# Beyond Prompts: Preserving Semantics in Diffusion-based Communication

**Wonjung Kim**
Seoul National University
dnjswnd116@snu.ac.kr

**Nakyung Lee**
Seoul National University
leena@cml.snu.ac.kr

**Sangwoo Hong**
Konkuk National University
swhong06@konkuk.ac.kr

**Jungwoo Lee**
Seoul National University
junglee@snu.ac.kr

## Abstract

Semantic communication seeks to transmit meaning rather than raw data. Recently, diffusion-based semantic communication has received a huge attention. Yet, recent diffusion-based approaches depend on external prompt generators, which can not preserve fine-grained details or require sending heavy image latents, incurring a robustness–bandwidth trade-off. We propose Textual Inversion-based Semantic Communication (TISC), a novel diffusion-based framework for semantic communication that learns compact, text-aligned latent tokens inside the model via textual inversion. These tokens serve as semantic carriers that condition generation, reducing reliance on external prompting and improving preservation of intended semantics under channel noise. TISC demonstrates consistently superior performance in communicating semantic meaning compared with strong baselines across a wide range of SNRs, and even under severely limited bandwidth.

## 1 Introduction

In recent years, the significance of information exchange has expanded, surpassing traditional data transmission to include a broad range of interactions between humans and machines. This change reflects the rapid evolution of communication technologies, which now must meet the complex demands of various edge devices enabled by artificial intelligence and autonomous systems. Within this rapidly transforming landscape, Semantic Communications (SC) has emerged as a crucial technology. Unlike traditional communication [1] paradigms that prioritize the accurate transmission of whole source data, SC aims to extract and convey the semantic information and context of the source data, ensuring that the core semantics remain intact throughout transmission. This approach not only optimizes resource utilization but also promotes the precise delivery of essential information, leveraging the properties of the desired task.

Among various SC methods, Joint Source-Channel Coding (JSCC) has emerged as a viable solution [2, 3]. By compressing the data while preserving semantic information under given channel conditions, JSCC has enabled more efficient semantic communication. Moreover, with developments in deep learning, this field has advanced further. Deep learning-based JSCC (DeepJSCC) [4, 5] optimizes compression and transmission processes simultaneously, offering greater robustness, especially under low signal-to-noise ratio (SNR) conditions. However, despite its benefits, JSCC faces limitations, particularly due to its complexity and low adaptability under various transmission conditions. To handle these limitations, generative models have been adopted in SC recently.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

Generative models in SC [6, 7] are utilized to leverage highly compressed semantic information to regenerate content, thus facilitating more bandwidth-efficient and expressive communication frameworks. Especially, utilizing Text-to-Image (T2I) diffusion models in SC has enabled significant performance gain, underscoring the importance of prompt-based systems. As a result, many existing diffusion-based SC studies [8, 9, 10] have also been conducted using the StableDiffusion model, a prominent T2I model. However, these studies face limitations as they typically employ only two extreme methods of communication: *i) transmitting prompts and generating the image based solely on the prompts,* or *ii) transmitting a combination of the original image's compressed latent and the prompt and using them for generation.* Transmitting only the prompts can lower the communication burden, but they show limited performance in preserving fine-grained details. On the other hand, sending latent can help generate the images with preserved details, but it requires significantly higher communication load than only sending the prompts. This limitation poses a significant challenge for semantic communication with preserved details, particularly under bandwidth limited scenarios. To address this problem, some studies have employed semantic maps instead of latent variables to convey conditional information. However, this approach still incurs high communication overhead, limiting its applicability in resource-constrained environments.

In this paper, we introduce Textual Inversion-based Semantic Communication (TISC), a novel diffusion-based framework designed for semantic-preserving image transmission. To effectively represent a specific image, TISC generates a semantic-preserving prompt by optimizing a new token embedding using the Textual Inversion (TI) technique [11]. TI learns a compact set of pseudo-word embeddings in the text encoder's vocabulary while keeping the backbone frozen. These vectors capture user- and content-specific semantics. The optimized tokens are integrated into the diffusion decoder's conditioning stream to guide image reconstruction under limited bandwidth and SNR. We demonstrate that this approach enables our model to preserve critical semantic information with minimal data transmission, resulting in the reconstruction of high-fidelity, meaningful outputs.

## 2 Related Works

### 2.1 Diffusion Models

Diffusion models are a class of generative models that learn to synthesize data by reversing a process that systematically corrupts a data sample with noise. Ultimately, the objective is to train a model capable of generating samples that are indistinguishable from the true data distribution. Denoising Diffusion Probabilistic Models (DDPMs) [12] have demonstrated the generation of high-fidelity samples by learning to reverse a fixed Markov noising process. However, operating directly in the high-dimensional pixel space makes DDPMs computationally expensive. To mitigate this, Latent Diffusion Models (LDMs) [13] were proposed. LDMs apply the diffusion process within a lower-dimensional latent space learned by a Variational Autoencoder (VAE), significantly reducing computational overhead for both training and inference.

### 2.2 Generative Semantic Communication

Following the success of DeepJSCC, research rapidly expanded from generic image transmission to task-specific scenarios (e.g., classification [14], text [15], and multimodal [16]), demonstrating strong low-SNR robustness through end-to-end neural mappings from pixels to complex channel symbols. However, the learned latent was forced to be both a compact semantic representation and channel-robust, constraining fine-texture preservation under tight bit-rate or SNR budgets [17].

To mitigate this limitation, a hybrid approach has been explored in [18], which attached generative heads to DeepJSCC outputs, improving perceptual realism but still relying on the fidelity of the initial autoencoder reconstruction. More recently, a unified solution [19] leverages large-scale diffusion backbones as universal decoders, transmitting only small semantic specifications and performing high-fidelity synthesis at the receiver. In parallel, modality transformation via vision–language models (VLMs) compresses images into compact captions that can be rendered by T2I diffusion models, achieving notable bit-rate savings [9]. However, caption-only pipelines still struggle with preserving fine-grained details, motivating representations that go beyond text alone.

# 3 TISC: Robust Textual Inversion for Semantic Communication

## 3.1 Overall Framework

We present the proposed overall framework of TISC in Figure 1. A key assumption is that both the transmitter and receiver share a pre-trained diffusion model as common knowledge. This model serves as a T2I generator, producing the target image for communication from random Gaussian noise through a series of denoising steps conditioned on a textual prompt.

The procedure consists of four steps. Step 1 generates an appropriate prompt using attention scores from BLIP [20]. Step 2 trains TI with the generated prompt and incorporates noise injection to enhance channel robustness. Step 3 transmits the generated prompt and the learned embedding via digital and analog communication, respectively. Finally, Step 4 reconstructs the intended image at the receiver using the received text and embedding information.

We model the communication environment as a single-input single-output (SISO) system operating over an additive white Gaussian noise (AWGN) channel. Specifically, TISC employs a hybrid communication scheme: (i) digital transmission of the standard text prompt and (ii) analog transmission of the newly trained token embedding.

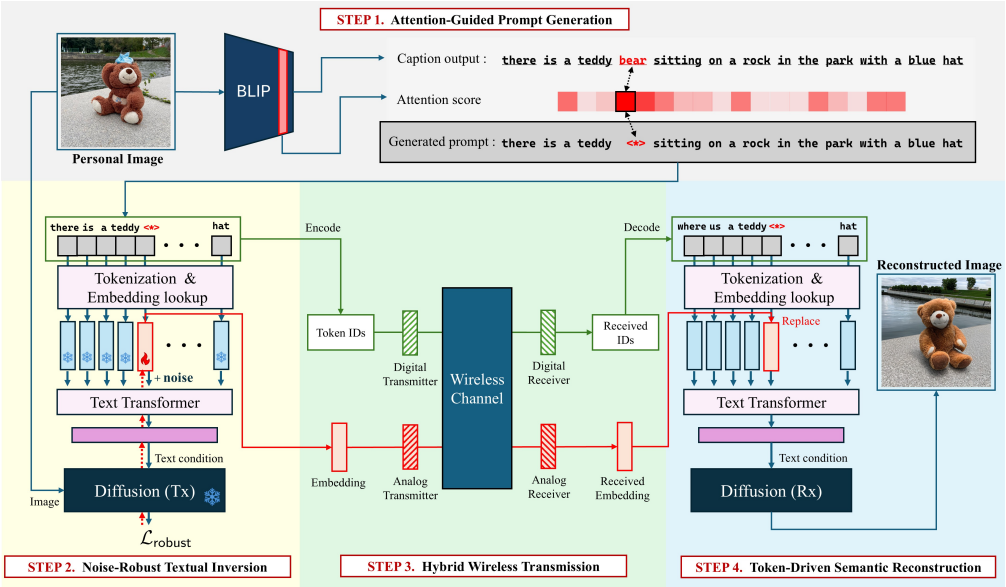



Figure 1: Proposed overall framework TISC comprising four steps of communication, with semantics flowing sequentially through each step.

## 3.2 Attention-Guided Prompt Generation

To obtain semantic-preserving and descriptive textual cues specific to an image, we first generate a caption using a BLIP-based image captioning model. We then compute self-attention over the caption tokens and select those with the highest attention weights. These selected tokens are then replaced with special tokens, obtained by TI.

It should be noted that we exclude special tokens such as <BOS> and <EOS> to mitigate the *attention sink* [21] phenomenon commonly observed in language models, where attention disproportionately concentrates on the initial tokens. Although such tokens may serve as structural anchors during training, they provide little semantic value. Thus, we restrict selection to content-bearing tokens in order to identify semantically informative cues for explainability. To ensure more reliable estimates of token importance, we average the attention weights across the last seven layers.

### 3.3 Prompt Optimization with Noise-Robust Textual Inversion

The primary objective of our system is to transmit semantic-preserving images by first converting their core semantic content into a compact, text-like representation. To achieve that goal, our approach is based on TI. TI learns an embedding, denoted as $v_{\langle * \rangle}$, to represent a novel concept within a pre-trained diffusion model. The optimization follows the model's inherent score-matching objective, which minimizes the difference between the ground-truth noise $\epsilon$ and the predicted noise $\epsilon_\theta$. The standard loss function is formulated as

$$\mathcal{L} = \mathbb{E}_{x,c,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(x_t, t, c(v_{\langle * \rangle}))\right\|_2^2\right], \tag{1}$$

where $x_t$ is the noisy image and $c(v_{\langle * \rangle})$ is the text conditioning containing the token $v_{\langle * \rangle}$. To enhance the robustness of this learned embedding against perturbations, we introduce artificial noise directly into the embedding vector during training. This is achieved by defining a noisy embedding $v'_{\langle * \rangle}$ as

$$v'_{\langle * \rangle} = v_{\langle * \rangle} + \delta, \tag{2}$$

where the noise $\delta$ is sampled from a zero-mean Gaussian distribution, $\mathcal{N}(0, \sigma^2 I)$, with $\sigma$ controlling the noise magnitude. By substituting this noisy embedding into the standard objective, we define our proposed robust loss as follows

$$\mathcal{L}_{\text{robust}} = \mathbb{E}_{x,c,\epsilon,t,\delta}\left[\|\epsilon - \epsilon_\theta(x_t, t, c(v'_*))\|_2^2\right] = \mathbb{E}_{x,c,\epsilon,t,\delta}\left[\left\|\epsilon - \epsilon_\theta(x_t, t, c(v_{\langle * \rangle} + \delta))\right\|_2^2\right]. \tag{3}$$

Optimizing this objective encourages a more stable representation by acting as a form of regularization, leading to a more generalized embedding that is less sensitive to channel noise. Throughout this process, only the embedding vector $v_{\langle * \rangle}$ is updated while the model weights $\theta$ remain frozen.

### 3.4 Hybrid Communication through Wireless Channel

After the robust TI process, the trained embedding and the generated prompt are transmitted through the hybrid wireless channel, which integrates both digital and analog transmission. At this stage, the two resources are treated differently. The text data are digitized by mapping indices from the tokenizer's vocabulary (token IDs) to bits and then modulated, while the embedding vector $v_{\langle * \rangle}$ is transmitted in the analog domain without quantization, where its continuous values are directly mapped onto channel symbols. This design separates the digital transmission of token IDs from the analog transmission of learned embeddings, making the overall system naturally align with the receiver's tokenizer inputs.

### 3.5 Token-Driven Image Reconstruction at the Receiver

Based on the two resources received through the wireless channel, the receiver performs image generation using a pre-trained T2I model. In this process, the prompt transmitted through the channel is used as the text condition, and the token embedding corresponding to $\langle * \rangle$ is filled with the embedding obtained from the channel. Guided by this text condition, the model iteratively denoises Gaussian noise to reconstruct an image that preserves the intended semantics, enabling the receiver to obtain the output image with preserved details.

## 4 Experimental Results

### 4.1 Datasets and Evaluation Settings

We evaluate our method on the task of personalized concept learning using the DreamBooth dataset [22], with all experiments built upon the pre-trained StableDiffusion v1.5 model. Our primary goal is to demonstrate the robustness of the learned embedding when transmitted over a noisy and resource-constrained communication channel.

To validate the performance of our method, we establish the following key settings. First, for noise-robust TI, we implement SNR of -5 dB, with QPSK modulation for the digital channel. Second, to represent the transmission constraint, we define a compression ratio $\rho$. $\rho$ is the proportion of transmitted resources to the input image dimensions, which can be formulated as $\rho = \frac{B}{C \times H \times W}$,

where $B$ denotes the total transmitted resources, and $C, H, W$ are the channel, height, and width of the input image, respectively. To maintain a consistently challenging environment, we fix this ratio at a highly constrained value of $\rho = \frac{1}{1024}$ for all experiments, which implies that all the communicated messages must be compressed to a size of $\frac{1}{1024}$. Furthermore, our system is evaluated against the following baselines using the metrics listed below.

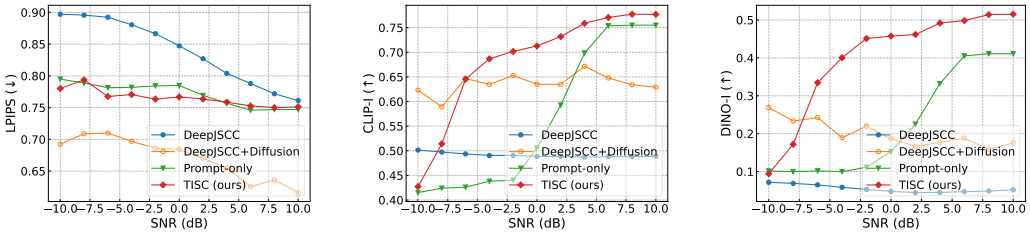
#### 4.1.1 Baselines

- DeepJSCC [4]: A pioneering deep learning-based semantic communication framework that directly maps image pixels to complex channel symbols via a CNN autoencoder. This end-to-end design effectively mitigates the "cliff effect" [1] observed in conventional digital schemes, where slight channel degradation lead to abrupt performance collapse.
- DeepJSCC+Diffusion [19]: An enhanced variant of DeepJSCC that incorporates a generative diffusion model at the receiver to improve perceptual quality. The received image is treated as a noisy input, which is iteratively refined through diffusion-based denoising, thereby restoring fine visual details.
- Prompt-only [9]: A language-driven transmission paradigm that reformulates image communication as a text-based task. VLM (e.g., BLIP) first generates a caption describing the source image, which is then digitally transmitted. At the receiver, a text-to-image generation model reconstructs the image conditioned on the received caption.

#### 4.1.2 Metrics

We evaluate the performance of SC with three metrics that measure image similarity. **LPIPS** [23] computes a perceptual distance from multi-layer CNN activations such as VGG or AlexNet, where lower values indicate higher perceptual fidelity. **CLIP-I** [24] measures cosine similarity between CLIP image embeddings and reflects high-level semantic agreement. **DINO-I** [25] reports cosine similarity in the self-supervised DINO feature space and emphasizes object and shape structure.

### 4.2 Simulation Results




(a) LPIPS (↓) (b) CLIP-I (↑) (c) DINO-I (↑)

Figure 2: Quantitative comparison of diffusion-aided SC baselines and our TISC across SNRs.

We compare our method with established baselines. Figure 2 summarizes quantitative performance across SNRs using LPIPS, CLIP-I, and DINO-I, and Figure 3 provides a reconstructed image.

Under an extreme compression setting (small $\rho$), vanilla **DeepJSCC** struggles to reconstruct the source even in the high-SNR regime. Even augmenting DeepJSCC with a pretrained diffusion prior (**DeepJSCC+Diffusion**) does not entirely resolve this problem. Since the channel output can show a significant deviation from the ground truth, the editing prior may generate false details or risk drifting from the original objective. Additionally, because instance-specific information is not provided, the pretrained model is limited in its ability to restore specific semantic-preserving information, which the transmitter wants to convey. On the other hand, the **Prompt-only** approach relies solely on text captions, which, even in high-SNR conditions, remain under-specified and fail to convey important fine-grained semantics. In contrast, our proposed **TISC** generates images that align with the transmitter's intended semantics across the various SNRs and consistently outperforms the baselines in both quantitative and qualitative evaluations.

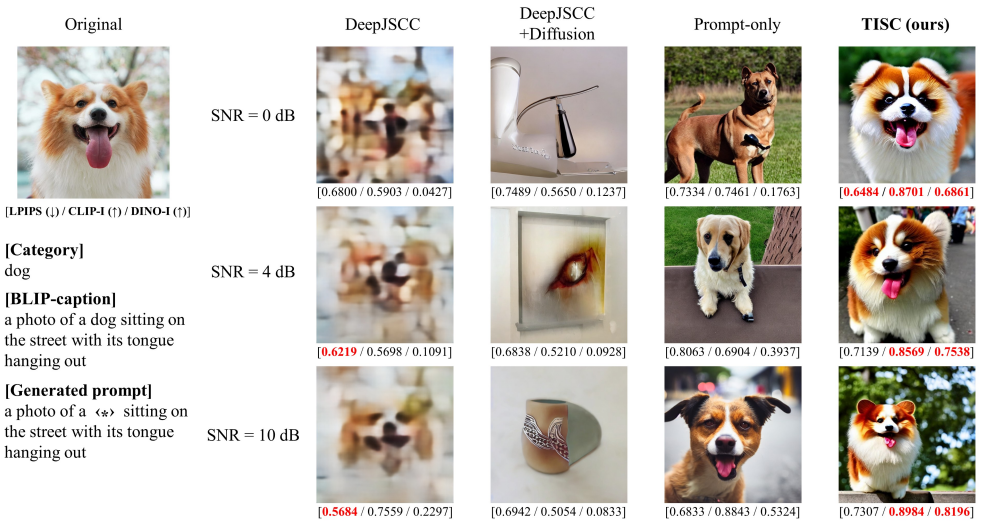
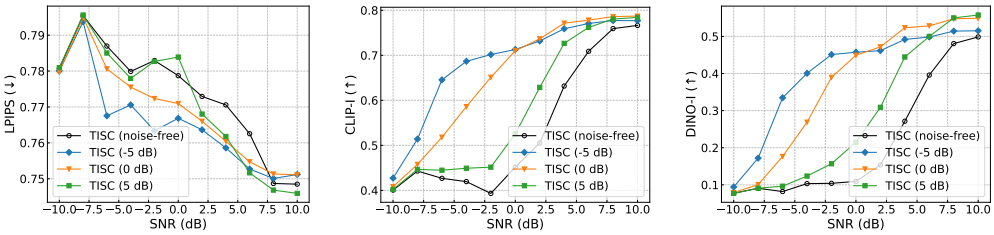



Figure 3: Qualitative reconstructions across SNRs for diffusion-aided SC baselines and our TISC.




(a) LPIPS (b) CLIP-I (c) DINO-I

Figure 4: Comparison of noise-robust TI performance across training-time embedding noise levels.

We also analyze how noise-robust TI affects the performance of TISC across varying SNRs. Models were trained at SNRs of -5, 0, and 5 dB, as well as in a noise-free setting. As shown in Fig. 4, the benefits of noise-robust TI are most pronounced in the low-SNR regime. We find that as the noise level for TI increases, the learned embeddings become more resilient. In particular, the model trained at -5 dB exhibits the strongest robustness across all SNRs, even those outside its training range. This robustness, however, comes at the cost of peak performance. In the high-SNR regime, the -5 dB trained model underperforms those trained at higher SNRs.

## 5 Conclusion

In this paper, we presented TISC, a diffusion-based framework for generative semantic communication that can preserve the fine-grained details. TISC reduces dependence on external prompt generators by coupling attention-guided prompt refinement with noise-robust textual inversion, learning compact, text-aligned tokens that better preserve user-specific semantics under channel impairments. Through simulations against strong diffusion-aided SC baselines, we demonstrate consistently superior semantic fidelity and perceptual quality across a broad range of SNRs, including scenarios with severely limited bandwidth. Future work will focus on efficient digital representations and channel coding for the learned tokens, integration with text-only SC (e.g., DeepSC), and extensions to MIMO and multi-user settings.

### Acknowledgments

This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(20%)), Institute of Information & communications Technol-

ogy Planning & Evaluation (IITP, RS-2021-II212068(10%), RS-2025-02305453(15%), RS-2025-02273157(15%), RS-2025-25442149(10%) RS-2021-II211343(10%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2025.

## References


[1] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[2] Jincheng Dai, Sixian Wang, Kailin Tan, Zhongwei Si, Xiaoqi Qin, Kai Niu, and Ping Zhang. Nonlinear transform source-channel coding for semantic communications. *IEEE Journal on Selected Areas in Communications*, 40(8):2300–2316, 2022.

[3] Sixian Wang, Jincheng Dai, Zijian Liang, Kai Niu, Zhongwei Si, Chao Dong, Xiaoqi Qin, and Ping Zhang. Wireless deep video semantic transmission, 2022.

[4] Eirina Bourtsoulatze, David Burth Kurka, and Denız Gündüz. Deep joint source-channel coding for wireless image transmission, 2019.

[5] Jia lin Xu, Tze-Yang Tung, Bo Ai, Wei Chen, Yuxuan Sun, and Deniz Gunduz. Deep joint source-channel coding for semantic communications, 2022.

[6] Lei Guo, Wei Chen, Yuxuan Sun, Bo Ai, Nikolaos Pappas, and Tony Q. S. Quek. Diffusion-driven semantic communication for generative models with bandwidth constraints. *IEEE Transactions on Wireless Communications*, 24(8):6490–6503, 2025.

[7] Eleonora Grassucci, Yuki Mitsufuji, Ping Zhang, and Danilo Comminiello. Enhancing semantic communication with deep generative models: An overview. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13021–13025, 2024.

[8] Xinfeng Wei, Haonan Tong, Nuocheng Yang, and Changchuan Yin. Language-oriented semantic communication for image transmission with fine-tuned diffusion model. In *2024 16th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1456–1461. IEEE, 2024.

[9] Hyelin Nam, Jihong Park, Jinho Choi, Mehdi Bennis, and Seong-Lyun Kim. 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)*, pages 13506–13510. IEEE, 2024.

[10] Giovanni Pignata, Eleonora Grassucci, Giordano Cicchetti, and Danilo Comminiello. Lightweight diffusion models for resource-constrained semantic communication. *arXiv preprint arXiv:2410.02491*, 2024.

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[14] Huiqiang Xie, Zhijin Qin, Geoffrey Ye Li, and Biing-Hwang Juang. Deep learning enabled semantic communication systems. *IEEE Transactions on Signal Processing*, 69:2663–2675, 2021.

[15] Huiqiang Xie, Zhijin Qin, Geoffrey Ye Li, and Biing-Hwang Juang. Deep learning enabled semantic communication systems. *IEEE transactions on signal processing*, 69:2663–2675, 2021.

[16] Guangyi Zhang, Qiyu Hu, Zhijin Qin, Yunlong Cai, Guanding Yu, and Xiaoming Tao. A unified multi-task semantic communication system for multimodal data. *IEEE Transactions on Communications*, 72(7):4101–4116, 2024.

[17] Jun Wang, Sixian Wang, Jincheng Dai, Zhongwei Si, Dekun Zhou, and Kai Niu. Perceptual learned source-channel coding for high-fidelity image semantic transmission. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pages 3959–3964. IEEE, 2022.

[18] Shunpu Tang, Qianqian Yang, Deniz Gündüz, and Zhaoyang Zhang. Evolving semantic communication with generative model. *arXiv preprint arXiv:2403.20237*, 2024.

[19] Selim F Yilmaz, Xueyan Niu, Bo Bai, Wei Han, Lei Deng, and Deniz Gündüz. High perceptual quality wireless image delivery with denoising diffusion models. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–5. IEEE, 2024.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[21] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2023.

[22] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir 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*, pages 22500–22510, 2023.

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[25] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

# A Implementation Details

## A.1 Hybrid Communication Details

For the digital path, a text prompt $s$ is first tokenized into a sequence of token IDs $\mathbf{i} = (i_1, \ldots, i_T)$ with $i_t \in \{0, \ldots, V-1\}$, where $V$ denotes the vocabulary size. Each token ID is converted into a bit sequence of length $\lceil \log_2 V \rceil$ and subsequently mapped onto $M$-QAM symbols $x_t \in \mathcal{X}_M$ that are normalized to satisfy $\mathbb{E}[|x_t|^2] = 1$. The AWGN channel output is given by

$$y_t = x_t + n_t, \quad n_t \sim \mathcal{CN}(0, \sigma^2), \tag{4}$$

where $\sigma^2$ is determined by the target $\mathrm{SNR} = E_b/N_0$. At the receiver, $y_t$ is demodulated to recover $\hat{i}_t$, and the sequence $\hat{\mathbf{i}}$ is detokenized to obtain the reconstructed text $\hat{s}$.

For the analog path, a placeholder token $\langle * \rangle$ is represented by its learned embedding $\mathbf{v} \in \mathbb{R}^D$. Prior to transmission, the embedding is normalized as

$$\tilde{\mathbf{v}} = \frac{\mathbf{v}}{\sqrt{\frac{1}{D}\|\mathbf{v}\|^2}}, \quad \frac{1}{D}\|\tilde{\mathbf{v}}\|^2 = 1, \tag{5}$$

so that its average power per dimension equals one. The AWGN channel output is expressed as

$$\mathbf{y}_a = \tilde{\mathbf{v}} + \mathbf{n}_a, \quad \mathbf{n}_a \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_D). \tag{6}$$

At the receiver, the noisy embedding $\mathbf{y}_a$ is directly injected into the text encoder. At this stage, we assume that the norm $\|\mathbf{v}\|$ is shared between the transmitter and the receiver in order to restore the original embedding.

## A.2 Training Configuration

For TI training, we employed the AdamW optimizer with a constant learning rate of $1 \times 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay of $0.01$. The model was trained for 200 steps with a batch size of 16, and the placeholder token was initialized using the pre-existing token "image". All experiments were conducted on a single NVIDIA H100 GPU.

## A.3 Evaluation Metrics

For metric computation, we adopted the same evaluation setup as described in the main text. Specifically, LPIPS was implemented using the VGG-16 backbone, CLIP-I was computed with the OpenAI CLIP ViT-B/32 model, and DINO-I was based on the ViT-S/16 checkpoint (facebook/dino-vits16). All images were resized to $256 \times 256$ and normalized to $[-1, 1]$ before evaluation. Each metric was averaged over the entire test set, and results are reported as mean values.

# B Qualitative Results

For the qualitative evaluation, we compare our model with the baselines across specific SNRs (0 dB, 4 dB, and 10 dB). The images shown for comparison are selected as the best among 64 generations, based on the highest CLIP-I score.As illustrated in Figures 5, 6, 7, and 8, the naive **DeepJSCC** model exhibits severe degradation at low SNR. As the SNR increases, the model produces a blurry but more recognizable reconstruction of the original image. However, in **DeepJSCC+Diffusion** model, when a diffusion-based image editing effect is applied to the unstable output, the resulting generation preserves the color palette but its structure diverges entirely from the original. Furthermore, the **Prompt-only** model fails to restore the original image's form at a low SNR of 0 dB. While its reconstruction of the shape improves with increasing SNR, it tends to miss the specific characteristics of target images for communication due to the inherent limitations in the expressive capacity of captions. In contrast, our proposed **TISC** successfully generates images that faithfully retain the original's distinct features across all tested SNR conditions. Notably, it achieves state-of-the-art performance in most cases when compared to the baselines, as demonstrated by the superior CLIP-I and DINO-I scores.

Original

[LPIPS (↓) / CLIP-I (↑) / DINO-I (↑)]

**[Category]**
backpack_dog

**[BLIP-caption]**
a photo of a bag on a window sill with buildings outside

**[Generated prompt]**
a photo of a ‹*› on a window sill with buildings outside

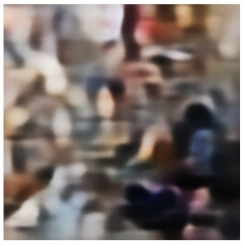
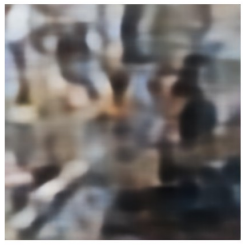
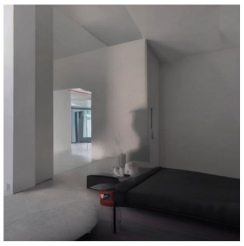
| | SNR = 0 dB | SNR = 4 dB | SNR = 10 dB |
|---|---|---|---|
| DeepJSCC |  |  | |
| | [0.7000 / 0.5356 / 0.0128] | [0.6755 / 0.5347 / 0.0175] | [**0.6261** / 0.4966 / 0.0677] |
| DeepJSCC +Diffusion |  | | |
| | [0.7002 / 0.3774 / 0.0887] | [**0.6565** / 0.5264 / 0.2262] | [0.6485 / 0.6000 / 0.1329] |
| Prompt-only | | | |
| | [**0.6955** / 0.4763 / 0.0360] | [0.7159 / 0.6416 / 0.1608] | [0.6878 / 0.6221 / 0.3020] |
| **TISC (ours)** | | | |
| | [0.7435 / **0.7539** / **0.4407**] | [0.7079 / **0.7578** / **0.3295**] | [0.6859 / **0.8115** / **0.4190**] |

Figure 5: Qualitative example for backpack dog.

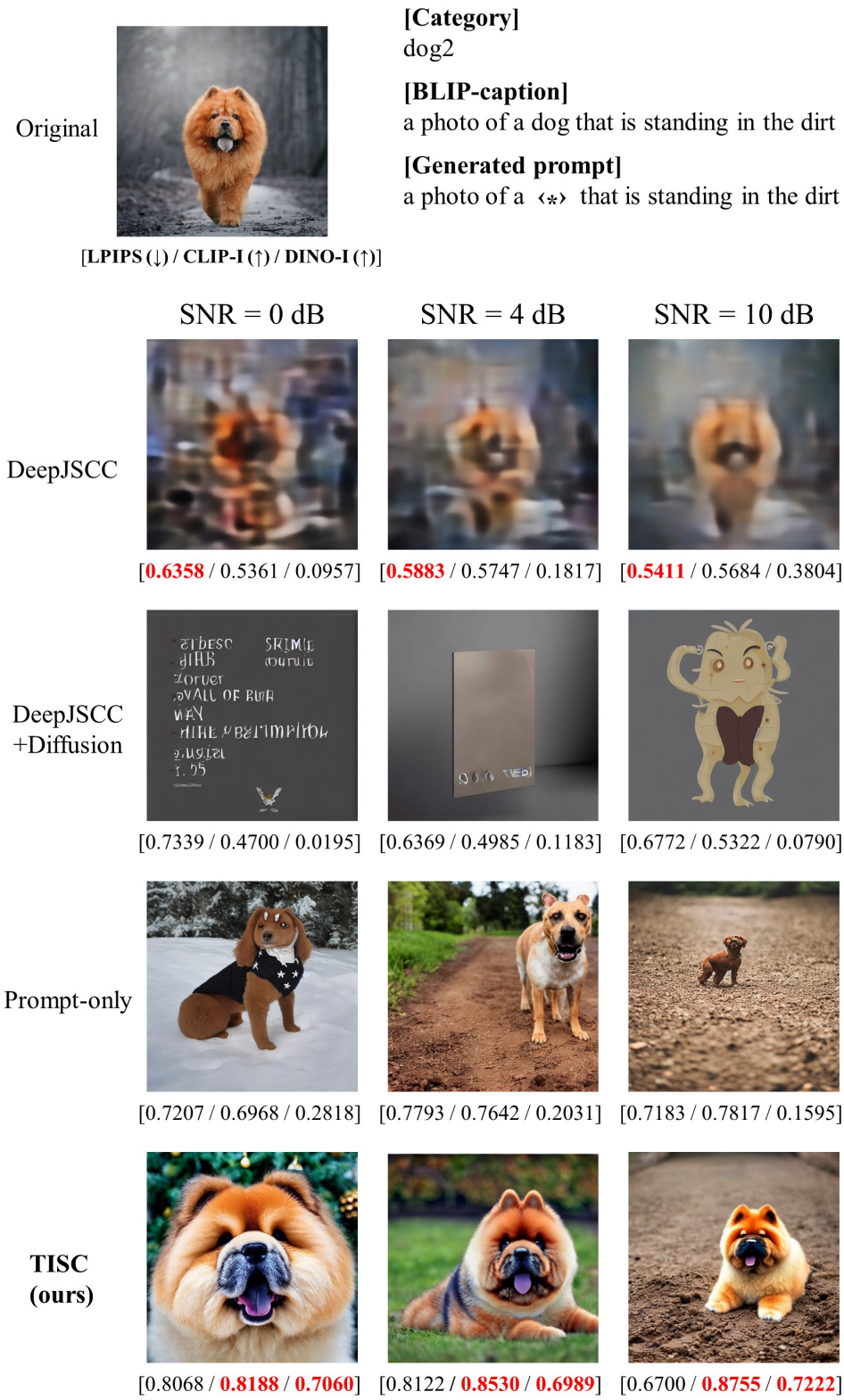



Figure 6: Qualitative example for dog2.

Original

[LPIPS (↓) / CLIP-I (↑) / DINO-I (↑)]

**[Category]**
monster_toy

**[BLIP-caption]**
a photo of a stuffed animal sits on a windowsill

**[Generated prompt]**
a photo of a stuffed animal sits on a ‹*›

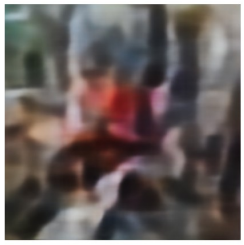
| | SNR = 0 dB | SNR = 4 dB | SNR = 10 dB |
|---|---|---|---|
| DeepJSCC | |  | |
| | [0.7191 / 0.5288 / 0.1113] | [**0.6877** / 0.4934 / 0.1679] | [**0.6531** / 0.5332 / 0.1825] |
| DeepJSCC +Diffusion | | | |
| | [**0.7078** / 0.4500 / 0.1120] | [0.7231 / 0.4980 / 0.1186] | [0.7155 / 0.5049 / 0.1929] |
| Prompt-only | | | |
| | [0.8100 / 0.4927 / **0.3766**] | [0.7225 / 0.7036 / **0.5165**] | [0.7370 / 0.7021 / 0.4595] |
| **TISC (ours)** | | | |
| | [0.7488 / **0.7310** / 0.3626] | [0.7900 / **0.7529** / 0.4300] | [0.7532 / **0.7769** / **0.5218**] |

Figure 7: Qualitative example for monster toy.

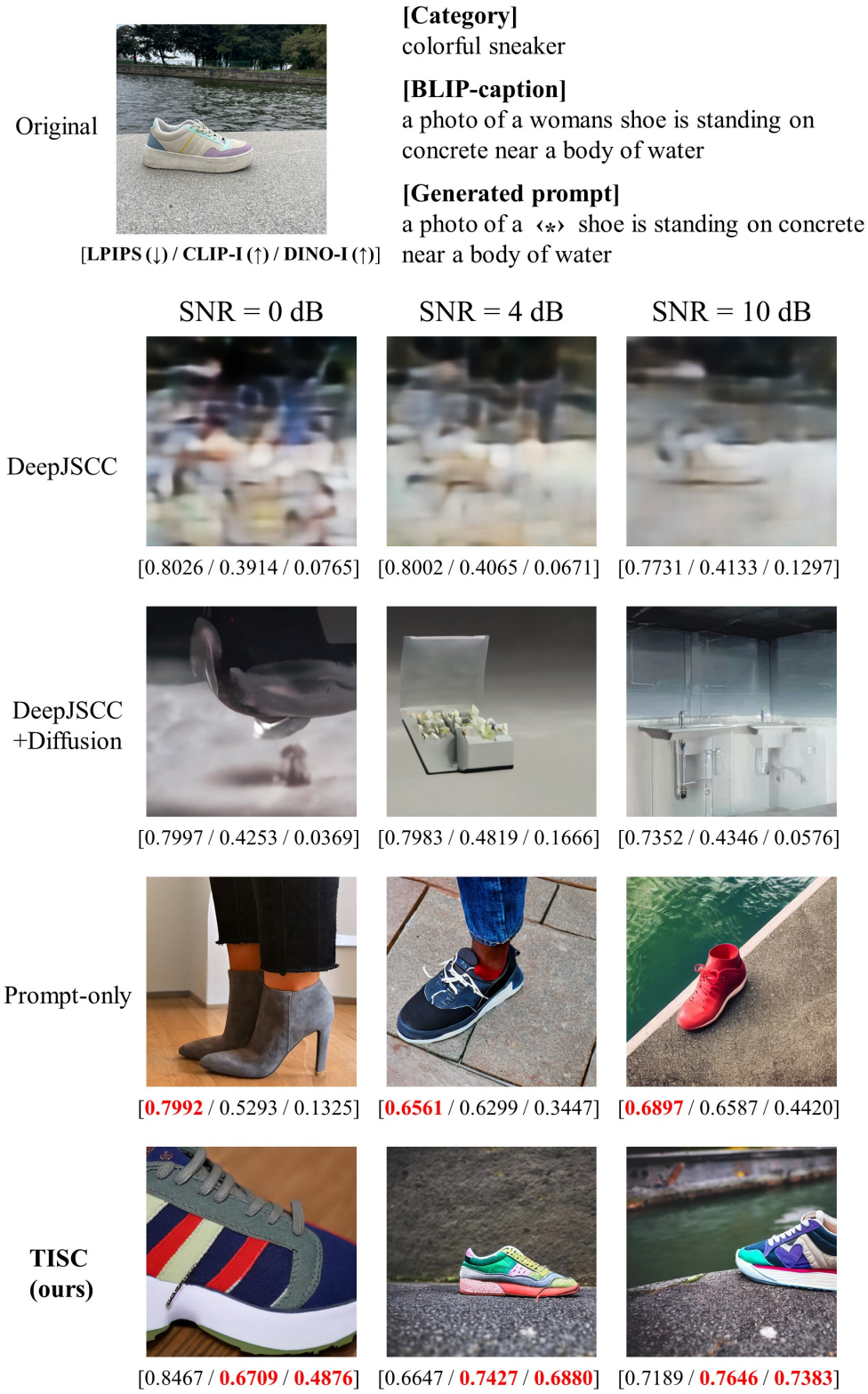



Figure 8: Qualitative example for colorful sneaker.