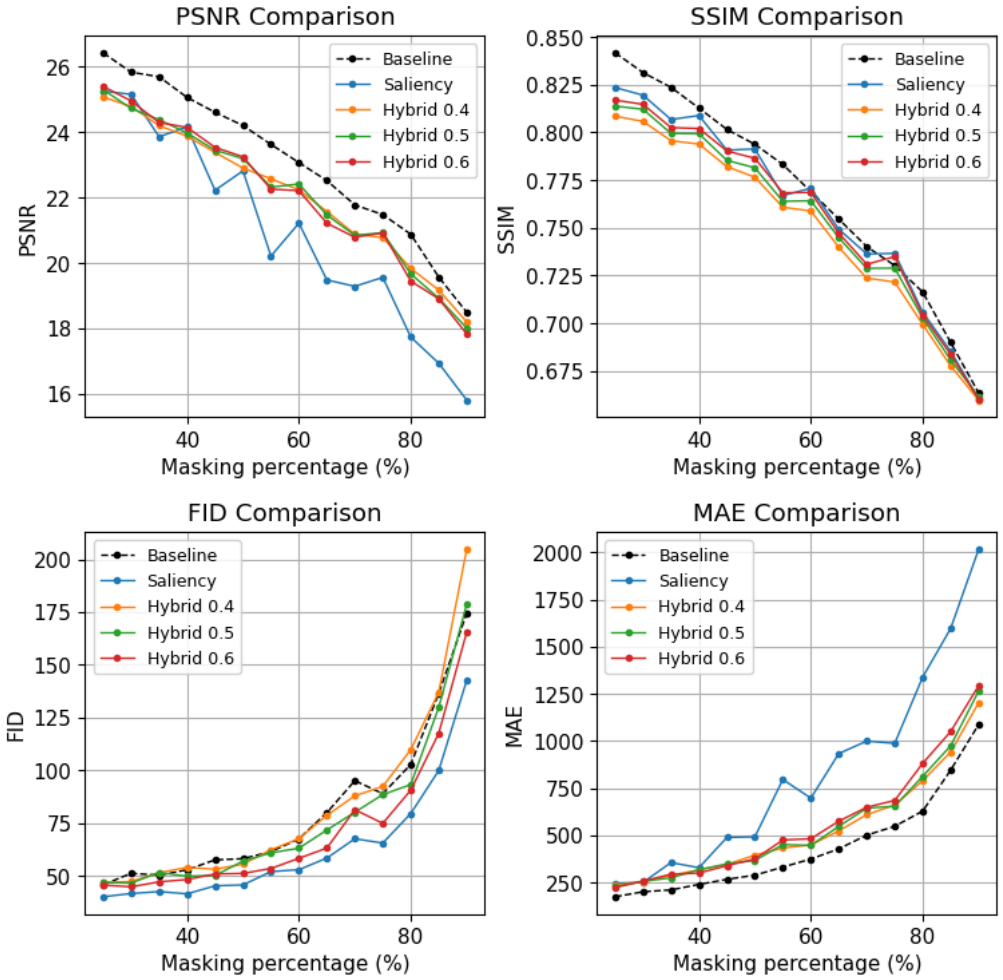



Figure 7: PSNR, SSIM, FID, MAE across different masking percentages

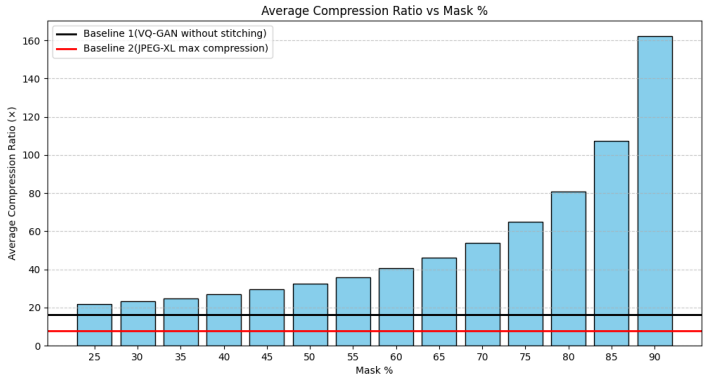
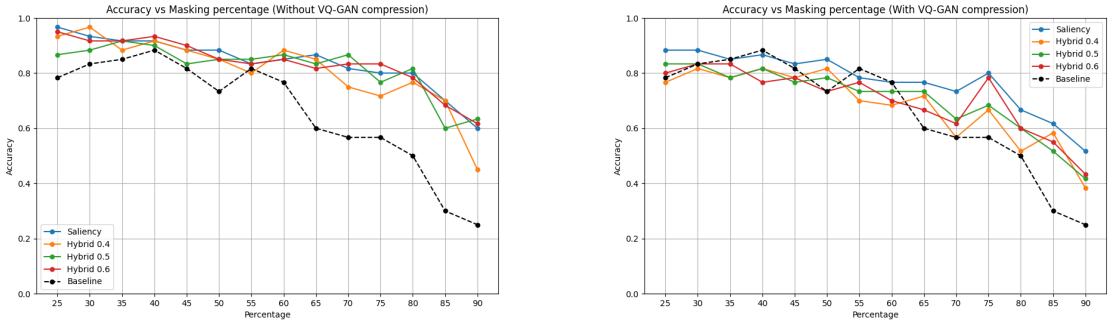



Figure 8: Compression ratio with baseline




(a) Without compression

(b) With compression

Figure 9: Task accuracy with baseline

## 6.3 Discussion

The evaluation results highlight the strengths and trade-offs of the implemented masking strategies. This applies both with and without VQ-GAN preprocessing and also the analysis covers a wide range of masking ratios. Overall, both hybrid and saliency-guided masking consistently outperform the baseline in preserving reconstruction quality and task performance. Without VQ-GAN, saliency-guided masking achieves the highest PSNR and SSIM values specially around low to moderate masking ratios (25–50%). This shows its effectiveness in prioritizing semantically important regions. At higher masking ratios over 60%, hybrid masking provides better reconstructions by combining saliency-driven and random masking. By this way it's possible to retains both critical and non-salient regions. Pixel-wise metrics such as MAE decrease slightly as masking increases. However, task accuracy and perceptual quality (FID) remain consistently higher for the introduced strategies and at the same time visual inspection confirms these findings. Reconstructions from hybrid and saliency masking are perceptually closer to the ground truth which means it preserves structural details and semantic content better. In Figure 10 illustrates the percentage gains achieved through both hybrid and saliency masking across different masking ratios.

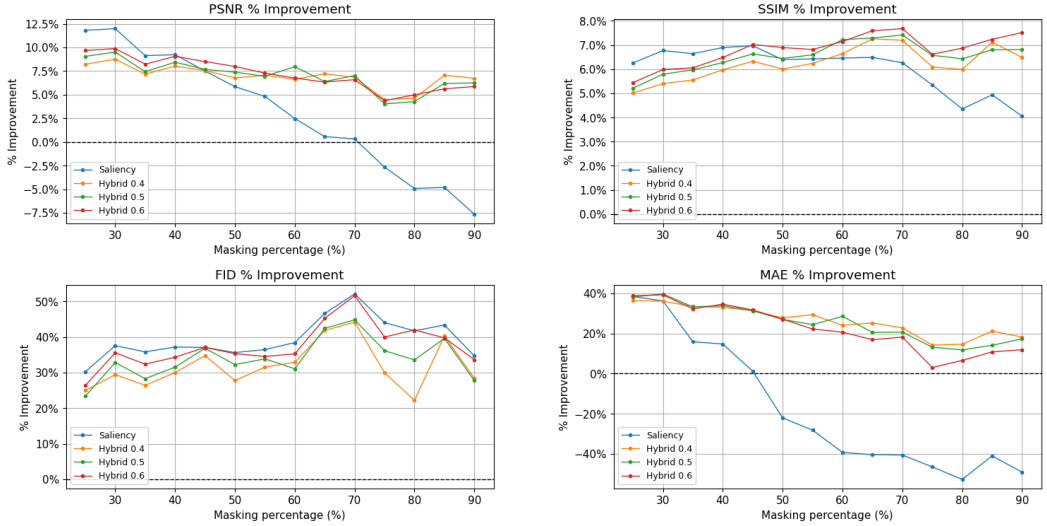



Figure 10: Performance improvements achieved through masking strategy

Task accuracy can sometimes fluctuate due to misclassifications, as a example airliner might be misidentified as a warplane. Despite this, overall accuracy tends to be higher for the introduced strategies because they preserve critical semantic information. The choice of masking strategy can depend on the size of the important region in the image. If the salient region is small, saliency masking with higher masking ratios is preferable because in this way you can get a very high compression ratio. If the important region is large, hybrid masking with different hybrid ratios works better, as it preserves both salient and non-salient content. Pure saliency masking may fail when Grad-CAM does not fully detect the salient region properly. In cases such as those, hybrid masking is more reliable because it balances saliency-driven and random masking, ensuring essential structures are retained.

Integrating VQ-GAN preprocessing introduces discrete latent compression and this can lead to minor reductions in pixel-level metrics (PSNR, SSIM, MAE) compared to the uncompressed setting. Even with this, both hybrid and saliency-guided masking retain strong task accuracy and perceptual quality. Saliency masking achieves the best FID scores in most cases and hybrid masking closely follows or surpasses the baseline in pixel-level metrics across all masking percentages. Saliency masking matches the baseline at low masking levels but underperforms at higher masking levels. The main advantage of VQ-GAN is its substantial compression ratio, ranging from 20× to 160× depending on the masking percentage. This makes it practical for bandwidth-constrained scenarios and also this method can be a very useful way to store images too. Because you can store codebook indices and later use them to convert them into images. The results show a clear trade-off, with only minor losses in reconstruction fidelity. These losses are offset by significant compression gains, while the masking strategies still preserve key structural and semantic information.

# 7 Conclusion and Future Work

In conclusion, hybrid and saliency-guided masking introduced in the research consistently perform better than the baseline across all masking ratios. Saliency-guided masking works best at moderate masking levels keeping the perceptual quality high. Also, when the saliency region is relatively small, saliency masking tends to work better. Hybrid masking performs better at high masking levels and when the saliency region is substantial. When these strategies are used with VQ-GAN, they achieve high compression without losing much visual quality or important information. This shows they are effective for efficient and semantically accurate image reconstruction and compression. This pipeline is not only useful for transmission but also it can utilized for other use cases such as storing image data.

For future work, several improvements can be made. More experiments can be done to test different hybrid masking ratios to find the best combination. This research was focused on a single class and It was because of the time and hardware limitations, but in future work with more time and better hardware should include multiple classes for better generalization. To improve robustness and performance of these models, testing on more datasets and using additional metrics can provide extra insights. Apart from those combining VQ-GAN with other compression or quantization methods may improve the balance between reconstruction quality and storage efficiency.

# References


Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847.

Chefer, H., Gur, S. and Wolf, L. (2021). Transformer interpretability beyond attention visualization, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791.

Chen, K.-H. and Lin, Y.-C. (2023). Adaptive vqvae: a learning-based image compression framework with vector quantization, *Proceedings of the 2023 11th International Conference on Computer and Communications Management*, ICCCM '23, Association for Computing Machinery, New York, NY, USA, p. 76–82.
**URL:** *https://doi.org/10.1145/3617733.3617746*

Cheng, Q., Qu, S. and Lee, J. (2024). Patch-aware vector quantized codebook learning for unsupervised visual defect detection, *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 586–592.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Duan, Z., Lu, M., Ma, J., Huang, Y., Ma, Z. and Zhu, F. (2024). Qarv: Quantization-aware resnet vae for lossy image compression, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(1): 436–450.

Duan, Z., Lu, M., Ma, Z. and Zhu, F. (2023). Lossy image compression with quantized hierarchical vaes, *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 198–207.

Erdemir, E., Tung, T.-Y., Dragotti, P. L. and Gündüz, D. (2023). Generative joint source-channel coding for semantic image transmission, *IEEE Journal on Selected Areas in Communications* **41**(8): 2645–2657.

Esser, P., Rombach, R. and Ommer, B. (2021). Taming transformers for high-resolution image synthesis, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878.

Guo, L., Chen, W., Sun, Y., Ai, B., Pappas, N. and Quek, T. Q. S. (2025). Diffusion-driven semantic communication for generative models with bandwidth constraints, *IEEE Transactions on Wireless Communications* **24**(8): 6490–6503.

Han, T., Tang, J., Yang, Q., Duan, Y., Zhang, Z. and Shi, Z. (2023). Generative model based highly efficient semantic communication approach for image transmission, *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R. (2022). Masked autoencoders are scalable vision learners, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T. and Zhai, X. (2021). An image is worth 16x16 words: Transformers for image recognition at scale, *International Conference on Learning Representations (ICLR)*.

Körber, N., Kromer, E., Siebert, A., Hauke, S., Mueller-Gritschneder, D. and Schuller, B. (2024). Egic: Enhanced low-bit-rate generative image compression guided bynbsp;semantic segmentation, *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XXXV*, Springer-Verlag, Berlin, Heidelberg, p. 202–220.
**URL:** *https://doi.org/10.1007/978-3-031-72761-0*$_1$2

Li, N. and Deng, Y. (2025). Goal-oriented semantic communication for wireless image transmission via stable diffusion, *ICC 2025 - IEEE International Conference on Communications*, pp. 3748–3753.

Li, Z., Zhou, Y., Wei, H., Ge, C. and Jiang, J. (2025). Toward extreme image compression with latent feature guidance and diffusion prior, *IEEE Transactions on Circuits and Systems for Video Technology* **35**(1): 888–899.

Liang, C., Li, D., Lin, Z. and Cao, H. (2024). Selection-based image generation for semantic communication systems, *IEEE Communications Letters* **28**(1): 34–38.

Liu, Y., Wang, X., Ning, Z., Zhou, M., Guo, L. and Jedari, B. (2024). A survey on semantic communications: Technologies, solutions, applications and challenges, *Digital Communications and Networks* **10**(3): 528–545.
**URL:** *https://www.sciencedirect.com/science/article/pii/S2352864823000925*

Lokumarambage, M. U., Gowrisetty, V. S. S., Rezaei, H., Sivalingam, T., Rajatheva, N. and Fernando, A. (2023). Wireless end-to-end image transmission system using semantic communications, *IEEE Access* **11**: 37149–37163.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models , *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 10674–10685.
**URL:** *https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042*

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626.

Szczepankiewicz, K., Popowicz, A., Charkiewicz, K., Nałecz-Charkiewicz, K., Szczepankiewicz, M., Lasota, S., Zawistowski, P. and Radlak, K. (2023). Ground truth based comparison of saliency maps algorithms, *Scientific Reports* **13**(1): 16887.

van den Oord, A., Vinyals, O. and Kavukcuoglu, K. (2017). Neural discrete representation learning, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, p. 6309–6318.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P. and Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 111–119.

Wu, H., Shao, Y., Ozfatura, E., Mikolajczyk, K. and Gündüz, D. (2024). Transformer-aided wireless image transmission with channel feedback, *IEEE Transactions on Wireless Communications* **23**(9): 11904–11919.

Yang, W., Du, H., Liew, Z. Q., Lim, W. Y. B., Xiong, Z., Niyato, D., Chi, X., Shen, X. and Miao, C. (2023). Semantic communications for future internet: Fundamentals, applications, and challenges, *IEEE Communications Surveys Tutorials* **25**(1): 213–250.

Zhang, H., Torres, F., Sicre, R., Avrithis, Y. and Ayache, S. (2024). Opti-cam: Optimizing saliency maps for interpretability, *Comput. Vis. Image Underst.* **248**(C).
**URL:** *https://doi.org/10.1016/j.cviu.2024.104101*