Language-oriented Semantic Communication for Image Transmission with Fine-Tuned Diffusion Model

Xinfeng Wei, Haonan Tong, Nuocheng Yang, and Changchuan Yin
Beijing Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, Beijing, China
Emails: {xinfengwei, hntong, yangnuocheng, and ccyin}@bupt.edu.cn

2024 16th International Conference on Wireless Communications and Signal Processing (WCSP) | 979-8-3503-9064-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/WCSP62071.2024.10827002

***Abstract*—Ubiquitous image transmission in emerging applications brings huge overheads to limited wireless resources. Since that text has the characteristic of conveying a large amount of information with very little data, the transmission of the descriptive text of an image can reduce the amount of transmitted data. In this context, this paper develops a novel semantic communication framework based on a text-2-image generative model (Gen-SC). In particular, a transmitter converts the input image to textual modality data. Then the text is transmitted through a noisy channel to the receiver. The receiver then uses the received text to generate images. Additionally, to improve the robustness of text transmission over noisy channels, we designed a transformer-based text transmission codec model. Moreover, we obtained a personalized knowledge base by fine-tuning the diffusion model to meet the requirements of task-oriented transmission scenarios. Simulation results show that the proposed framework can achieve high perceptual quality with reducing the transmitted data volume by up to 99% and is robust to wireless channel noise in terms of portrait image transmission.**

***Keywords*—Semantic communication, generative model, transformer, portrait transmission**

I. Introduction

In the future, data traffic of merging services is expected to continue increasing, posing significant challenges to resource-limited wireless networks. Especially, image transmission in extremely resource-constrained (i.e., spectrum and power) and harsh environments remains challenging. Therefore, transmitting images using fewer network resources has potential applications across various scenarios.

Semantic communication is an emerging research paradigm expected to enable efficient data transmission [1]–[4]. In [5], the deep joint source and channel coding (DeepJSCC) schemes have been developed to effectively compress and transmit images by optimizing the joint coding scheme to adapt to wireless channels. DeepJSCC encodes images based on the semantic features of the data to be transmitted, maintaining the visual information while achieving efficient image transmission.

However, in some cases (i.e., task-oriented communication), transmitting all features of an image is unnecessary. For instance, in the retail industry, a customer's facial expressions can reveal their preference for a product. In AR/VR applications, user expressions can enhance gaming or human-computer interaction experiences. In these scenarios, the image receiver's focus is not on the people's specific identity in the image but rather on the people's state, expressions, and other contextual information. Therefore, In order to reduce the transmission volume, information unrelated to the portrait state should be removed, and the remaining effective information can be compressed into text form to meet the needs of such communication scenarios.

Recently, generative artificial intelligence (GenAI) models have seen significant advancements, a large amount of innovative methods utilizing GenAI models have been introduced in the realm of semantic communication. [6] proposed an image transmission method based on descriptive information contained in images. In this method, the transmitter sends only the descriptive text information extracted from the original image using an image-to-text (I2T) algorithm to the receiver, and then the receiver reconstructs the image based on the received information using an image generation model. Textual data has the characteristic of conveying a large amount of information with very little data. In noisy channels, even a single letter error in the decoded text at the receiving end can lead to a meaning vastly different from the original. Therefore, it is necessary to enhance the noise resistance of text transmission methods. Moreover, without specific control input, it is difficult for generative models to generate high-fidelity portraits or styles from text alone.

This paper aims to address the above issues, the main contributions are:

- We propose a novel text-to-image semantic communication system. In the system under consideration, the transmitter converts images to text, transmits the text using a deep learning-based end-to-end text communication method, and reconstructs the images at the receiver using a text-to-image generative model.
- We designed transformer-based text transmission codec, enhancing the robustness of text transmission over noisy channels.
- We performed few-shot fine-tuning on the base diffusion model to generate high-fidelity portrait images so as to meet the needs of portrait image communication scenarios.

Simulation results demonstrate that in the task of portrait transmission, the proposed method achieves high perceptual similarity while effectively reducing the data volume and exhibiting robustness to noise. The remainder of this paper is organized as follows. The system model is presented in Section II. We give a detailed description of the proposed

This work was supported in part by Beijing Natural Science Foundation under Grant L223027, the National Natural Science Foundation of China under Grants 62471056, 61629101 and 61671086, the 111 Project under Grant B17007, in part by BUPT Excellent Ph.D. Students Foundation under Grant CX2021114, and China Scholarship Council.

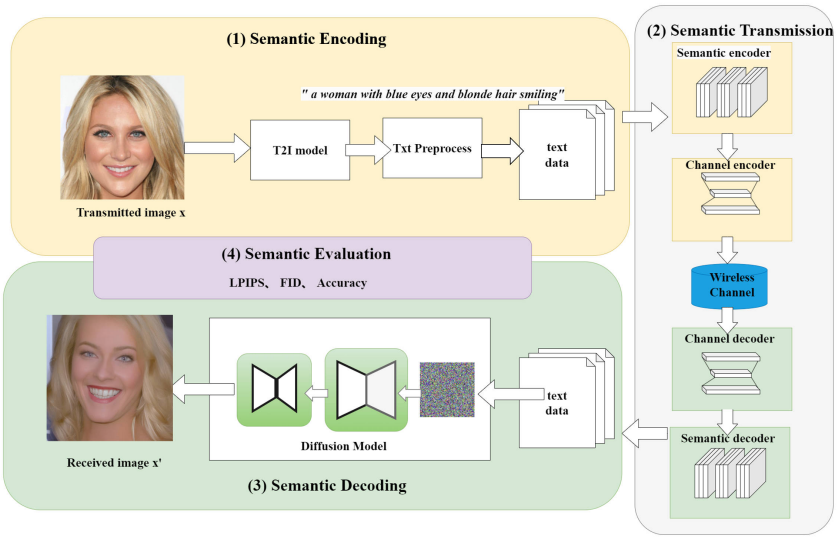

Fig. 1: The framework of semantic communication for networks.

semantic encoder and decoder model in section III. The simulation results are presented and analyzed in Section IV. The conclusion is given in Section V.

II. System Model And Problem Formulation

As shown in Fig. 1, our proposed framework consists of three main modules: semantic encoder through the img2txt model, semantic transmission involving the encoding and decoding of the text, and semantic decoder to reconstruct images based on received text.

A. Semantic Encoder

At the transmitter, the transmitter generates descriptive text data from the input image $\boldsymbol{v}$ using a pre-trained image-to-text (Img2Txt) model. The text data retains semantic alignment with the image data which contains the intended object information of the source image. The text data is presented as a sentence $\boldsymbol{s}$ in a specific order and is given by:

$$\boldsymbol{s} = \mathrm{I2T}(\boldsymbol{v}) = \left(\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{|\mathbf{S}|}\right), \tag{1}$$

where the function $\mathrm{I2T}(\cdot)$ represents an I2T encoder (i.e., bootstrapping language-image pre-training(BLIP) [7]). Through modality conversion, the data volume can be significantly reduced, and data redundancy can be greatly minimized.

B. Semantic Transmission

As shown in Fig.2, the sentence $\boldsymbol{s} = \left(s_1, s_2, \cdots, s_{|S|}\right)$ is fed into the transmission model, which consists of a transmitter and a receiver. The transmitter includes a semantic encoder and a channel encoder, both are implemented with neural networks. On the transmitting side, the semantic encoder extracts semantic information from $\boldsymbol{s}$ and maps to symbols $\boldsymbol{x}$, which are then transmitted over the physical channel by the channel encoder. Let the neural network parameters of the semantic encoder and the channel encoder be denoted as β and α, respectively. The encoded symbols$\boldsymbol{x}$ can be represented as:

$$\boldsymbol{x} = C_\alpha\left(S_\beta(\boldsymbol{s})\right), \tag{2}$$

where $S_\beta(\cdot)$ represents the semantic encoder with parameter β and $C_\alpha(\cdot)$ represents the channel encoder with parameter α. The encoded symbols $\boldsymbol{x}$ are transmitted over the physical channel, assuming $\boldsymbol{x}$ is normalized. The wireless channel is represented as $P_h(\boldsymbol{Y} \mid \boldsymbol{x})$, with $\boldsymbol{x}$ being the input and $\boldsymbol{y}$ being output. The transmission process through the wireless channel is given by

$$\boldsymbol{y} = \boldsymbol{h}\boldsymbol{x} + \boldsymbol{n}, \tag{3}$$

where h indicates the channel gain, and $\boldsymbol{n} \sim \mathcal{N}|(0, \sigma_n^2|\boldsymbol{I})$ is additive white Gaussian noise (AWGN) with variance σ_n^2 and $\boldsymbol{I}$ being identity matrix. The receiver comprises a channel decoder and a semantic decoder to recover the transmitted symbols and decode the sentence. The decoded sequence is given by

$$\hat{\mathbf{s}} = S_\chi^{-1}\left(C_\delta^{-1}(\boldsymbol{y})\right). \tag{4}$$

C. Semantic Decoder

We equip the receiver with a conditional image generative model, specifically utilizing Stable Diffusion [8] (SD), which has proven to be highly effective in generating images from textual descriptions. SD synthesizes an image starting from random noise and gradually refines it through a denoising process that is guided by a text prompt. This model consists of three main components: an encoder, which generates latent vectors into latent space from RGB image input; a denoiser, which performs the diffusion denoising process; and a decoder, which reconstructs the image from the latent vectors into RGB space. As depicted in Fig. 1, the diffusion model, under the guidance of the conditional text, iteratively reduce the noise by denoiser.

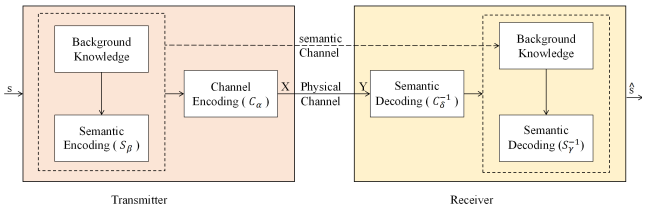

Fig. 2: The framework of end-to-end text transmission system.

III. Networks architecture

A. Text Transmission model

We use a transformer-based joint source-channel coding model, DeepSC [9], as the text transmission codec. To improve the effectiveness of Deep-SC in noisy channels, we introduced the bidirectional and Auto-Regressive transformers model (BART) [10], resulting in BART-SC. The BART-SC is shown in Fig.3. The transmitter is composed of a semantic encoder and a channel encoder, the input sequence $\boldsymbol{s}$ is encoded by the semantic encoder, compressed by the channel encoder for transmission. The receiver consists of a channel decoder and a semantic decoder. The channel decoder decompress the received data, and finally semantic decoder reconstructs the original semantic information. Specifically, the core of the encoder section is a module consisting of three transformer layers, each including multi-head self-attention mechanisms and a feed-forward neural network [11]. The channel encoder compresses the output of the encoder into a low-dimensional representation using two fully connected neural networks. The channel decoder recovers the received low-dimensional representation to its original high-dimensional form using fully connected neural networks and performs normalization through a layerNorm layer. The final output is a probability distribution over the target vocabulary, generated through a fully connected layer.

The BART model is a transformer-based sequence-to-sequence pre-training model that learns to compress, generate, and decompress text by applying random masking and text generation to the original input sequences. BART employs a special masking method called "noise mask" to generate noise in the input and enable self-supervised training. Additionally, BART uses an autoregressive approach during training, where the decoder generates the output step by step to ensure semantic consistency between the generated output and the original input. Based on the above reasons, we introduce the BART module into the DeepSC. The loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(\mathbf{s}, \hat{\mathbf{s}}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) + \lambda \mathcal{L}_{\text{MI}}(\boldsymbol{x}, \boldsymbol{y}; T, \boldsymbol{\alpha}, \boldsymbol{\beta}), \quad (5)$$

The first component is a cross-entropy loss function that measures the difference between $\hat{\mathbf{s}}$ and $\mathbf{s}$.

$$\mathcal{L}_{\text{CE}}(\mathbf{s}, \hat{\mathbf{s}}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\chi}, \boldsymbol{\delta}) = -\sum_{l=1} q(w_l) \log(p(w_l)) + (1 - q(w_l)) \log(1 - p(w_l)), \quad (6)$$

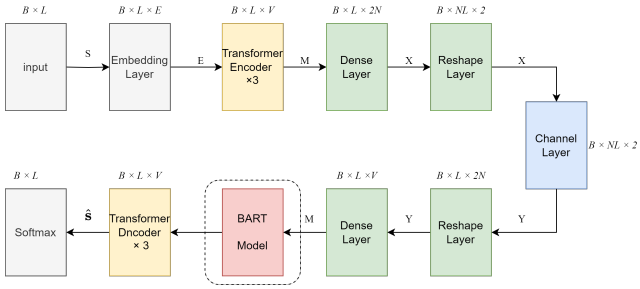

Fig. 3: The proposed neural network structure for end-to-end text transmission system.

For the l-th word in the estimated sentence, denoted as w_l, $q(w_l)$ represents its actual probability, while $p(w_l)$ represents the predicted probability of w_l appearing in the sentence $\hat{\mathbf{s}}$.

The second component is a mutual information loss function that quantifies the amount of information shared between the transmitted symbol, which aims to maximize the data rate during transmitter training. $I(X;Y)$ is the mutual information between the transmitted symbols x and the received symbols y.

$$\begin{aligned} \mathrm{I}(x,y) &= \int x \times y \times p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \\ &= \mathbb{E}_{p(x,y)} \left[\log \frac{p(x,y)}{p(x)p(y)} \right], \end{aligned} \quad (7)$$

We can optimize the encoder by maximizing the mutual information, which is expressed as:

$$\mathcal{L}_{\text{MI}}(\boldsymbol{x}, \boldsymbol{y}; T) = \mathbb{E}_{p(x,y)} [f_T] - \log \left(\mathbb{E}_{p(x)p(y)} \left[e^{f_T} \right] \right), \quad (8)$$

where f_T is composed by a neural network, in which the inputs are samples from $p(x,y)$, $p(x)$, and $p(y)$.

We implement a cross-training strategy to train BART-SC, we alternate training between the channel encoder/decoder and the semantic encoder/decoder models. Specifically, first, train the channel model and then freeze its parameters. Next, train the semantic model, freeze its parameters, and then retrain the channel model. This process can be repeated until both the semantic communication model and the channel model converge.

B. Image generation with fine-tuned stable diffusion

At the receiver, SD is utilized to generate the original image based on the received text. SD operates in a latent diffusion model, meaning it works within an autoencoder framework. Specifically, images are first encoded into a latent space by an encoder $\mathcal{E}$. The diffusion and reverse processes are applied in this latent space, and the resulting latent representations are then decoded back into image space by a decoder $\mathcal{D}$. More precisely, in the diffusion process, noise is progressively added to the original latent tensor x_0 which is converted by autoencoder from input image. The model iteratively adds Gaussian noise to x_0:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I\right), t = 1, .., T, \quad (9)$$

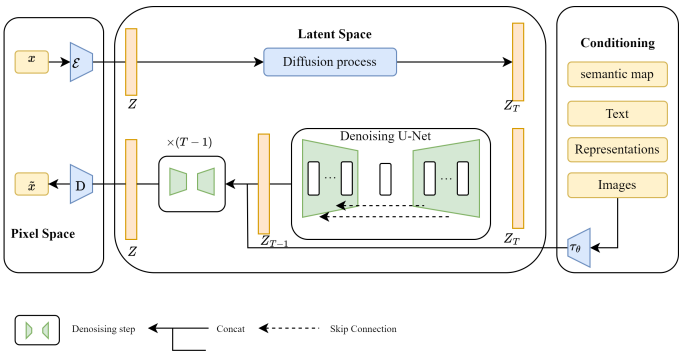

Fig. 4: The architecture of Stable Diffusion model.

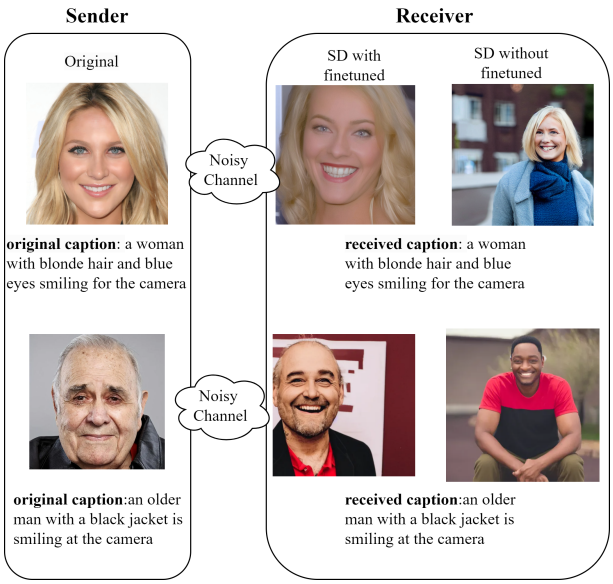

Fig. 5: Visual results. On the left side two randomly selected samples along with text which extracted by Img2Txt model. The generated images are on the right. The SNR is 9 dB.

where $q\left(x_t \mid x_{t-1}\right)$ is the conditional density of x_t given x_{t-1}, and $\{\beta_t\}_{t=1}^{T}$ are hyperparameters. T denotes the diffusion step. The reverse process is central to model training. The model learns to recover the clean latent tensor x_0 from a noisy latent tensor x_t:

$$p_\theta\left(x_{t-1} \mid x_t\right)=\mathcal{N}\left(x_{t-1} ; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right), \tag{10}$$

for $t = T, ..., 1$, which allows to generate a valid signal x_0 from the standard Gaussian noise step by step. Finally, x_0 is decoded back to the RGB space by autoencoder's decoder to get the generated image.

In a task-oriented communication scenario, SD can struggle to generate specific human portraits or styles without precise control inputs, which makes standard SD less capable of meeting the needs of users, particularly in situations where users have specific requirements for the generated style. Therefore, to produce specific portrait images, we enhanced the model by fine-tuning SD with additional sample data.

We use dreambooth [12] to fine-tune SD on a portrait dataset [13]. We follow the same loss objective as LoRA [14]. Let $\Theta(\cdot)$ denote all parameters of a model, G_s denotes the generative model after s iterations, then the hypothesis class at iteration s is:

$$\mathcal{G}_s=\left\{G \mid \operatorname{rank}\left(\Theta(G)-\Theta\left(G_s\right)\right) \leq R\right\}, \tag{11}$$

where R denotes the rank of weight updates and in practice we choose $R = 128$ to balance efficiency and image quality. This fine-tuning process refines the model's understanding of facial features, expressions, and textures, allowing it to produce images that not only resemble portraits more closely but also convey the subtle qualities and realism that are essential in high-quality portraiture.

C. Semantic Evaluation Metrics

For text transmission, we use the bilingual evaluation understudy (BLEU) [15] score, which is usually used to evaluate the quality of text produced by machine translation systems. It measures how closely the generated text matches reference text by comparing overlapping n-grams in machine translation, which can be described as:

$$\log \mathrm{BLEU}=\min\left(1-\frac{l_{\hat{\mathbf{s}}}}{l_{\mathbf{s}}}, 0\right)+\sum_{n=1}^{N} u_n \log p_n, \tag{12}$$

where u_n is the weight of n-grams and p_n is the n-grams score, which is:

$$p_n=\frac{\sum_k \min\left(C_k(\hat{\mathbf{s}}), C_k(\mathbf{s})\right)}{\sum_k \min\left(C_k(\hat{\mathbf{s}})\right)}, \tag{13}$$

where $C_k(\cdot)$ is the frequency count function for the k-th elements in n-th grams.

For image generation, semantic communication prioritizes semantic-level fidelity over pixel-level fidelity. We evaluate it based on both the quality of the reconstruction and the accuracy of the semantics.

- Quality of image reconstruction: We use learned perceptual image patch similarity (LPIPS) [16] metric. which assess the perceptual similarity of intended and generated images using additional neural networks, to evaluate the quality of semantically decoded images.
- Accuracy of semantics: The accuracy of classifications for expressions, gender, and age can be employed as key indicators when it comes to evaluating the accuracy of semantic reconstruction in portrait image transmission. Therefore, we measure the accuracy of key semantic information reconstruction (including age, gender, and expression) using several neural network classification models.

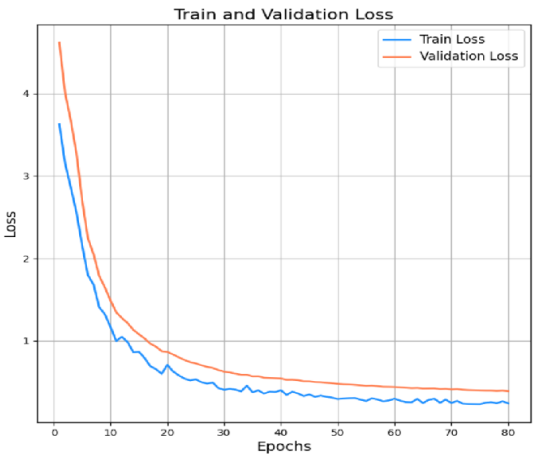

Fig. 6: Loss values vs. the number of training epochs

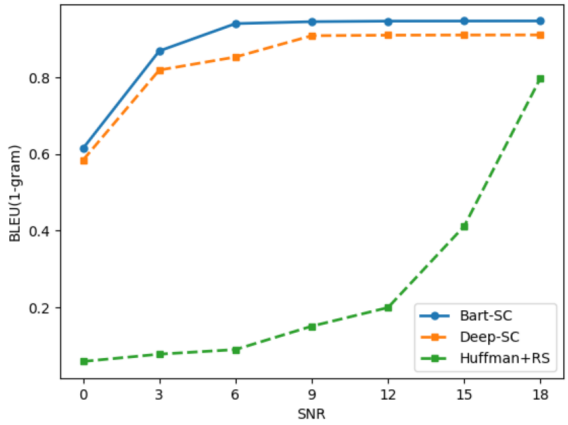

Fig. 7: BLEU score vs. SNR for the same total number of transmitted symbols, with Huffman coding with RS (30,42) in 64-QAM, DeepSC and our method.

IV. Simulation And Performance Analysis

We used the BLIP model to generate textual descriptions of 3000 images from the CelebA [13] dataset with 3000 sentences. Then, we train the BART-SC model on these 3000 sentences with fading channel with $h = 0.9$ and SNR from 5 to 10 dB. We employed the Stable Diffusion XL (SDXL) model [17] as the base model with pre-trained weights and we fine-tuned it with Dreambooth to generate high-quality portrait images. The fine-tuning is performed on the default hyperparameters, with a learning rate of 1e-5, and a maximum training epoch deployment of 500. We fine-tuned the SDXL for 10 epochs using 10 images from the CelebA dataset. For comparison, we evaluate the performance of the proposed Gen-SC by comparing it with standard SD without fine-tuned and traditional text communication link using Huffman coding, (5,7) Reed-Solomon (RS) coding, and 64 quadrature amplitude modulation (QAM), respectively.

In Fig. 6, we showed that the training loss and validation loss change as the number of training epochs varies. Training loss and validation loss decrease as epochs increase. This is due to the fact that the model calculates the loss function to measure the error between the predicted results and the actual labels during training. As the number of training epochs increases, the model parameters are progressively optimized.

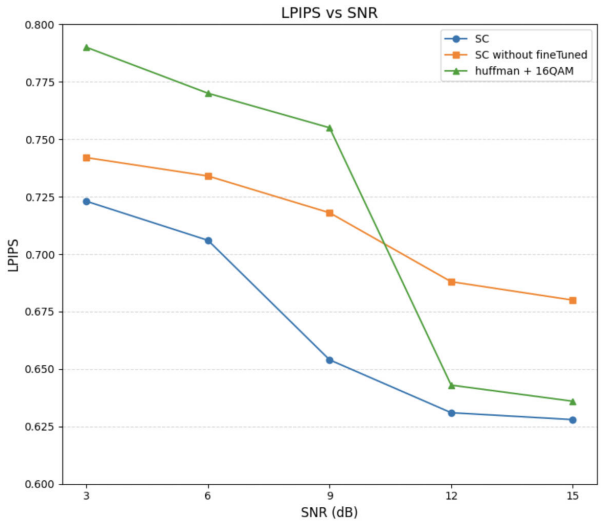

Fig. 8: LPIPS vs. SNR.

In Fig. 7, we showed the relation between the BLEU score and SNR under the same number of transmitted symbols over the AWGN channel. In the meanwhile, the traditional method uses 64-QAM modulation. All BLEU increases as SNR increases. This is due to the fact that when SNR increases, the distortion from noisy channels decreases. The BLEU of huffman+RS has the lowest in lower SNR than Bart-SC and Deep-SC. This is due to the fact that traditional methods are sensitive to channel error rates caused by channel variations while BART-SC and Deep-SC adopt JSCC that integrates source coding and channel coding into a single model through deep neural networks training, enabling better adaptation to the channel conditions. BART-SC outperforms Deep-SC in each SNR. This is due to the fact that BART model learns how to recover the original sequence from a noisy sequence during pre-training, which can enhance the model's robustness to noisy channels.

In Fig. 8, we show how the LPIPS changes as the SNR varies. We observed that the average LPIPS decreases as the SNR increases. Compared to traditional text transmission methods, Gen-SC contributes to a reduction of up to 0.1 in average LPIPS under low SNR conditions, with this reduction diminishing as the SNR increases. This indicates the potential benefits of optimizing Gen-SC levels based on given channel conditions for future research. Additionally, the fine-tuned stable diffusion model performs better in reconstructing images compared to the non-fine-tuned model, as shown in Fig.5.

In Fig. 9, we show the accuracy of age, gender, and expression in reconstructed images changes as the SNR varies. We observed that even in low SNRs, the accuracy of all three classifications remains above 80%. This reflects the effectiveness of Gen-SC in communication over noisy channels. Among

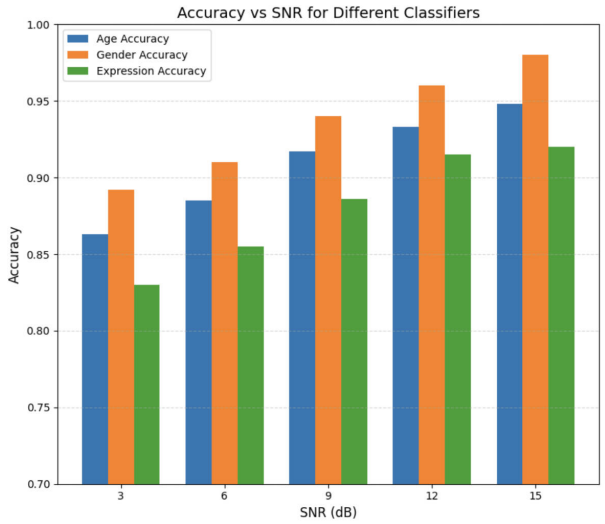

Fig. 9: Accuracy of age, gender and expression in reconstructed images versus the SNR

them, gender classification achieved the highest accuracy while expression classification had the lowest accuracy. This is due to the fact that SD tends to reconstruct some neutral expressions as negative expressions. For instance, a "serious look" might be reconstructed as an image with a frowning expression. The accuracy of age reconstruction was intermediate, with the testing revealing that the granularity for age reconstruction is relatively coarse.

V. Conclusion

In this paper, we proposed a Gen-SC framework for scenarios involving portrait transmission to achieve an efficient and robust semantic communication system. The main processes in Gen-SC include converting images into text, utilizing a transformer-based text transmission model, and employing a diffusion model for image reconstruction. Experimental results indicate that the proposed framework can significantly reduce the amount of transmitted data while preserving the semantic information. Additionally, the transformer-based transmission model provides better robustness to noisy wireless channels during text transmission than baseline methods. Finally, the scheme of fine-tuning the diffusion model enhanced the perceptional similarity in portrait image generation.

References

[1] W. Gong, H. Tong, S. Wang, Z. Yang, X. He, and C. Yin, "Adaptive bitrate video semantic communication over wireless networks," in *2023 International Conference on Wireless Communications and Signal Processing (WCSP)*, 2023.

[2] H. Tong, Z. Yang, S. Wang, Y. Hu, W. Saad, and C. Yin, "Federated learning based audio semantic communication over wireless networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[3] Y. Zeng, X. He, X. Chen, H. Tong, Z. Yang, Y. Guo, and J. Hao, "Dmce: Diffusion model channel enhancer for multi-user semantic communication systems," in *ICC 2024 - IEEE International Conference on Communications*, 2024, pp. 855–860.

[4] Q. Pan, H. Tong, J. Lv, T. Luo, Z. Zhang, C. Yin, and J. Li, "Image segmentation semantic communication over internet of vehicles," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.

[5] E. Bourtsoulatze, D. Burth, and D. GÃ¼ndÃ¼z, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[6] H. Nam, J. Park, J. Choi, and M. Bennis, "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)*. IEEE, 2024, pp. 13 506–13 510.

[7] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[8] R. Rombach, A. Blattmann, and D. Lorenz, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[9] H. Xie, Z. Qin, G. Li, and B. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, and A. Mohamed, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[11] A. V. aand N. Shazeer and N. Parmar, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 500–22 510.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[14] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[15] M. Belghazi, A. Baratin, and S. Rajeshwar, "Mutual information neural estimation," in *International conference on machine learning*. PMLR, 2018, pp. 531–540.

[16] R. Zhang, P. Isola, and A. Efros, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[17] D. Podell, Z. English, K. Lacey, and A. Blattmann, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.