Efficient and Robust Semantic Image Communication via Stable Cascade

Bilal Khalid [1] **Pedro Freire** [1] **Sergei K. Turitsyn** [1] **Jaroslaw E. Prilepsky** [1]

Abstract

Diffusion Model (DM) based Semantic Image Communication (SIC) systems face significant challenges, such as slow inference speed and generation randomness, that limit their reliability and practicality. To overcome these issues, we propose a novel SIC framework inspired by Stable Cascade, where extremely compact latent image embeddings are used as conditioning to the diffusion process. Our approach drastically reduces the data transmission overhead, compressing the transmitted embedding to just 0.29% of the original image size. It outperforms three benchmark approaches — the diffusion SIC model conditioned on segmentation maps (GESCO), the recent Stable Diffusion (SD)-based SIC framework (Img2Img-SC), and the conventional JPEG2000 + LDPC coding — by achieving superior reconstruction quality under noisy channel conditions, as validated across multiple metrics. Notably, it also delivers significant computational efficiency, enabling over $3\times$ faster reconstruction for 512×512 images and more than $16\times$ faster for 1024×1024 images as compared to the approach adopted in Img2Img-SC.

1. Introduction

Semantic communication (SemCom) is a transformative approach that focuses on effectively conveying the meaning of information rather than transmitting raw bit data (Strinati & Barbarossa, 2021). The goal is to communicate the essential information the receiver needs to complete its task successfully. This also makes it bandwidth efficient as significantly less data has to be transmitted across the communication channel (Luo et al., 2022; Qin et al., 2021).

[1]Aston Institute of Photonic Technologies, Aston University, Birmingham, UK. Correspondence to: Bilal Khalid <r.khalid4@aston.ac.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Workshop on Machine Learning for Wireless Communication and Networks, Vancouver, Canada, 2025.

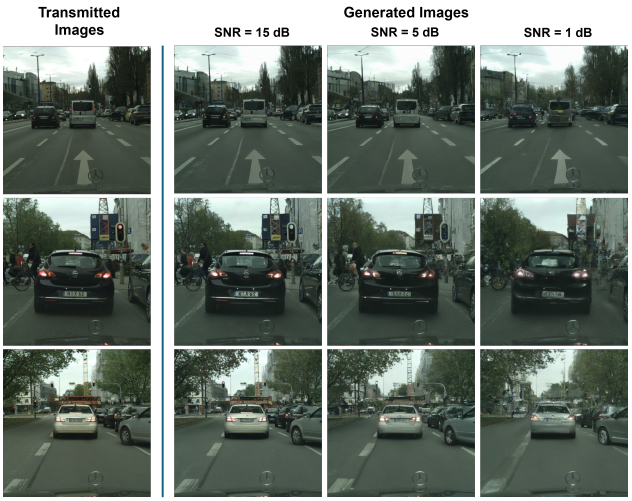

Figure 1. 1024×1024 Image reconstructions using our model under different channel SNR conditions. Even at an SNR of 1 dB, images are faithfully reconstructed and perceptually very similar to the transmitted images.

The advancement of Deep Learning (DL) and generative AI has enabled the emergence of SemCom as a viable alternative to traditional communication. DL and generative AI models are used for extracting the relevant semantic information at the transmitter end as well as for deciphering the meaning behind this information at the receiver end. Deep learning-based Joint Source-Channel Coding (DeepJSCC) (Bourtsoulatze et al., 2019) was one of the first approaches to incorporate DL in wireless system design. Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Diffusion Models (DMs) and Flow-based Generative Models (FGMs) are the major generative AI techniques now commonly used in SemCom systems (Xia et al., 2025). Out of these, DMs have shown great potential at Semantic Image Communication (SIC) tasks because of their exceptional ability to synthesize high-quality images (Dhariwal & Nichol, 2021). However, one drawback of DMs is that they are inherently slower at inference because of their iterative nature. The introduction of Latent Diffusion Models (LDMs) (Rombach et al., 2022) has alleviated this problem by performing the diffusion process in a compressed latent space instead of the original pixel space, enabling fast and high-resolution image generation via diffusion.

Several DM-based SIC systems have been implemented in

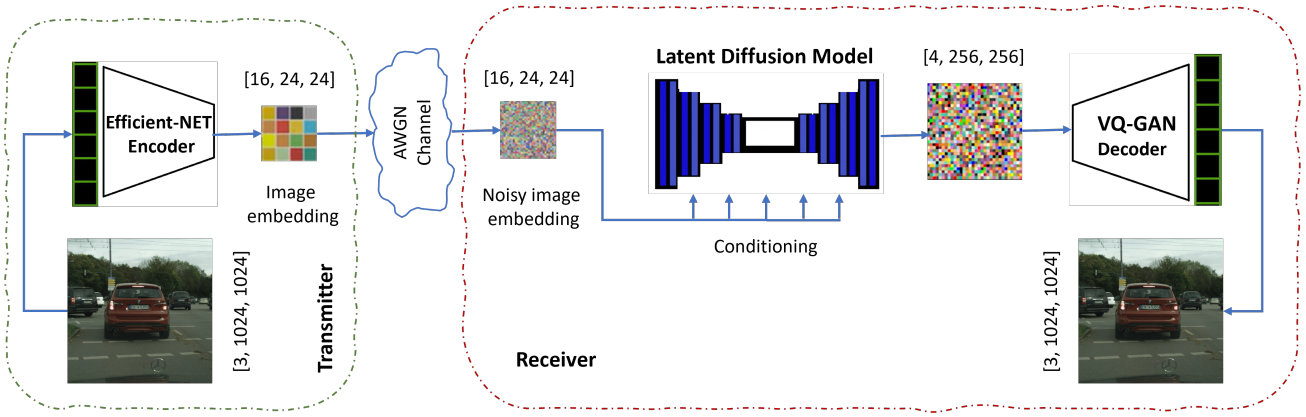

Figure 2. Our system model. At the transmitter side, a compact image embedding Z of size [16, 24, 24] is extracted from an image X of size [3, 1024, 1024]. Z is transmitted across the physical channel. The receiver uses the noisy embedding $\hat{Z}$ as conditioning for the LDM. Finally, the VQGAN decoder is used to project the image back into pixel space.

recent years. In (Grassucci et al., 2023), segmentation maps are used to guide the diffusion process. In (Yilmaz et al., 2024), the primary image structure is transmitted using the DeepJSCC technique, whereas fine details are generated using the diffusion model. (Jiang et al., 2024) also use a diffusion model to refine the reconstruction obtained after image decoding. However, inference using these approaches is time-consuming. Recently, LDMs have been used for SIC to speed up the inference process. In (Nam et al., 2024; Cicchetti et al., 2024), text conditioning is used to guide the generative process of Stable Diffusion's text-to-image model (Rombach et al., 2022). In (Cicchetti et al., 2024), the generation process starts from a noisy version of image embedding instead of pure noise. Although efficient in terms of bandwidth, these models struggle to faithfully reconstruct the intended image and suffer from generation randomness. (Chen & Yang, 2024) denoise a noisy image embedding using an LDM, and the clean embedding is then used to reconstruct the image using a semantic decoder. Instead of predicting the noise in the image, (Yang et al., 2025) use a diffusion model to predict the source image in a few denoising steps directly. Both of these models reduce inference time but operate at a lower compression factor as compared to our proposed method.

In this paper, we propose a novel SIC model inspired by Stable Cascade (SC) (Pernias et al., 2023), a multistage text-to-image LDM that operates in a much smaller latent space than Stable Diffusion (SD). Our approach achieves the trifecta of high compression efficiency, fast inference, and perceptually aligned image reconstruction, which is missing in existing DM-based SIC systems. In our method, a highly compressed image embedding is extracted using a semantic encoder and transmitted across the physical channel. The noisy embedding is then given as a conditioning signal to the LDM of SC that projects it into a higher dimensional latent space where the semantic decoder operates. Results indicate that we outperform benchmark models and as shown in Figure 1, generate consistent reconstructions even under extremely poor channel Signal-to-Noise Ratio (SNR) conditions.

2. Proposed Framework

In this section, the proposed system model is explained. The model is built upon the architecture of SC that has three stages, i.e., stages A, B and C. As discussed below, our model is based on stage A and a finetuned stage B that is trained to work with noisy conditioning.

Stage A is a Vector Quantized Generative Adversarial Network (VQGAN) (Esser et al., 2021) with parameters Θ that compresses the image space by a factor of 4. The relationship between an input image $X \in \mathbb{R}^{3\times1024\times1024}$ and the output of VQGAN encoder X_{VG} is given as:

$$X_{\text{VG}} = f_{\Theta}(X). \tag{1}$$

If f_{Θ}^{-1} represents the VQGAN decoder, the image can be reconstructed from the compressed latent space using

$$f_{\Theta}^{-1}(X_{\text{VG}}) \approx X. \tag{2}$$

Stage B is a LDM that learns to generate the latent space X_{VG} given a highly compressed latent representation Z of X. This compact embedding is obtained via the EfficientNet-V2 encoder (Tan & Le, 2019). During the forward process in training, the latents X_{VG} are noised according to the following relation:

$$X_{\text{VG},t} = \sqrt{\bar{\alpha}_t} \cdot X_{\text{VG},t} + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon. \tag{3}$$

Here, $\bar{\alpha}_t$ specifies the noise schedule whereas ϵ is the noise sampled from a standard normal distribution $N(0,1)$. At

any time-step t, with noised latents $X_{\text{VG},t}$ and noisy conditional embedding $\hat{Z}$, the LDM is trained to predict the noise $\bar{\epsilon}(X_{\text{VG},t}, t, \hat{Z})$. The training objective is to minimize the loss function L, defined as the Mean-Squared Error (MSE) between the predicted and actual noise:

$$L = \mathbb{E}_{(X_{\text{VG},t},t,\hat{Z},\epsilon)} \left[\|\epsilon - \bar{\epsilon}(X_{\text{VG},t}, t, \hat{Z})\|_2^2 \right]. \quad (4)$$

Text embedding is also used as conditioning for Stage B in the original SC paper (Pernias et al., 2023). However, as noted in the paper itself, it has no significant impact on the reconstruction quality of stage B as the conditioning provided by the image embedding is much stronger. Thus, we do not consider text conditioning in our model. The fine-tuning of Stage B conditioned on $\hat{Z}$ makes it robust to channel impairments. Moreover, we do not consider stage C either as it is primarily responsible for text-to-image generation.

3. System Model

Figure 2 shows the three phases of our system model i.e. semantic information extraction at the transmitter, noisy channel transmission, and image reconstruction at the receiver.

3.1. Semantic Feature Extraction

As in (Pernias et al., 2023), we utilize the pretrained EfficientNet-V2 image encoder to extract a compact image embedding. An input RGB image $X \in \mathbb{R}^{N\times H\times W}$ is encoded into a compressed embedding $Z = \mathcal{E}(X)$[1]. Despite its compact size, this embedding contains well-generalized feature representations that provide stronger guidance to the diffusion model as compared to text embeddings. As a result, the reconstructed image is very similar to the original one, with differences in fine details only. Although image generation based solely on text conditioning is highly efficient in terms of bandwidth, it may result in reconstructions that are semantically quite different from the source image (Nam et al., 2024). Furthermore, as compared to segmentation map-based conditioning (Grassucci et al., 2023), image embeddings offer better reconstruction fidelity. Although segmentation maps retain spatial structure, they often lose crucial details such as texture, color, and fine-grained features. Additionally, because they provide only class-level information, the same segmentation map can yield multiple plausible reconstructions, introducing variability. To achieve reliable, predictable, and efficient SIC, we propose using rich image embedding as a more effective conditioning signal, ensuring reduced generation randomness and high-fidelity reconstruction of transmitted images.

[1]The dimensionalities of X; N is the number of channels, i.e. 3 for RGB, and H and W stand for the height and width pixel resolution respectively.

3.2. Communication Channel

To maintain conformity with most previous works (Grassucci et al., 2023; Yilmaz et al., 2024; Chen & Yang, 2024; Yang et al., 2025), we consider the widely adopted additive white Gaussian noise (AWGN) channel in our simulations. The extracted image embedding Z is transmitted across the AWGN channel where the noise ϵ is sampled from a zero-mean normal distribution $N(0, \sigma^2)$ with variance σ^2. If P denotes the received signal power, the channel conditions are characterized by the Signal-to-Noise Ratio (SNR):

$$\text{SNR} = 10 \log \left(\frac{P}{\sigma^2} \right) \text{(dB)}. \quad (5)$$

Depending upon the SNR level, noise is added to Z and the distorted embedding $\hat{Z}$ is obtained as

$$\hat{Z} = Z + \epsilon. \quad (6)$$

3.3. Image Reconstruction

The noisy image embedding $\hat{Z}$ is used as a conditioning signal to the diffusion model at the receiver side. It should be noted that in (Cicchetti et al., 2024), a text-conditioned diffusion model starts sampling from a noisy version of the image embedding, whereas, in our model, a significantly more compressed image embedding is used purely as a conditioning signal. After the conditional denoising process is complete, the output of the LDM is the predicted latent space $\hat{X}_{\text{VG}}$ where the VQGAN decoder operates. Finally, in accordance with Equation (2), the generated image $\hat{X}$ is obtained using $f_\Theta^{-1}(\hat{X}_{\text{VG}}) = \hat{X}$.

4. Experimental Evaluation

4.1. Model Training

We train our model using the Cityscapes dataset (Cordts et al., 2016). The dataset contains 3000 training, 500 validation, and 1500 test images. All images are resized to 1024×1024 resolution. We finetune the pre-trained stage B checkpoint for 15000 steps using a batch size of 4, learning rate of 1×10^{-4}, and AdamW optimizer. To improve generalization and robustness, the SNR is randomly selected to be between $1-20$ dB. At each training step, image embeddings are extracted and transmitted across the AWGN channel. The model is trained to use the noisy embeddings as conditioning to reconstruct images with the objective of minimizing the MSE loss in accordance with Equation (4). In addition to the Cityscapes dataset, we also evaluate our model's performance on the DIV2K dataset (Agustsson & Timofte, 2017), which is composed of highly diverse images. We do not finetune our model again for this dataset to investigate how well it generalizes on completely different and unseen data. All the training and simulations have been performed

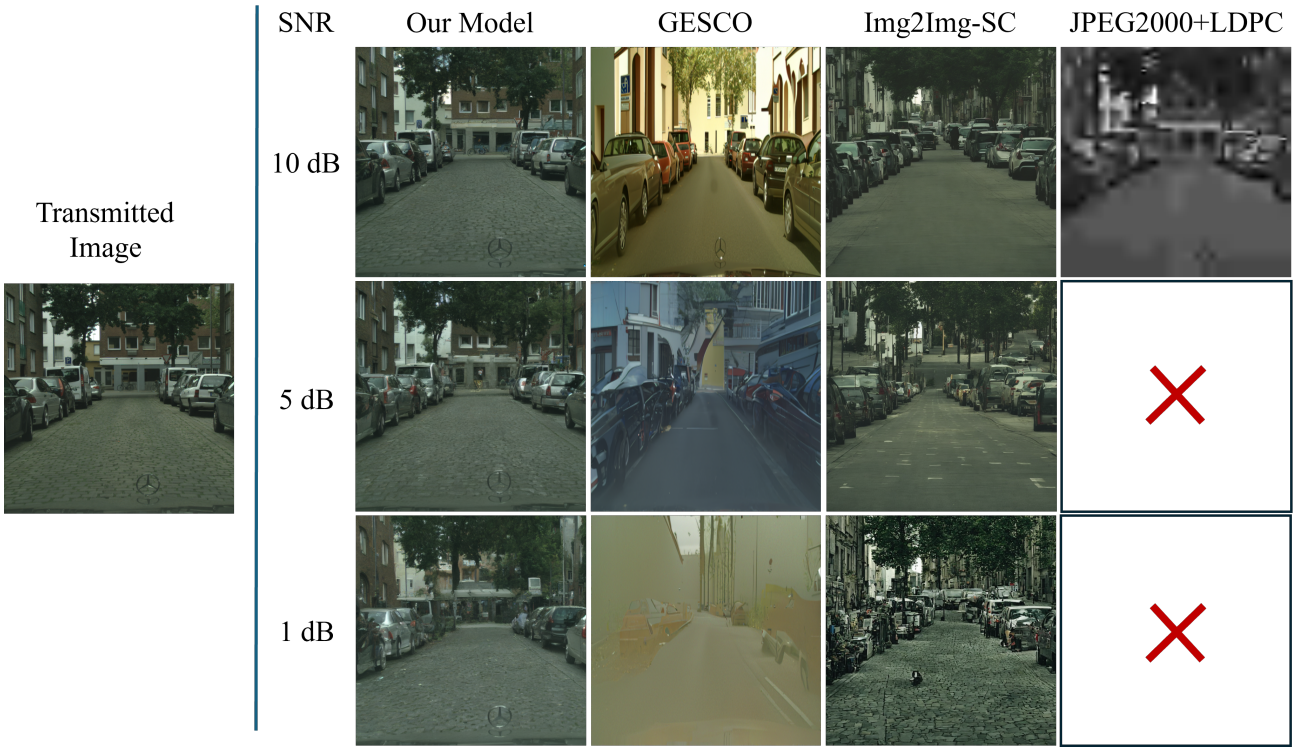

Figure 3. Image reconstructions using our model, GESCO, Img2Img-SC and JPEG2000+LDPC in low SNR conditions. It can be observed that our model generates the most semantically similar images with the least generation randomness. The red crosses indicate that the JPEG2000+LDPC system was unable to recover the image at the corresponding SNR.

using a single NVIDIA RTX A6000 (48-GB) GPU. All code scripts and fine-tuned model weights will be accessible at: https://github.com/abilalk02/SC-SIC.

4.2. Simulation Settings

We compare the performance of our model with (i) the diffusion SIC model conditioned on segmentation maps (GESCO) (Grassucci et al., 2023), (ii) the Stable Diffusion-based SIC model that transmits text and image embeddings (Img2Img-SC) (Cicchetti et al., 2024), and (iii) the conventional JPEG2000 compression with Low-Density Parity-Check (LDPC) error correction approach. For evaluation, we generate 100 samples using each model with channel SNR values of $1, 5, 10, 15$ and 20 dB respectively. All samples are of resolution 512×512, except for GESCO, where the resolution is 256×512[2]. For sampling with GESCO and Img2Img-SC, 1000 and 30 denoising steps are used, respectively, as in the original papers. For JPEG2000+LDPC, Quadrature Amplitude Modulation (QAM) is used and the LDPC coding rate is set to $1/2$ following the method described in (Bourtsoulatze et al., 2019).

[2]It was not possible to generate 512×512 images using GESCO without altering the model architecture.

Performance Metrics: To evaluate the perceptual and semantic similarity between the original and generated images, we calculate the Learned Perceptual Image Patch Similarity (LPIPS) score (Zhang et al., 2018), Fréchet Inception Distance (FID) score (Seitzer, 2020) and Structural Similarity Index Measure (SSIM) (Wang et al., 2004). We also measure the Peak Signal-to-Noise Ratio (PSNR) to evaluate pixel-level similarity between images. Lower values of LPIPS and FID indicate better performance, whereas higher values of SSIM and PSNR indicate better performance.

4.3. Results

4.3.1. Image Reconstruction Quality

We first evaluate the reconstruction quality of our model against existing approaches, including GESCO, Img2Img-SC, and the JPEG2000+LDPC framework. Figure 3 shows the reconstruction of a transmitted image at the receiver end using the four models under low SNR conditions. Our model consistently achieves the most accurate reconstructions of the original image. Even at extremely low SNR levels of 5 dB and 1 dB, it preserves object clarity and recognizability. In contrast, the reconstruction quality of GESCO deteriorates rapidly as SNR decreases, leading to significant visual degradation. Moreover, the output pro-

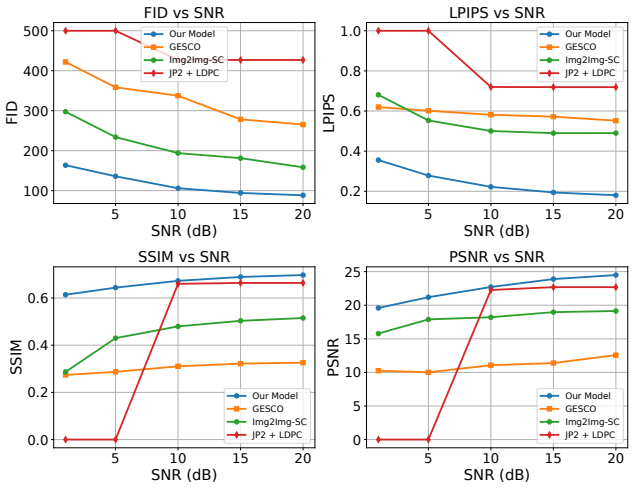

Figure 4. Performance comparison between our model, GESCO, Img2Img-SC and JP2+LDPC at different SNRs.

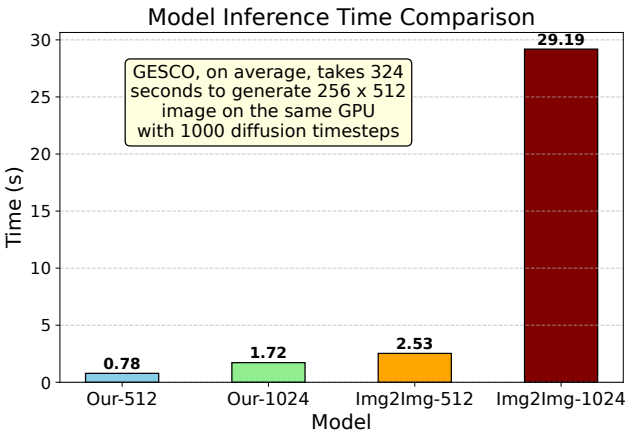

Figure 5. Inference time comparison of our model with GESCO and Img2Img-SC.

duced by Img2Img-SC is loosely tied to the original image because text conditioning introduces significant randomness in the generation process. Finally, the conventional JPEG2000+LDPC produces heavily distorted output, and error correction completely fails at low SNR, as was observed earlier (Bourtsoulatze et al., 2019; Jiang et al., 2024). For cases where it fails to reconstruct the images, we set the PSNR and SSIM scores to 0, whereas LPIPS and FID scores are assigned an arbitrary maximum value of 1 and 500 respectively.

The comparison across performance metrics on the Cityscapes test data, shown in Figure 4, also reveals that our model achieves the best results. In terms of FID and LPIPS, on average, our model improves on the results of the next-best approach from Img2Img-SC by 43% and 55%, respectively. Similarly, in terms of SSIM and PSNR, our model gives the best results, maintaining good performance even at low SNR. For SNR greater than 10 dB, JPEG2000+LDPC achieves comparable PSNR and SSIM to our model even though its reconstructions are heavily distorted, have artifacts, and lack details. This can be attributed to the fact that JPEG2000 compression preserves low-frequency components and structural integrity. PSNR and SSIM primarily assess pixel-level accuracy and structural similarity, respectively. In contrast, LPIPS and FID are more sensitive to perceptually significant distortions, capturing the loss of fine details, reduced realism, and unnatural textures. Thus, high PSNR and SSIM scores can misleadingly overestimate the performance of JPEG2000+LDPC, failing to reflect the perceptual degradation. Moreover, as discussed, the conventional method fails to reconstruct the images at low SNR. Overall, our model improves SSIM by 56% and PSNR by 23% as compared to Img2Img-SC. The results of our model improve further when generating 1024×1024 images.

4.3.2. Inference Speed and Bandwidth Efficiency

In terms of computational complexity, we evaluate both inference latency and the dimensionality of the transmitted data. As shown in Figure 5, the model from (Grassucci et al., 2023), which does not utilize an LDM, exhibits significantly higher latency, requiring 5 minutes and 24 seconds for image reconstruction with $T = 1000$ denoising steps. Our method achieves substantially lower inference time, just 0.78 seconds for 512×512 images, making it $3\times$ faster than Img2Img-SC. For 1024×1024 images, our model accelerates reconstruction further, achieving speeds over $16\times$ faster than that of Img2Img-SC.

Table 1. Dimensionality Comparison

Transmitted Data	Dimensionality	Compression Ratio	% of original
Original Image	$[3, 512, 512]$	–	–
Our Model	$[16, 12, 12]$	341	0.29%
Img2Img-SC	$[4, 64, 64]$	48	2.08%
DIFFSC	$[8, 32, 32]$	96	1.04%
CASC	$[8, 32, 32]$	96	1.04%

Moreover, in terms of dimensionality, Table 1 shows that we achieve a higher Compression Ratio (CR) as compared to other state-of-the-art DM-based SIC systems. Following the definition in (Jiang et al., 2024), where CR is defined as the ratio of the input image's dimensionality to that of its encoded representation, our approach compresses an RGB image of size $[3, 512, 512]$ into a compact embedding of $[16, 12, 12]$, achieving an exceptional CR of 341 – meaning that the transmitted data is only 0.29% of the original image size. This highlights the remarkable bandwidth efficiency of our method.

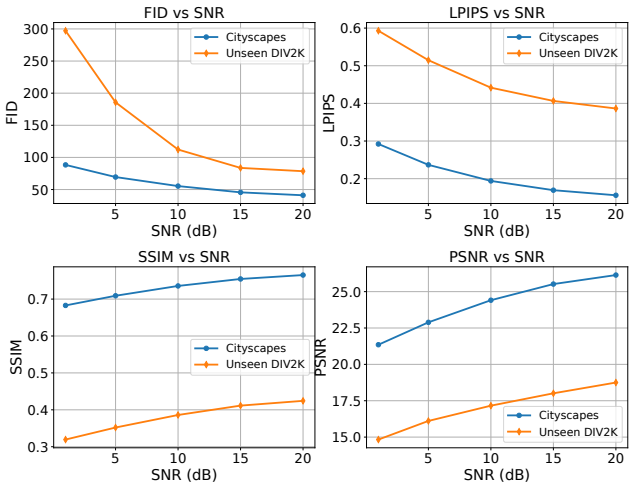

Figure 6. Performance of our model on unseen DIV2K data.

4.3.3. Reconstruction Predictability

We assess reconstruction predictability across varying SNR conditions using the LPIPS metric. For each case, we simulate image transmission 25 times with fixed parameters, computing the mean (μ) and standard deviation (σ) of LPIPS scores across all pairwise comparisons of generated images. As shown in Table 2, our model achieves the lowest average LPIPS score and standard deviation, $(\mu \pm \sigma) = (0.173 \pm 0.003)$ at SNR$=$ 20dB, indicating minimal generation randomness. Thus, the proposed model is able to reconstruct images reliably and consistently.

Table 2. Predictability Comparison

SNR (dB)	LPIPS Score $(\mu \pm \sigma)$			
	Our-1024	**Our-512**	**GESCO**	**Img2Img-SC**
20	0.173 ± 0.003	0.205 ± 0.005	0.401 ± 0.014	0.520 ± 0.011
15	0.195 ± 0.003	0.223 ± 0.006	0.433 ± 0.012	0.541 ± 0.017
10	0.229 ± 0.003	0.264 ± 0.008	0.424 ± 0.017	0.522 ± 0.012
5	0.287 ± 0.004	0.314 ± 0.009	0.575 ± 0.021	0.554 ± 0.019
1	0.351 ± 0.006	0.371 ± 0.013	0.613 ± 0.017	0.578 ± 0.019

4.3.4. Generalization on Unseen Data

We also analyze the performance of our model, trained on the Cityscapes dataset, on entirely unseen data. For this purpose, we use the DIV2K dataset that contains diverse images, including landscapes, people, architecture, and animals. Figure 6 indicates that there is a significant degradation in performance on this new data across all four metrics. For example, at an SNR of 15 dB, LPIPS increases from 0.17 to 0.4, whereas FID increases from 45 to 83, indicating a substantial loss in perceptual quality. However, a closer look at the generated images, Figure 7, reveals that much of this degradation may be attributed to the sharp differences in the colors between the original and generated images. The model does fairly well to reconstruct these unseen images and mitigate the effects of noise, but since it is finetuned on the Cityscapes dataset, the generated images have a color tone that resembles very closely to that of the images in the said dataset. These results suggest that fine-tuning a Stable Cascade model on a single large and highly diverse dataset may enable it to handle a wide range of image types with strong performance.

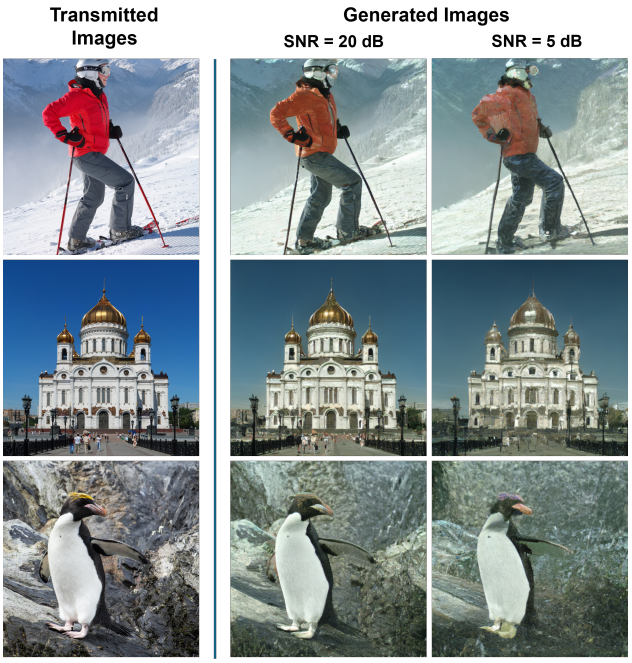

Figure 7. Image reconstructions on unseen DIV2K data. It can be seen that the model does well to mitigate the noise and reconstruct semantically similar images considering that it was not finetuned for this dataset.

4.3.5. Ablation Studies

Finally, we perform ablation tests to compare the performance of our fine-tuned model against the original Stable Cascade model in the semantic image communication scenario. Figure 8a shows that without fine-tuning, the original model's performance degrades sharply with decreasing SNR. In particular, at SNR less than 10 dB, the images generated using the original model are heavily corrupted by noise. This is also evident from Figure 9, which shows that the original model is unable to mitigate the channel effects. These findings validate our training approach and demonstrate the substantial performance gains achieved by fine-tuning the model to work with noisy image embedding as a conditioning signal.

We also analyze the impact of increasing the size of the extracted image embedding on the generation quality for

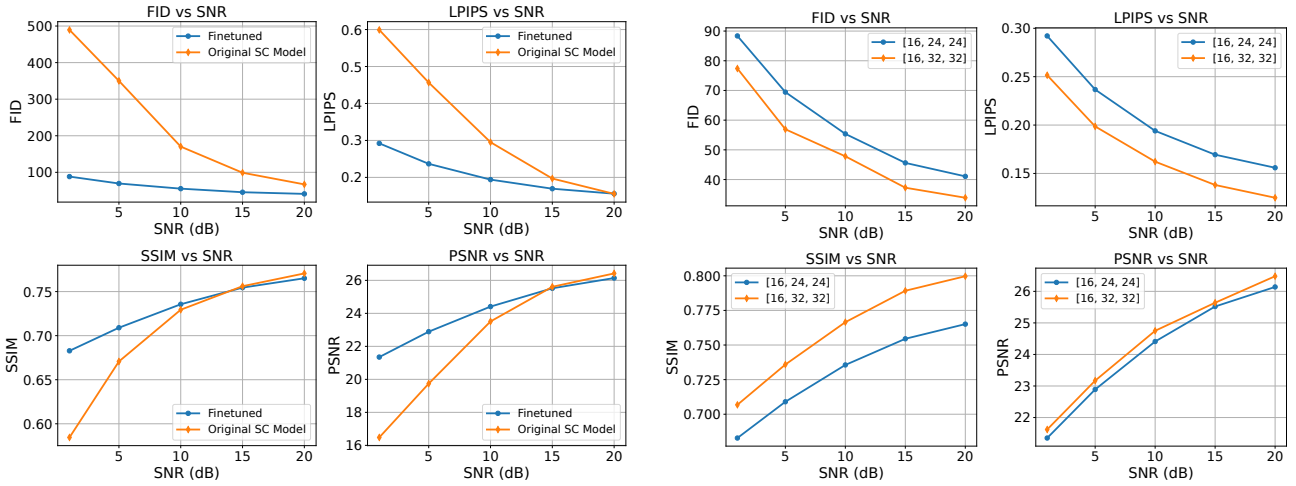

Figure 8. Results of ablation experiments highlighting (a) the performance gains obtained via fine-tuning and (b) the impact of increasing the embedding size from [16, 24, 24] to [16, 32, 32] on performance metrics.

Figure 9. Images reconstructed by the original Stable Cascade model. It can be seen that without proper fine-tuning, the original model fails to deal with the effects of channel noise.

1024×1024 images. It can be seen from Figure 8b that there is a noticeable improvement in performance across all four performance metrics when the embedding size is increased from $[16, 24, 24]$ to $[16, 32, 32]$. Quantitatively, on average, LPIPS, FID, and SSIM scores improve by greater than 10%. However, these improvements come at a cost to the compression ratio that drops from 341 to 192. Hence, there is an understandable tradeoff between performance and bandwidth efficiency

5. Conclusion

In this paper, we introduce a novel DM-based SIC framework that leverages the Stable Cascade architecture to achieve an exceptional balance of speed, compression, and fidelity under noisy channel conditions. Our method transmits a highly compact image embedding, only 0.29% of the original size, and reconstructs 512×512 images in just 0.78 seconds – $3\times$ faster than Img2Img-SC. Extensive evaluations using perceptual quality metrics, including LPIPS, SSIM, and FID, demonstrate the noise robustness of our approach and its superiority over existing benchmarks. Additionally, our framework minimizes generation randomness by achieving an LPIPS score variance of only 0.003 at SNR greater than 10dB, ensuring faithful and consistent image reconstruction. Future work may explore further optimizations to minimize inference time and extend the framework to high-fidelity semantic video communication.

Acknowledgements

This research has received funding from the European Union's Horizon Europe research and innovation programme MSCA-DN NESTOR (G.A. 101119983). The authors also acknowledge EPSRC project TRANSNET (EP/R035342/1). Experiments were run on Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

References

Agustsson, E. and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Bourtsoulatze, E., Kurka, D. B., and Gündüz, D. Deep joint source-channel coding for wireless image transmission.

IEEE Transactions on Cognitive Communications and Networking, 5(3):567–579, 2019.

Chen, W. and Yang, Q. Casc: Condition-aware semantic communication with latent diffusion models. *arXiv preprint arXiv:2411.06552*, 2024.

Cicchetti, G., Grassucci, E., Park, J., Choi, J., Barbarossa, S., and Comminiello, D. Language-oriented semantic latent representation for image transmission. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2024.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Grassucci, E., Barbarossa, S., and Comminiello, D. Generative semantic communication: Diffusion models beyond bit recovery. *arXiv preprint arXiv:2306.04321*, 2023.

Jiang, Z., Liu, X., Yang, G., Li, W., Li, A., and Wang, G. Diffsc: Semantic communication framework with enhanced denoising through diffusion probabilistic models. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 13071–13075. Institute of Electrical and Electronics Engineers Inc., 2024. ISBN 9798350344851. doi: 10.1109/ICASSP48485.2024.10448094.

Luo, X., Chen, H.-H., and Guo, Q. Semantic communications: Overview, open issues, and future research directions. *IEEE Wireless Communications*, 29(1):210–219, 2022.

Nam, H., Park, J., Choi, J., Bennis, M., and Kim, S.-L. Language-oriented communication with semantic coding and knowledge distillation for text-to-image generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13506–13510. IEEE, 2024.

Pernias, P., Rampas, D., Richter, M. L., Pal, C. J., and Aubreville, M. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023.

Qin, Z., Tao, X., Lu, J., Tong, W., and Li, G. Y. Semantic communications: Principles and challenges. *arXiv preprint arXiv:2201.01389*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Seitzer, M. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

Strinati, E. C. and Barbarossa, S. 6g networks: Beyond shannon towards semantic and goal-oriented communications. *Computer Networks*, 190:107930, 2021.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Xia, L., Sun, Y., Liang, C., Zhang, L., Imran, M. A., and Niyato, D. Generative AI for Semantic Communication: Architecture, Challenges, and Outlook. *IEEE Wireless Communications*, 32(1):132–140, 2025. doi: 10.1109/MWC.003.2300351.

Yang, P., Zhang, G., and Cai, Y. Rate-adaptive generative semantic communication using conditional diffusion models. *IEEE Wireless Communications Letters*, 14(2): 539–543, 2025.

Yilmaz, S. F., Niu, X., Bai, B., Han, W., Deng, L., and Gündüz, D. High perceptual quality wireless image delivery with denoising diffusion models. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 1–5. IEEE, 2024.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.