SING: Semantic Image Communications using Null-Space and INN-Guided Diffusion Models

Jiakang Chen*†, *Graduate Student Member, IEEE*, Selim F. Yilmaz*,
Di You*, *Graduate Student Member, IEEE*, Pier Luigi Dragotti, *Fellow, IEEE*, and Deniz Gündüz, *Fellow, IEEE*

***Abstract*—Joint source-channel coding systems based on deep neural networks (DeepJSCC) have recently achieved remarkable performance in wireless image transmission. Existing methods primarily minimize pixel-level distortion between the transmitted image and its reconstruction at the receiver, often overlooking perceptual quality. This can lead to severe perceptual degradation when transmitting images under extreme conditions, such as low bandwidth compression ratios (BCRs) and low signal-to-noise ratios (SNRs). In this work, we propose SING, a novel two-stage JSCC framework that formulates the recovery of high-quality source images from corrupted reconstructions as an inverse problem. Depending on the prior knowledge of the DeepJSCC encoder/decoder and channel, SING either approximates the stochastic degradation by a linear proxy (SING-Zero), or learns a nonlinear proxy using invertible neural networks (INNs) (SING-INN). Both approaches enable the seamless integration of pretrained unconditional diffusion models as a generative prior, enhancing perceptual quality. Experimental results demonstrate that SING outperforms DeepJSCC and other approaches, delivering superior perceptual quality even under extremely challenging conditions, including scenarios with significant distribution mismatches between the training and test data.**

***Index Terms*—Semantic communications, joint source-channel coding, inverse problems, diffusion models, invertible neural networks.**

I. Introduction

Shannon's separation theorem [1] laid the foundations for modern communication systems. In a typical point-to-point digital communication system, the encoding process at the transmitter involves two steps: source coding followed by channel coding and modulation. The source encoder compresses the input signal by removing possible inherent redundancies, while the channel encoder adds structured redundancy to the compressed bits to correct potential errors caused by the noisy communication channel. For example, in wireless image transmission applications, which will be the main focus of this paper, image compression algorithms such as JPEG or BPG are used to remove redundancies in the input image, and channel coding methods like LDPC or polar codes help to compensate for errors over the channel. Separation theorem states that this two-step procedure is optimal in the asymptotic infinite blocklength regime for ergodic sources and channels [2]. However, the theoretical optimality of the separation principle does not apply to practically relevant finite blocklength regimes as well as in most multi-user setups [3]. In these scenarios, it is well-known that joint source-channel coding (JSCC) methods that integrate source compression and channel coding into a single step can potentially provide better performance, particularly under extreme bandwidth and latency constraints.

Despite potential gains, designing practical JSCC schemes has long been a challenge [4]. Recently, JSCC methods that benefit from modern deep learning (DL) architectures, known as DeepJSCC [5–11] have demonstrated remarkable results for a variety of source modalities and channels. DeepJSCC models the encoding and decoding processes of the communication system as an autoencoder, allowing it to extract complex features from the source data and implicitly incorporate the process of combating bandwidth compression and channel noise into its encoding process, thereby significantly improving performance. Additionally, DeepJSCC avoids the cliff effect and achieves smooth degradation, meaning that even if the channel quality is below the target SNR, images can still be decoded within reasonable reconstruction quality. These characteristics give DeepJSCC significant advantages over traditional separation schemes, particularly in challenging channel modeling and estimation scenarios [12].

Standard DeepJSCC methods typically focus on pixel-level distortion or structural similarity, which fail to account for the semantic similarity between the input and the reconstructed images. It has been recently shown that the perceptual quality of the reconstructed image can be better captured by the learned perceptual image patch similarity (LPIPS) [13] metric, which exhibits remarkable parallels with human perception. This is crucial in semantic communication, especially in image/video transmission contexts, where the receiver might be more concerned with specific tasks like classification or retrieval rather than precise reconstruction of the original signal. In extreme image compression situations, the receiver might be more interested in generating outputs that share similar semantics as the source signal rather than its precise reconstruction.

In this study, we propose a semantic communication scheme targeting extreme channel environments. Specifically, this scheme aims to improve the perceptual quality of existing DeepJSCC methods. We focus on extreme channel settings

Jiakang Chen, Selim F. Yilmaz, Di You, Pier Luigi Dragotti, and Deniz Gündüz are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mails: (jiakang.chen21, s.yilmaz21, di.you22, p.dragotti, d.gunduz)@imperial.ac.uk)

We acknowledge funding from the European Union's Horizon 2020 Marie Skłodowska Curie Innovative Training Network Greenedge (GA. No. 953775), from the UKRI for the projects INFORMED-AI (EP/Y028732/1), and the SNS JU project 6G-GOALS (Grant Agreement No. 101139232).

For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

*These authors contributed equally to this work. †Corresponding author.

because, under very low channel signal-to-noise ratios (SNR) and extremely low bandwidth compression ratios (BCRs), DeepJSCC systems optimized by minimizing pixel-level distortion often experience significant degradation in perceptual quality. To address this issue, we aim to incorporate diffusion models [14], the state-of-the-art generative models, to achieve high perceptual quality reconstructions.

The method we propose does not rely on a specific implementation of the DeepJSCC encoder/decoder and can enhance the perceptual quality of any JSCC system. This is because our method directly acts on the output of the original DeepJSCC decoder, treating the restoration of high perceptual quality from the decoder output as an inverse problem, and employing unconditional generative diffusion models. If we have no prior information on the DeepJSCC architecture, we employ the denoising diffusion null-space model (DDNM) [15], where we approximate the nonlinear degradation process caused by DeepJSCC and the channel as a known linear transformation and construct a pseudo-inverse based on this approximation to guide the pre-trained unconditional diffusion model to restore the degraded output. We call this method SING-Zero as it is a zero-shot scheme oblivious of the specific DeepJSCC architecture. If we have partial information about the DeepJSCC architecture and the underlying channel, we can perform a second stage method, called SING-INN, to further refine the perceptual quality. Inspired by the InDigo architecture in [16], we directly model the nonlinear degradation process using an invertible neural network (INN) combined with the DDNM method to guide the diffusion process. By employing a more precise degradation model, the SING-INN method further enhances the perceptual quality compared to SING-Zero.

We remark that SING does not require any modification of the underlying physical layer encoding and decoding functions, and can be deployed at the application layer at the receiver end. For SING-Zero, all we need is a powerful pretrained diffusion model for the transmitted images, while SING-INN further requires a small dataset of transmitted and reconstructed image pairs for training together with the underlying communication status.

We also note that there is a growing interest in using diffusion models to solve inverse problems. However, applying such solvers in our setting raises three challenges under which existing off-the-shelf diffusion-based methods are not directly applicable. Firstly, unlike many standard image restoration benchmarks, where the degradation operator is often assumed to be known and relatively simple, the forward operator in semantic communication is highly nonlinear, stochastic, and admits no closed-form expression. This is because it encapsulates the entire DeepJSCC pipeline. Secondly, practical deployment demands a *receiver-only* solution that treats the communication pipeline as a black box, precluding the use of methods that require accessing the forward operator. Thirdly, distribution mismatch, where the DeepJSCC system was trained on data statistically different from the images being transmitted, compounds the difficulty by producing heavily out-of-distribution decoder outputs. The two-stage architecture of SING is designed to address these three challenges: SING-Zero provides a zero-shot approach via a linear surrogate, while SING-INN overcomes the linearity assumption by learning the nonlinear degradation from observed input–output pairs, without requiring access to the encoder or decoder architecture.

A. Contributions

Our main contributions can be summarized as follows:

1) We propose a two-stage generative restoration method for DeepJSCC-based wireless image transmission, accommodating scenarios where prior knowledge of the DeepJSCC encoder/decoder functions may or may not be available at the decoder.
2) In the absence of prior information, we present SING-Zero, where we adapt the DDNM framework [15] by approximating the nonlinear degradation induced by DeepJSCC and the noisy communication channel as a linear transformation, constructing a pseudo-inverse to guide an unconditional diffusion model for restoring degraded outputs and improving perceptual quality.
3) When partial information about the DeepJSCC system and channel is available, we propose SING-INN, where we further refine the restoration process using INN-based modeling inspired by InDigo+ [16]. This approach enhances perceptual quality by accurately capturing nonlinear degradations due to communications combined with DDNM guidance.
4) Extensive experiments demonstrate that our method outperforms traditional DeepJSCC and recent generative DeepJSCC approaches under varying SNR and bandwidth conditions, achieving superior generalization without requiring modifications at the transmitter.

The rest of the paper is organized as follows. In Section II, we review related works on JSCC and also provide an overview of diffusion models. In Section III, we define the image transmission problem over noisy channels and introduce key concepts. In Section IV, we describe our novel methodology, including the use of DeepJSCC, linear approximations, and INN-guided diffusion models. In Section V, we present our experimental results, showing the performance improvements of our method. Finally, we conclude in Section VI.

II. Related Work

In recent years, JSCC techniques based on deep learning have been extensively studied [5–11]. We refer the readers to [4] for a recent survey on the topic. In [5], deep neural networks were first used for wireless image transmission in the context of JSCC. This approach is known as DeepJSCC. Here, an autoencoder architecture is employed for wireless communication, where the encoder converts the source image directly into channel symbols similar to analog communication. Correspondingly, the decoder is used to reconstruct the image directly from the channel symbols corrupted by channel noise. Notably, DeepJSCC has been shown to have smooth degradation with channel quality, avoiding the *cliff effect* while outperforming traditional separation-based digital communication techniques. DeepJSCC has then been extended

to incorporate feedback [6, 17] and applied to various communication scenarios such as OFDM [9, 18] and MIMO [19, 20]. In addition to wireless image transmission, DeepJSCC has also been extended to video transmission [8, 21]. However, these widely applicable DeepJSCC methods often focus only on classical pixel-level distortion metrics, and largely neglect perceptual and semantic quality, which is crucial for next-generation semantic communication.

While the aforementioned DeepJSCC schemes are largely driven by reconstruction distortion, recent works have also considered other aspects such as explainability, controllability, and privacy in semantic communication. A trustworthy image semantic communication framework with GenAI is proposed in [22], where images are mapped to human-interpretable semantic tokens and a generative model provides explainability and controllability at the receiver side. Complementarily, task-oriented and privacy-preserving perspective on semantic communication has been studied in [23–25] highlighting how semantic representations can be designed to balance task performance and privacy leakage. These works are complementary to our setting: they focus on *what* semantic information should be transmitted and how to make it explainable and privacy-aware, whereas SING focuses on *how* to restore high-perceptual-quality images from severely degraded DeepJSCC outputs under extreme channel conditions. To tackle this restoration problem, we build on a complementary line of work that leverages powerful neural generative models as priors for JSCC and related inverse problems.

In particular, generative models based on neural networks, such as variational autoencoders [26] (VAEs), generative adversarial networks [27] (GANs), and diffusion models [14] have demonstrated powerful capabilities in generating high perceptual quality data. Within the JSCC literature, these generative models have been incorporated in various approaches. For instance, JSCC schemes have been implemented through VAE [28], adversarial training for secure JSCC in the presence of eavesdroppers [23], GAN-style loss for training decoders [9], and diffusion denoising models to enhance reconstruction quality at the decoder [29]. These works can be viewed as early attempts to inject generative modeling into JSCC, typically by using the generative component as an auxiliary loss or module within an otherwise standard DeepJSCC encoder–decoder. In the next paragraph, we turn to a more closely related line of research that explicitly formulates semantic image communication as an inverse problem driven by a strong generative prior (e.g., InverseJSCC and its extensions), which is also the line that SING builds upon.

In [30], pre-trained generative models are used to address the classical inverse problem in compressed sensing. Similarly, [31] achieves outstanding performance in single-image super-resolution by leveraging generative models, highlighting their effectiveness in solving practical inverse problems. The similarity between communication systems and inverse problems indicates the potential for applying powerful generative models to JSCC, as reflected in recent works such as [23, 24, 32, 33]. Notably, the authors of [32] proposed InverseJSCC, treating the recovery of high-quality source images from degraded reconstructions as an inverse problem. By utilizing a pre-trained StyleGAN-2 generator and the intermediate latent optimization (ILO) method [34], they achieve notable perceptual quality results for image reconstruction under extreme channel conditions. In [35], the approach in [32] is extended to a dynamic generative scenario, where the receiver, rather than relying on a fixed pre-trained generative model, keeps fine-tuning its model using the images it has received in the past. This way the model continuously adapts to the statistics of the images being transmitted.

Diffusion models have made significant advances in the field of visual generation. As a result, some studies have attempted to leverage powerful diffusion models to solve inverse problems, particularly in image restoration tasks. Among these, we are particularly interested in zero-shot methods because they allow us to directly apply diffusion models, trained in an unconditioned manner, to image restoration tasks without the need for supervised tuning of the diffusion model to a specific degradation. For example, [36] guides the diffusion model outputs to restored images by replacing the low-frequency components in the denoised output with those from the reference degraded image. In [37], the authors further extend the solvers of diffusion models through approximate posterior sampling, effectively addressing the problem of general noisy image restoration.

Due to the powerful generative capabilities of diffusion models, they have also been applied in semantic communications. In [29], the authors propose a hybrid JSCC scheme that combines traditional transmission methods with a process of adding noise to input images through diffusion and then transmitting them via DeepJSCC, achieving high-quality reconstruction at the decoder. In [38], the authors apply the noise-removal properties of diffusion models to wireless communication to help the receiver mitigate the impact of channel noise. In [39], the authors consider the application of diffusion models in wireless communication under strict bandwidth constraints. In this approach, the signal transmission process through the channel is treated as the forward process in a diffusion model and a VAE-based upsampling and downsampling pair is used to reduce the bandwidth requirements by leveraging the advanced compression capabilities of VAEs. In [40], the authors propose conditional diffusion using some semantic description of the image (e.g., text or edge map), and employ a semantics-guided channel denoising approach allowing dynamic adaptation to channel conditions. More recently, DiffCom [41] casts the received channel signal as a natural condition for diffusion posterior sampling and jointly optimizes the transmitter and receiver for robust semantic communication.

In [42], the authors propose CommIN, a JSCC scheme that treats image reconstruction as an inverse problem, using INNs and diffusion models to enhance perceptual quality, outperforming DeepJSCC in extreme conditions. Their INN-based degradation modeling effectively separates coarse and detail components, allowing diffusion models to refine reconstructions. Similarly, in [33], the authors introduce a DeepJSCC framework with denoising diffusion models, transmitting lower-resolution images and restoring details at the receiver, achieving superior perceptual quality. By leveraging range–null space decomposition, their method reconstructs high-fidelity images with

notable improvements in perception- and distortion-oriented metrics. These works can be regarded as representative inverse-problem-based semantic communication schemes equipped with generative priors; SING follows this high-level viewpoint but introduces a two-stage design (SING-Zero and SING-INN) tailored to extreme-channel DeepJSCC scenarios and to transmitter-agnostic deployment.

Overall, existing semantic image communication methods can be broadly grouped into three categories. The first category comprises classical DeepJSCC architectures [5, 6, 8, 9, 19, 21], which primarily optimize distortion-oriented metrics and have recently started to incorporate more complex perceptual losses. The second category focuses on trustworthy and task-oriented semantic communication, emphasizing explainability and privacy aspects [22–25]. The third category consists of schemes that couple powerful generative models (VAE, GAN, diffusion, INN) with JSCC in an inverse-problem or generative-JSCC formulation, including StyleGAN-based InverseJSCC and its dynamic extensions [32, 35], diffusion-aided receivers [29, 33], CommIN [42], and joint generative frameworks [40, 41]. Our proposed framework, SING, is complementary to these lines of works: it adopts the inverse-problem and generative-prior viewpoint of the third category, but, unlike joint generative JSCC designs, it is implemented purely at the receiver side as a *transmitter-agnostic* add-on that can be plugged into any pre-trained DeepJSCC (or, more generally JSCC) system (including legacy separation-based communication protocols), and specifically targets the restoration of high perceptual quality under extreme channel conditions. As such, SING is fully compatible with current wireless standards, and can be deployed as an application layer add-on without modifying the underlying communication system.

A. A Brief Review of Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) [14] represent a powerful class of generative models that excel at learning complex data distributions by progressively refining noisy inputs. These models operate through two complementary processes: a forward diffusion process, which gradually corrupts data by adding noise, and a reverse denoising process, which reconstructs the original data by removing the noise in a step-by-step manner.

The forward process in DDPM is designed to transform data into a tractable distribution, typically a Gaussian distribution. This transformation occurs over T discrete time steps. Starting from an initial data sample $\mathbf{x}_0$ from the target distribution, at each time step t, Gaussian noise is incrementally added to produce the noisy state $\mathbf{x}_t$. The process is defined by:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where β_t is a variance schedule that controls the noise level at each time step. Over the sequence of T steps, the data distribution is gradually morphed into a simple Gaussian distribution. We note that, any intermediate state $\mathbf{x}_t$ can be directly sampled from the original data $\mathbf{x}_0$ using the expression:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{2}$$

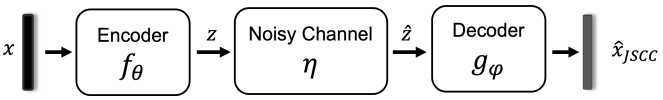


Fig. 1: Overview of JSCC-based communication systems.

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is a Gaussian noise term. This direct sampling formula is central to the forward process as it simplifies the generation of noisy data at any time step.

The reverse process is where the generative power of DDPM is realized. It aims to invert the forward process, starting from Gaussian noise and iteratively denoising it back to the original data distribution. This reverse process is modeled as a learned sequence of denoising steps. The conditional distribution of each reverse step can be computed using Bayes' theorem:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right) + \sigma_t\mathbf{z}, \tag{3}$$

where $\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}$, and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ represents additional Gaussian noise, ensuring that the reverse step maintains the stochasticity necessary for generative diversity.

To predict the noise term $\boldsymbol{\epsilon}$ at each time step, DDPM employs a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$. The training objective of this network is to minimize the difference between the true noise $\boldsymbol{\epsilon}$ and the predicted noise $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$, given by:

$$\nabla_{\boldsymbol{\theta}}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||_2^2, \tag{4}$$

where $\mathbf{x}_0$ represents the original data, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is the noise added during the forward process.

The generative process in DDPM is executed by iteratively applying the reverse denoising steps, starting from an initial noise sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As the model progresses through each time step, it gradually refines the noise into a data sample resembling those in the training distribution. This iterative denoising effectively transforms random noise into a meaningful data sample, thereby showcasing the model's ability to learn and reproduce complex data distributions.

III. Problem Statement

We consider point-to-point wireless transmission of images over a noisy channel (see Fig. 1). As in [5], we model the transmitter and receiver as two neural networks. Specifically, the transmitter is the encoder and the receiver is the decoder of an autoencoder system. The transmitter maps the source signal, a real vector $\mathbf{x} \in \mathbb{R}^m$ to the complex channel input vector $\mathbf{z} \in \mathbb{C}^k$ using a parameterised encoding function $\mathbf{z} = f(\mathbf{x}; \boldsymbol{\theta})$. Here m represents the source bandwidth and k the channel bandwidth. We define the bandwidth compression ratio (BCR) as $\rho \triangleq k/m$ channel uses per source sample. This ratio reflects the compression applied to the signal by the communication system. We expect the performance to improve as more channel bandwidth becomes available, i.e., as ρ increases. In the image transmission problem, $m = H \times W \times C$, where H, W and C are the height, width and, number of colour channels of the input image, respectively. The channel input vector must satisfy the following average power constraint:

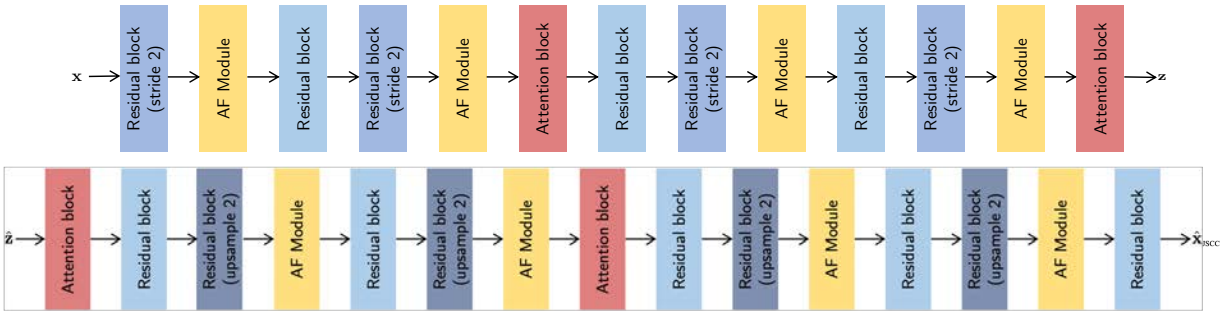


Fig. 2: Encoder architecture of our method (**top**), which is used to encode the input image $\mathbf{x}$ at the transmitter. Decoder architecture of our method (**bottom**), which is used to reconstruct the input image from the noisy channel output $\hat{\mathbf{z}}$.

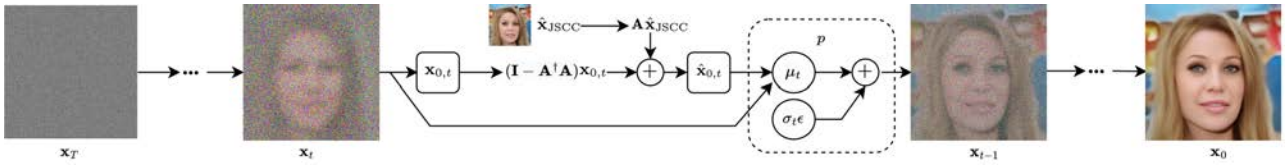


Fig. 3: Overview of the SING-Zero with null-space-based image restoration procedure.

$$\frac{1}{k}\mathbb{E}_{\mathbf{z}}\left[\|\mathbf{z}\|_2^2\right] \leq \bar{P}. \tag{5}$$

This is satisfied with equality by normalising the signal $\tilde{\mathbf{z}}$, output of the last layer of the encoder, as follows:

$$\mathbf{z} = \sqrt{k\bar{P}}\frac{\tilde{\mathbf{z}}}{\sqrt{\tilde{\mathbf{z}}^H\tilde{\mathbf{z}}}}, \tag{6}$$

where H is the Hermitian transpose. The channel introduces random corruptions to the transmitted symbols. We consider a complex AWGN channel, in which the received signal can be expressed as follows:

$$\hat{\mathbf{z}} = \eta\left(\mathbf{z}, \sigma^2\right) = \mathbf{z} + \mathbf{n}_C, \tag{7}$$

where the entries of $\mathbf{n}_C$ are independent identically distributed (i.i.d.) samples from a complex Gaussian distribution with variance σ^2: $\mathbf{n}_C \sim \mathcal{CN}\left(0, \sigma^2\mathbf{I}\right)$. The receiver uses a parameterised decoding function $\hat{\mathbf{x}}_{\text{JSCC}} = g\left(\hat{\mathbf{z}}; \boldsymbol{\phi}\right)$ to get the degraded reconstruction $\hat{\mathbf{x}}_{\text{JSCC}}$ from $\hat{\mathbf{z}}$. To represent the condition of the noisy channel, we define channel SNR as follows:

$$\text{SNR} \triangleq 10\log_{10}\frac{\bar{P}}{\sigma^2}\ \text{dB}. \tag{8}$$

The encoder and decoder are jointly optimised by minimising:

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}\mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{\mathbf{x}}|\mathbf{x}}}[d(\mathbf{x}, \hat{\mathbf{x}}_{\text{JSCC}})], \tag{9}$$

where d can be any distortion measure.

In our simulations, we will employ the same DeepJSCC architecture as in [32]. We note that DeepJSCC models can be optimized directly for any differentiable distortion metric and can be readily adapted for a range of multimedia transmission tasks, such as video transmission as demonstrated in [8]. Following the methodology described in [43], we adopt a composite loss function, $\mathcal{L}$, to simultaneously optimize the *distortion-perception trade-off* [32] for the reconstructed images, formulated as:

$$\mathcal{L}\left(\hat{\mathbf{x}}_{\text{JSCC}}, \mathbf{x}\right) = \text{MSE}(\hat{\mathbf{x}}_{\text{JSCC}}, \mathbf{x}) + \lambda \cdot \text{LPIPS}(\hat{\mathbf{x}}_{\text{JSCC}}, \mathbf{x}), \tag{10}$$

where LPIPS [13] is employed as a perceptual quality metric. The parameter λ regulates the balance between the LPIPS and MSE losses. Designed to assess the similarity of distorted patches, LPIPS calculates the distance in the feature space of a deep neural network (DNN) model initially trained for image classification. In contrast to the MSE distortion metric which operates on a point-wise basis, perception-oriented distortion metrics like LPIPS [13] and MS-SSIM [44] evaluate the similarity between the original and reconstructed images using proxies to human perception.

In the following section, we introduce our JSCC schemes SING-Zero and SING-INN, which leverage diffusion models at the receiver to effectively reduce the semantic distortion between the source signal and its reconstruction, in order to improve the perceptual quality.

IV. Methodology

A. DeepJSCC as an Inverse Problem

We aim to improve the perceptual quality of the images reconstructed by the DeepJSCC encoder-decoder pair by formulating the reconstruction process as an inverse problem. Specifically, we interpret the received degraded signal $\hat{\mathbf{x}}_{\text{JSCC}}$ as degraded measurements $\mathbf{y} = \mathcal{H}(\mathbf{x})$ in a conventional inverse problem, where the forward operator $\mathcal{H}$ encapsulates the entire DeepJSCC communication chain, including the encoder, noisy channel, and decoder. Formally, $\mathcal{H}$ is expressed as:

$$\mathbf{y} = \mathcal{H}(\mathbf{x}, \eta) = g_\phi\left(\eta\left(f_\theta(\mathbf{x}), \sigma^2\right)\right) = \hat{\mathbf{x}}_{\text{JSCC}}, \tag{11}$$

representing a nonlinear stochastic process with AWGN, η. Figure 2 depicts the employed DeepJSCC encoder and decoder architectures, which can be substituted with other DeepJSCC-based or traditional source/channel coding techniques.

With the forward model in place and given measurements of $\mathbf{y}$, one typically seeks to restore the original signal $\mathbf{x}$ by simultaneously imposing a fidelity constraint and a proper prior

on $\mathbf{x}$. The fidelity constraint is usually the MSE between the output of $\mathcal{H}$ and the measurements $\mathbf{y}$:

$$\mathrm{MSE}(\mathcal{H}(\mathbf{x},\eta),\mathbf{y}) = \|\mathcal{H}(\mathbf{x},\eta)-\mathbf{y}\|_2^2. \tag{12}$$

For the prior, we adopt diffusion models, which offer a highly effective way to model the underlying structure of $\mathbf{x}$.

This approach involves two main challenges: (1) the nonlinear nature of the degradation process, which lacks a closed-form expression, and (2) the computational complexity of incorporating a diffusion model as a prior. To tackle these challenges, we propose a two stage scheme depending on the availability of prior knowledge of the coding system and the channel. When no information about the DeepJSCC architecture or communication channel is available, we approximate the degradation process as a fixed linear transformation. When partial information about the DeepJSCC system and channel is available—evidenced by the received degraded images—we directly model the nonlinear degradation process using INNs. This approach allows us to capture and leverage the nonlinear properties of the system more effectively. In both cases, we utilize pre-trained unconditional diffusion models to significantly reduce computational complexity while maintaining reconstruction quality.

A salient feature of both stages is an explicit measurement-consistency mechanism that limits the generative freedom of the diffusion prior. In SING-Zero, the range–null space decomposition constrains the range component of each intermediate estimate to match the received signal; the diffusion model operates exclusively in the complementary null space. In SING-INN, the learned INN further tightens this constraint by capturing nonlinear degradation structure, reducing the residual degrees of freedom that the diffusion model infers. Together, these mechanisms confine the generative process to plausible completions of *unobserved* information rather than unconstrained synthesis, thereby mitigating hallucination risk.

B. SING-Zero: Using a Linear Proxy for the Communication Chain

Following DDNM [15, 33], we use a known, deterministic, linear operator $\mathbf{A}$ as a proxy for the (stochastic, nonlinear) forward process $\mathcal{H}(\mathbf{x},\eta)$. The approximation can be expressed as

$$\hat{\mathbf{x}}_{\mathrm{JSCC}} \approx \mathbf{A}\mathbf{x}. \tag{13}$$

The proxy $\mathbf{A}$ can be any linear operator with a tractable pseudo-inverse, such as decolorization or mean-pooling-based downsampling [15]. In this paper, we use factor-2 mean-pooling because it is simple, training-free, and independent of the DeepJSCC architecture; its numerical comparison with a learned linear alternative validate our choice and is reported in Section V-C.

Given $\mathbf{A}$, we denote with $\mathbf{A}^\dagger$ its pseudo-inverse. Given a received degraded image $\hat{\mathbf{x}}_{\mathrm{JSCC}}$, the image restoration aims to produce $\hat{\mathbf{x}}$ with minimal distortion while enhancing perceptual quality. To achieve this, we impose a consistency constraint $\mathbf{A}\hat{\mathbf{x}} \equiv \hat{\mathbf{x}}_{\mathrm{JSCC}}$ and ensure the perceptual quality by requiring $\hat{\mathbf{x}}$ to follow the distribution of ground truth images. We employ the zero-shot image restoration method DDNM to achieve that.

Algorithm 1 SING-Zero: Null Space Sampling

Require: The received degraded image $\hat{\mathbf{x}}_{\mathrm{JSCC}}$, degradation operator $\mathbf{A}$, and its pseudo-inverse $\mathbf{A}^\dagger$.

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
for $t = T,\dots,1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ if $t>1$, else $\mathbf{z}=\mathbf{0}$
  $\mathbf{x}_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t))$
  $\hat{\mathbf{x}}_{0,t} = \mathbf{x}_{0,t} - \mathbf{A}^\dagger\left(\mathbf{A}\mathbf{x}_{0,t} - \hat{\mathbf{x}}_{\mathrm{JSCC}}\right)$
  $\mathbf{x}_{t-1} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_{0,t} + \sigma_t\mathbf{z}$
end for
return $\mathbf{x}_0$

We use the neural network in DDPM, denoted by $\boldsymbol{\epsilon}_\theta$, trained via a progressive denoising task with T-step forward and backward processes. At timestep t, the noise in $\mathbf{x}_t$ is predicted by the neural network $\boldsymbol{\epsilon}_\theta$ as $\epsilon_t = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)$. DDNM uses and guides the pretrained DDPM $\boldsymbol{\epsilon}_\theta$ to improve image quality and consistency without additional training [15].

Specifically, the sample $\mathbf{x}$ is decomposed using the range-null space decomposition: $\mathbf{x} \equiv \mathbf{A}^\dagger\mathbf{A}\mathbf{x} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}$. Here, $\mathbf{A}^\dagger\mathbf{A}\mathbf{x}$ is the range component and $(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}$ is the null space component. Let $\mathbf{x}_{0,t}$ denote the estimated $\mathbf{x}_0$ at diffusion timestep t. For a received degraded image $\hat{\mathbf{x}}_{\mathrm{JSCC}}$, we form a solution for $\hat{\mathbf{x}}$ that fulfills the consistency constraint: $\hat{\mathbf{x}} = \mathbf{A}^\dagger\hat{\mathbf{x}}_{\mathrm{JSCC}} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{x}_{0,t}$, where $\mathbf{x}_{0,t}$ is refined iteratively using the method shown in Algorithm 1. The hyperparameter $\bar{\alpha}$ defines the noise schedule. This process ensures that modifying $\mathbf{x}_{0,t}$ does not break the consistency constraint, allowing us to enhance perceptual quality while preserving the transmitted image's consistency with DeepJSCC decoder's output. Finally, $\hat{\mathbf{x}}$ is unflattened back to its original dimensions of $H \times W \times C$.

C. SING-INN: Modelling the Forward Process using INNs

Although the DDNM-based SING-Zero is effective even without any system knowledge, its linear proxy does not fully capture the strongly nonlinear DeepJSCC encoder/decoder and can limit the perceptual gains at extreme BCR/SNR values. For this reason, inspired by our recent work CommIN [42], we improve our approach using INNs and generative diffusion models. The key insight here is that INN splits the signal (typically an image) into a coarse version $\mathbf{c}$ and details $\mathbf{d}$. In [16], the forward part of the INN is trained to mimic the degradation process in an inverse problem so that $\mathbf{c}$ is very close to the degraded image. In this paper, instead, we treat the communication process composed of the transmission, corruption with channel noise and reconstruction as the degradation process and mimic it using an INN (see Fig. 5).

Specifically, building upon the CommIN framework [42], we introduce a conditional INN, where the channel SNR serves as the conditioning input. However, contrary to CommIN, this design allows a single INN to model the degradation process across various SNRs in a unified manner, thereby eliminating the need to train separate models for each channel condition. Consequently, our method extends the advantages of CommIN—effective modeling of the communication degradation process—while significantly enhancing its usability and efficiency in real-world scenarios.

Algorithm 2 SING-INN: Null Space Sampling with INN

Require: The received degraded image $\hat{\mathbf{x}}_{\text{JSCC}}$, the linear operator $\mathbf{A}$ approximating the degradation, and its pseudo-inverse $\mathbf{A}^{\dagger}$, the pre-trained INN and the actual channel SNR.

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{0,t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t))$
5: $\quad \hat{\mathbf{x}}_{0,t} = \mathbf{x}_{0,t} - \mathbf{A}^{\dagger}\left(\mathbf{A}\mathbf{x}_{0,t} - \hat{\mathbf{x}}_{\text{JSCC}}\right)$
6: $\quad \hat{\mathbf{x}}_{t-1} = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_{0,t} + \sigma_t \mathbf{z}$
7: $\quad \mathbf{c}_t, \mathbf{d}_t = \mathit{INN}_{Forward}(\hat{\mathbf{x}}_{0,t}, \text{SNR})$
8: $\quad \tilde{\mathbf{x}}_{0,t} = \mathit{INN}_{Inverse}(\hat{\mathbf{x}}_{\text{JSCC}}, \mathbf{d}_t, \text{SNR})$
9: $\quad \mathbf{x}_{t-1} = \hat{\mathbf{x}}_{t-1} - \zeta \nabla_{\mathbf{x}_t} \|\tilde{\mathbf{x}}_{0,t} - \hat{\mathbf{x}}_{0,t}\|_2^2$
10: **end for**
11: **return** $\mathbf{x}_0$

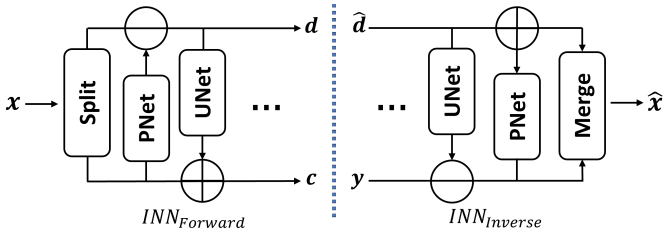


Fig. 4: General architecture of conditional INN in this work.

As illustrated in Fig.5, we train a conditional INN, where the channel SNR is used as a conditioning input, to simulate the degradation process. Under a given channel SNR, the INN decomposes the original signal $\mathbf{x}$ into the estimated degraded observation $\hat{\mathbf{x}}'_{\text{JSCC}}$ ($\approx \hat{\mathbf{x}}_{\text{JSCC}}$) and the lost details $\mathbf{d}$. Due to the invertibility of the INN, under the same channel SNR, it is possible to perfectly recover the original image $\mathbf{x}$ if one has access to both $\hat{\mathbf{x}}'_{\text{JSCC}}$ and $\mathbf{d}$. By incorporating the SNR as a condition, a single INN can be used to handle varying channel conditions in a unified manner, thereby simplifying the training and inference processes. During inference, the pre-trained conditional INN guides the diffusion process, ensuring integration of observed measurements $\hat{\mathbf{x}}_{\text{JSCC}}$ and preservation of intricate details.

The basic architecture of our conditional INN is shown in Fig. 4. The original image is split into two parts by a splitting operator. Then, the Prediction Network (PNet), conditioned on the coarse part and channel SNR, aims to predict the detail part. Meanwhile, the Update Network (UNet), conditioned on the detail part and channel SNR, is used to adjust the coarse part to make it smoother. The PNet and UNet are applied alternatively to generate the coarse and detail parts, $\mathbf{c}$ and $\mathbf{d}$, respectively. Notably, PNet and UNet can be any complex nonlinear functions without compromising the invertibility of the INN. In our implementation, each PNet/UNet is composed of j CondBlocks, each of which can be described as

$$\alpha = \text{FC}(\text{SNR}), \tag{14}$$

$$x_{j+1} = \text{Conv}_2\Big(\text{ReLU}\big(\text{Conv}_1(x_j)\big)\Big) \odot \alpha + x_j, \tag{15}$$

where $\text{FC}(\cdot)$ denotes a fully connected layer, and $\odot$ indicates channel-wise multiplication.

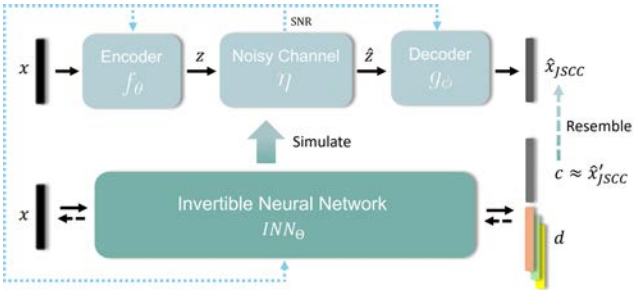


Fig. 5: The key insight of SING-INN is to simulate the degradation process of the communication pipeline (top) with a conditional INN (bottom), which can transform the input original signal $\mathbf{x}$ into the simulated degraded observation $\hat{\mathbf{x}}'_{\text{JSCC}}$ and estimated lost details $\mathbf{d}$ under a certain channel SNR. At testing stage, $\mathbf{d}$ is estimated using a diffusion model.

To model the degradation process, we impose that $\mathbf{c}$ resembles $\hat{\mathbf{x}}_{\text{JSCC}}$. Given a training set $\left\{\mathbf{x}^i, \hat{\mathbf{x}}^i_{\text{JSCC}}\right\}_{i=1}^N$, which contains N high-quality input images and their noisy counterparts recovered by DeepJSCC, we optimize our INN with the following loss function:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{c}^i - \hat{\mathbf{x}}^i_{\text{JSCC}}\right\|_2^2, \tag{16}$$

where Θ denotes the learnable parameters of our INN. Once we constrain one part of the output of $\mathit{INN}_{Forward}(\mathbf{x})$ to be close to $\hat{\mathbf{x}}_{\text{JSCC}}$, due to invertibility, the other part of the output, $\mathbf{d}$, will inevitably represent the detailed information lost during the degradation process. We note that to train our INN model, we do not need to have access to DeepJSCC architecture, but only to a set of original and degraded image pairs.

In Algorithm 2, we show how DDNM and INN are combined in SING-INN to improve the quality of the received image $\hat{\mathbf{x}}_{\text{JSCC}}$. Specifically, given $\hat{\mathbf{x}}_{\text{JSCC}}$ as defined in (11), we first apply the main steps of DDNM in line 4-6, then, given $\hat{\mathbf{x}}_{0,t}$ which is a preliminary estimate consistent with $\hat{\mathbf{x}}_{\text{JSCC}}$ under the same linear degradation model of Algorithm 1, we improve the consistency by using INN in steps 7-9 (in blue). In these additional steps, given a channel SNR, we decompose the intermediate result $\hat{\mathbf{x}}_{0,t}$ with $\mathit{INN}_{Forward}$ into coarse $\mathbf{c}_t$ and detail part $\mathbf{d}_t$, and replace the coarse part $\mathbf{c}_t$ with the received degraded image $\hat{\mathbf{x}}_{\text{JSCC}}$. The INN-optimized $\tilde{\mathbf{x}}_{0,t}$ is generated by the inverse process $\mathit{INN}_{Inverse}$. Thus, the INN-optimized $\tilde{\mathbf{x}}_{0,t}$ is composed of the coarse information $\hat{\mathbf{x}}_{\text{JSCC}}$ and the details generated by the diffusion process. To incorporate the INN-optimized $\tilde{\mathbf{x}}_{0,t}$ into the DDNM algorithm, we update $\mathbf{x}_t$ with the guidance of the gradient of $\|\tilde{\mathbf{x}}_{0,t} - \hat{\mathbf{x}}_{0,t}\|_2^2$ as shown in step 9. We note that the operation $\mathit{INN}_{Inverse}(\hat{\mathbf{x}}_{\text{JSCC}}, 0, \text{SNR})$ is the nonlinear counterpart of the pseudo-inverse $\mathbf{A}^{\dagger}$. Therefore, step 8 in Algorithm 2 is the nonlinear equivalent of line 5 in Algorithm 1. With the help of INN, our algorithm effectively estimates the details lost in the nonlinear degradation of the communication process and does not require the knowledge of the closed-form expression of the degradation model.

D. Detailed Comparison with InverseJSCC

We provide a systematic comparison with InverseJSCC [32] to clarify the advantages of our approach. Regarding the modeling assumptions, InverseJSCC relies on a StyleGAN prior, which assumes that all real images lie perfectly on the GAN's learned manifold. StyleGAN-based inversion projects the reconstruction onto a low-dimensional learned manifold; and thus, any image outside this manifold cannot be exactly represented, which limits reconstruction quality. Our method avoids this by leveraging a diffusion model, which defines a full-dimensional implicit prior over the pixel space, providing higher expressive capacity and higher sample fidelity [45]. Concerning adaptability and reconstruction mechanism, InverseJSCC's flexibility comes from its test-time optimization, which can incorporate any differentiable degradation $\mathcal{H}(\cdot)$. However, this reliance on gradient optimization within the latent space provides no explicit mechanism to recover lost information. Our method, in contrast, employs a pre-trained INN. The INN not only models complex and nonlinear degradations, but, more crucially, is designed to explicitly estimate the 'lost details' (i.e., the null-space information) during its forward pass and use them for a principled, invertible reconstruction. This also leads to a profound difference in inference complexity. InverseJSCC requires sample-specific, complex optimization at test time. This process involves backpropagating gradients from the output sample through the entire forward operator and multiple GAN layers, which is computationally demanding. Our method, while also iterative (due to diffusion sampling), involves a feed-forward inference process at each step. Its guidance mechanism requires only a simple gradient step, avoiding the deep and complex backpropagation loop inherent to InverseJSCC, thus leading to a more efficient and robust pipeline. These design differences are reflected in the quantitative results in Section V-F and the complexity comparison in Table V, where SING achieves a favorable PSNR–LPIPS trade-off under distribution shift and substantially lower inference time than InverseJSCC.

E. Distinction from Joint Generative JSCC Frameworks

It is also important to clarify the conceptual distinction between our proposed SING framework and other promising generative-based frameworks such as GenerativeJSCC [32], SGD-JSCC [40] and DiffCom [41]. Methods in the latter category typically propose a joint redesign of the entire transmitter-receiver pipeline. For instance, GenerativeJSCC trains a specific, generative-aware encoder, while SGD-JSCC and DiffCom explicitly require a differentiable channel model and access to the encoder's structure to perform posterior sampling reconstruction. This 'transmitter-aware' paradigm assumes full control over the end-to-end communication system.

In contrast, SING is designed as an encoder/decoder-agnostic, receiver-only add-on. It operates by treating the output of any underlying communication system—even a frozen, legacy (separation-based), or black-box one—as a fixed measurement to be restored. SING does not require any modification to the original communication system or access to its internal parameters. It can therefore be deployed without modifying an existing legacy or standard-compliant communication pipeline.

Because SING addresses a different and more constrained problem (i.e., post-processing a legacy signal at the receiver) than frameworks that jointly optimize the entire pipeline, a direct quantitative comparison would conflate two distinct problem settings. Our goal is to demonstrate improvement over a given baseline, not to compete with fully redesigned systems.

V. Experimental Results

To evaluate the performance of our proposed approach, we compare our two-stage method SING-Zero and SING-INN with the state-of-the-art JSCC methods: DeepJSCC [8] and InverseJSCC [32]. For fair comparisons, we use the same pre-trained DeepJSCC for all three approaches. We utilize the 512×512 CelebA-HQ dataset [46] following [32], comprising $30\,000$ high-resolution celebrity images, divided into training, validation, and testing sets with a ratio of $8:1:1$.

A. Implementation Details

We apply standard hyperparameters for our method that are commonly used in the literature to train DeepJSCC [5, 7]. Specifically, we use a learning rate of 1×10^{-4}, a batch size of 32, and an average power constraint of $\bar{P} = 1.0$ (refer to Eq. (5)). For all the methods analyzed, PyTorch Adam optimizer [47] is utilized to minimize the training loss described in Eq. (4). During training, the network is subjected to channel noise $\mathbf{n}$ determined by a uniformly sampled SNR ranging from -5 to 5 dB. Our INN is composed of 4 pairs of PNet/UNet, and each PNet/UNet consists of $J = 2$ CondBlocks. For training INN, we use a learning rate of 5×10^{-5} and a batch size of 32 on CelebA-HQ training dataset. Moreover, Table I summarizes the training and testing datasets used in each step, while Table II outlines the data preparation pipeline for clarity.

TABLE I: Training and testing dataset used for each step.

		Diffusion	**DeepJSCC**	**INN**
Experiments in Figs. 6 to 10	Train	FFHQ	CelebA	CelebA
	Validate/Test	FFHQ	CelebA	CelebA
Experiments in Figs. 11 and 12	Train	FFHQ	TinyImageNet	CelebA
	Validate/Test	FFHQ	TinyImageNet	CelebA

We assume perfect channel knowledge at the encoder and the decoder. We experiment on channel SNRs of $\{-5, -3, -1, 1, 3, 5\}$ dB and BCR values $\rho = \{0.0013, 0.0052\}$, which correspond to highly challenging communication scenarios. Empirically, we set $T = 1000$ and step size ζ to 0.3, 0.4, and 0.5 for SNR = $\{-5, -3\}$ dB, $\{-1, 1\}$ dB, and $\{3, 5\}$ dB respectively. We analyze the performance of the proposed scheme using the pixel-wise metric PSNR, the perceptual similarity metric LPIPS [13], as well as the distributional quality metric FID [48].

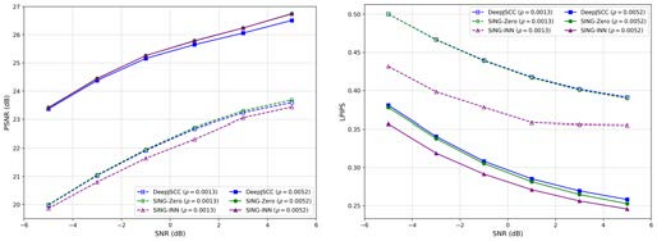
B. Results and Analysis

Fig. 6 presents the performance comparison of DeepJSCC, trained and tested on the CelebA-HQ dataset, with the proposed

TABLE II: Data preparation pipeline.

Step	Description
1	Split the clean images into train/validate/test datasets without any overlap.
2	Using the dataset listed under **DeepJSCC** in Table I, we train DeepJSCC on the training split only and subsequently **freeze** the encoder–decoder.
3	Using the dataset listed under **INN** in Table I, we construct pairs for every split (train/validation/test): for each clean image x_i, a degraded counterpart x_i^{JSCC} is generated by passing x_i through the frozen DeepJSCC pipeline obtained from **Step 2** with an AWGN channel at a randomly sampled SNR.
4	Train the INN on the pairs $\{x_i, x_i^{\mathrm{JSCC}}\}$ constructed in **Step 3**. Validation and test pairs are generated from the corresponding splits using the same approach.

two-stage methods SING-Zero and SING-INN. The results are reported in terms of PSNR and LPIPS as functions of channel SNR (dB) for $\rho = 0.0052$ and the more constrained bandwidth of $\rho = 0.0013$. In both cases, higher PSNR indicates better image quality, whereas lower LPIPS reflects superior perceptual similarity between the ground truth and restored images.



(a) PSNR versus SNR (higher better) (b) LPIPS versus SNR (lower better)

Fig. 6: Performance comparison in terms of PSNR and LPIPS for $\rho = 0.0013$ and 0.0052.

For a BCR of $\rho = 0.0052$, both SING-Zero and SING-INN slightly improve PSNR. While SING-Zero achieves a modest enhancement in LPIPS, SING-INN further boosts perceptual quality by explicitly modeling nonlinear degradations. Under more extreme degradation conditions, such as $\rho = 0.0013$, the fully zero-shot SING-Zero struggles to improve LPIPS due to the difficulty of approximating severe nonlinear degradations as linear transformations. In contrast, SING-INN effectively models nonlinear degradations, significantly improving LPIPS compared to DeepJSCC yielding visibly superior perceptual quality. Fig. 7 provides visual comparisons of reconstructed images alongside their corresponding original and DeepJSCC outputs at $\rho = 0.0013$ for SNR=-1 dB. These results demonstrate that our methods effectively restore perceptual details lost under extreme channel conditions in DeepJSCC. FID measurements in Table III confirm the same trend, with the strongest gains at the tightest bandwidth ($\rho = 0.0013$) and lowest SNRs.

C. Analysis of the Degradation Operator

To justify our choice of the degradation operator $\mathbf{A}$ in Section IV-B, we compared the mean-pooling downsampling by a factor of 2 used in SING-Zero with a learned alternative.

TABLE III: FID (lower better) for $\rho = 0.0013, 0.0052$ and different SNR values.

ρ	SNR	DeepJSCC	SING-Zero	SING-INN
0.0013	-5	123.46	123.64	109.78
	-3	116.01	116.06	104.65
	-1	109.04	109.06	98.53
	1	102.58	102.58	93.19
	3	97.29	97.19	91.16
	5	91.97	91.90	87.18
0.0052	-5	95.36	95.12	90.73
	-3	90.22	90.01	84.37
	-1	83.94	83.67	80.06
	1	80.10	80.03	78.04
	3	77.09	77.00	75.31
	5	74.54	74.50	73.23

(a) Ground Truth

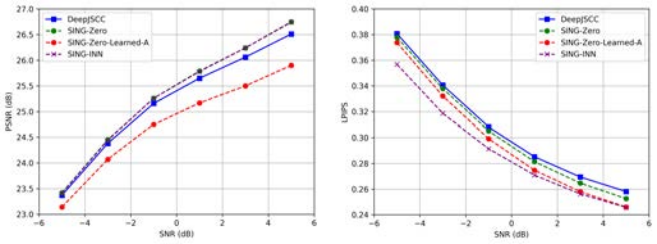
(b) DeepJSCC at SNR=-1 dB

(c) SING-Zero at SNR=-1 dB

(d) SING-INN at SNR=-1 dB

Fig. 7: Original and reconstructed CelebA-HQ images by DeepJSCC and our proposed SING-Zero and SING-INN for $\rho = 0.0013$ and SNR=-1 dB.

Using the same dataset as for INN training (the set constructed in Step 3 of Table II), we attempt to learn a data-driven linear transform $\mathbf{A}$. Directly parameterizing $\mathbf{A}$ as a dense matrix at image resolution is impractical due to its size, so instead we model realistic image downsampling by training a 3×3 convolutional kernel followed by downsampling by a factor of 2. As shown in Fig. 8, the learned operator yields a small improvement in LPIPS over mean-pooling. However, it significantly reduces PSNR, and more importantly, ties $\mathbf{A}$ to the training distribution, thereby eroding the zero-shot nature of SING-Zero. Thus, the learned alternative does not provide a significant improvement over mean-pooling. We therefore adopt



(a) PSNR versus SNR (higher better) (b) LPIPS versus SNR (lower better)

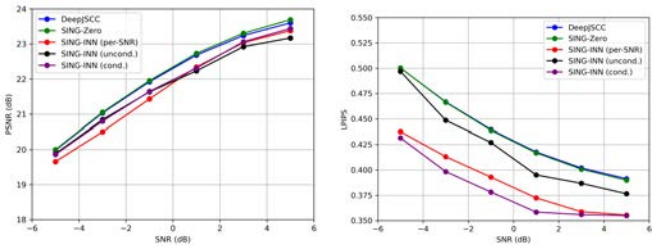
Fig. 8: Performance comparison of mean-pooling and learned degradation operators in terms of PSNR and LPIPS at $\rho = 0.0052$.

mean-pooling downsampling by a factor of 2 for **A**, which offers competitive performance while preserving simplicity and generality.

D. Analysis of INN Design and SNR Generalization

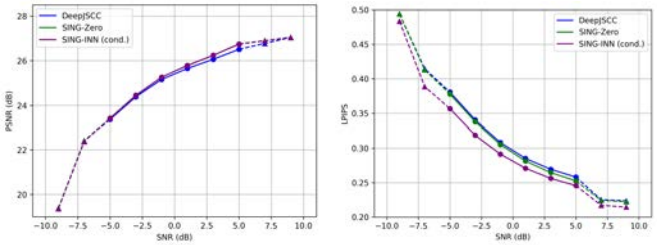
TABLE IV: Ablation studies on different INN architectures.

SING-INN (cond.)	A single conditional INN shared across all SNRs (ours)
SING-INN (uncond.)	A single unconditional INN trained jointly on the full SNR range
SING-INN (per-SNR)	Separate INNs, each trained on a single SNR



(a) PSNR versus SNR (higher better) (b) LPIPS versus SNR (lower better)

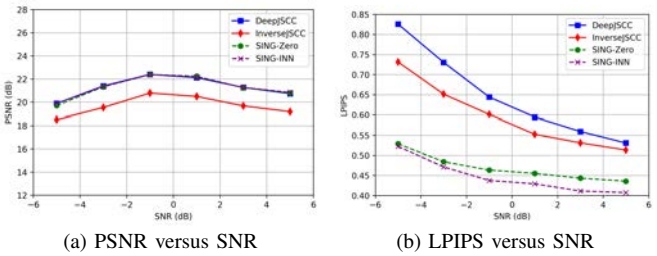
Fig. 9: Performance comparison among different INN architectures in terms of PSNR and LPIPS for $\rho = 0.0013$.



(a) PSNR versus SNR (higher better) (b) LPIPS versus SNR (lower better)

Fig. 10: Performance comparison of INNs for SNR generalization (trained on [-5, 5] dB, tested on (-10, 10) dB) in terms of PSNR and LPIPS for $\rho = 0.0052$.

To disentangle the contribution of the INN design and to verify adaptability to varying channel conditions, we add ablation studies over three INN configurations summarized in Table IV, and report their performances in Fig. 9. The single conditional INN consistently provides the best accuracy with



(a) PSNR versus SNR (higher better) (b) LPIPS versus SNR (lower better)

Fig. 11: Performance comparison when DeepJSCC is trained on TinyImageNet in terms of PSNR and LPIPS for $\rho = 0.0052$.

only one model; training a single unconditional INN over the full SNR range underfits and trails the conditional variant, while the per-SNR setting attains competitive in-range accuracy but requires one model per SNR and thus scales poorly. Beyond in-range behavior, we stress-test SNR generalization by training on $[-5, 5]$ dB and evaluating across $(-10, 10)$ dB (Fig. 10). The conditional INN exhibits smooth, monotonic trends outside the training interval with no catastrophic collapse; performance degrades gracefully as the test SNR moves further from the seen range.

E. Complexity and Latency

To quantify the computational overhead of our two-stage framework, we report in Table V the parameter counts, GPU memory usage, and per-image inference time (RTX 3090, 512×512 resolution) for all methods. As expected, DeepJSCC is the most efficient model, requiring only 3.12 GB of memory and 0.07 s per image. InverseJSCC is dominated by its test-time optimization and becomes the slowest method, with 15.76 GB memory and 137.76 s per image. Our SING variants fall between these two extremes. SING-Zero adds a moderate cost (4.42 GB, 4.75 s) compared to DeepJSCC, reflecting the diffusion-based prior. SING-INN increases latency further (15.75 GB, 23.03 s) due to the additional INN-based refinement, but it is still more than one order of magnitude faster than InverseJSCC while achieving significantly better reconstruction quality and robustness under extreme channel conditions.

TABLE V: Comparison of Model Efficiency

Method	GPU Mem. (GB)	Infer. Time (s)	Params (M)
DeepJSCC	3.12	0.07	144.29
InverseJSCC	15.76	137.76	174.66†
SING-Zero	4.42	4.75	159.79
SING-INN	15.75	23.03	160.11‡

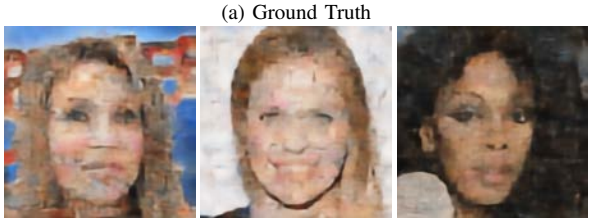
† 174.66 = 30.37 (StyleGAN2) + 144.29 (DeepJSCC).
‡ 160.11 = 159.79 (DDPM) + 0.32 (INN).

F. Analysis of Robustness to Distribution Shift.

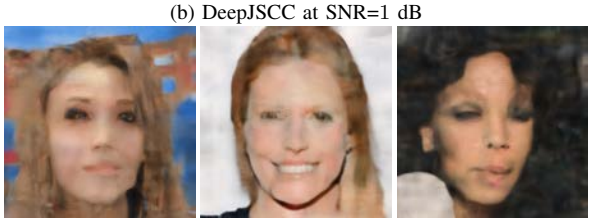
We analyze robustness under the challenging distribution shift setting where DeepJSCC is trained on TinyImageNet (low-resolution natural scenes) and tested on CelebA-HQ (high-resolution human faces), revealing critical differences in their underlying mechanisms.

Specifically, we trained the baseline DeepJSCC model on TinyImageNet, a dataset with only 100,000 natural images (less than 1% of the 14,000,000 images in standard ImageNet) and a resolution of just 64x64. We then tested this model on the same

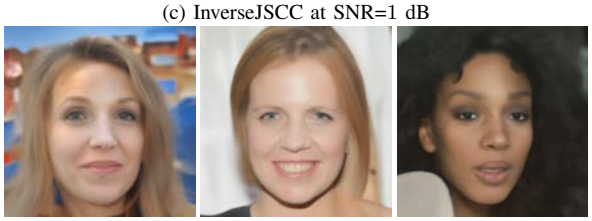
(a) Ground Truth



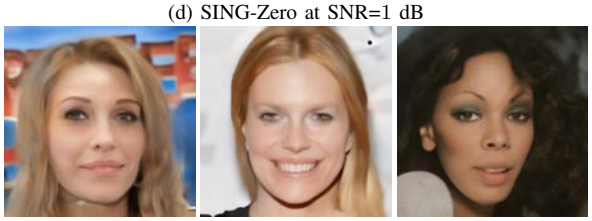
(b) DeepJSCC at SNR=1 dB



(c) InverseJSCC at SNR=1 dB



(d) SING-Zero at SNR=1 dB



(e) SING-INN at SNR=1 dB

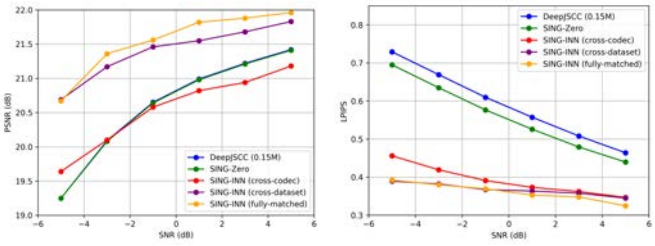
Fig. 12: Original and reconstructed CelebA-HQ images using DeepJSCC trained on TinyImageNet, along with results from InverseJSCC, SING-Zero, and SING-INN under $\rho = 0.0052$ and SNR=1 dB.

high-resolution 512x512 CelebA-HQ dataset used in previous experiments. This substantial distributional mismatch between training and test data significantly amplifies the difficulty of this inverse problem.

We compare DeepJSCC, StyleGAN-2-based InverseJSCC, and our proposed SING-Zero and SING-INN methods under this new setting where DeepJSCC was trained on TinyImageNet and tested on CelebA-HQ, as shown in Fig. 11. For fairness, the generative priors—StyleGAN-2 and DDPM—were both trained on the FFHQ dataset. We can observe that while InverseJSCC achieves better LPIPS than DeepJSCC across all channel SNR values, it sacrifices PSNR performance. Our fully zero-shot SING-Zero method significantly outperforms InverseJSCC in LPIPS across all SNR levels while maintaining comparable PSNR performance. The SING-INN approach further enhances perceptual quality, especially at lower SNR values, where improvements in LPIPS over InverseJSCC become increasingly pronounced. Fig. 12 provides visual comparisons of reconstructed images when DeepJSCC is trained on TinyImageNet alongside their corresponding original and DeepJSCC outputs for SNR=1 dB at $\rho = 0.0052$.

Traditional DeepJSCC struggles in this setting because its encoder learns feature extractors specific to the training distribution; when presented with semantically very different images (e.g., faces), it produces latent vectors that lie outside the support of the decoder's training distribution. The decoder cannot invert these vectors, resulting in severe artifacts. InverseJSCC only partially remedies this. Its GAN prior forces the output onto a face-specific manifold. However, the initial reconstruction it receives is already corrupted with out-of-distribution artifacts. The optimization must then resolve this conflict, often discarding valid fine-grained details that do not conform to the manifold, resulting in PSNR loss and over-smoothed textures. Our methods, SING-Zero and SING-INN, remain robust due to a principled separation between measurement consistency and the generative prior. First, our unconditional DDPM prior models the full pixel distribution, not a low-dimensional manifold, giving it the capacity to adapt better to the new image domain. Second, and most critically, our reconstruction mechanism (via INN or range-null space projection) strictly enforces the received measurements. The diffusion prior is applied only to the complementary, unobserved null space, allowing it to plausibly fill in lost information without ever overwriting the valid data that survived. This separation allows SING-Zero and SING-INN to preserve both perceptual and distributional quality even under this severe distribution shift.

G. Analysis of Cross-Codec and Cross-Dataset Generalization



(a) PSNR versus SNR (higher better)

(b) LPIPS versus SNR (lower better)

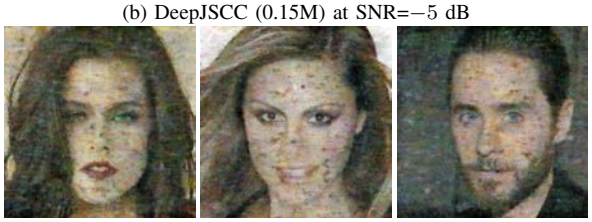
Fig. 13: Performance comparison for cross-codec and cross-dataset generalization of SING-INN in terms of PSNR and LPIPS.

So far, the INN has been trained and tested under matched conditions: the same codec generates both the training pairs and the test-time degraded images. In practice, however, a pre-trained INN may encounter a codec or image domain unseen during training. We therefore evaluate SING-INN under systematic mismatches between its training and deployment conditions.

At test time, images from CelebA-HQ (512×512) are transmitted through the lightweight original version of DeepJSCC [5] ($\sim$0.15 M parameters, trained on CIFAR-10 at 32×32, BCR $= 1/12$, without explicit channel-SNR knowledge at the encoder/decoder). This codec differs from the 144 M-parameter DeepJSCC used in Sections V-A–V-F in architecture scale

(a) Ground Truth

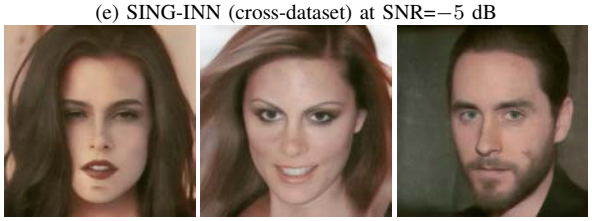
(b) DeepJSCC (0.15M) at SNR=−5 dB



(c) SING-Zero at SNR=−5 dB

(d) SING-INN (cross-codec) at SNR=−5 dB

(e) SING-INN (cross-dataset) at SNR=−5 dB



(f) SING-INN (fully-matched) at SNR=−5 dB

Fig. 14: Original and reconstructed CelebA-HQ images using DeepJSCC (0.15M) trained on CIFAR10, along with results from SING-Zero, SING-INN (cross-codec), SING-INN (cross-dataset) and SING-INN (fully-matched) under $\rho = 1/12$ and SNR=−5 dB.

(0.15M vs. 144M parameters), training data (CIFAR-10 vs. CelebA-HQ), input resolution (32×32 vs. 512×512), and operating regime (standard vs. extreme BCR). We compare three INN training configurations in Table VI.

The full results across all SNR values are shown in Figure 13, while visual comparisons of the reconstructed images under this setting at $\mathrm{SNR} = -5\,\mathrm{dB}$ are provided in Figure 14. These results establish a clear performance hierarchy:

$$\text{Fully matched} > \text{Cross-dataset} > \text{Cross-codec} > \text{SING-Zero},$$

TABLE VI: INN training configurations for the generalization study. In all three cases, the test images are CelebA-HQ (512×512) transmitted through the 0.15M-parameter DeepJSCC (trained on CIFAR-10, BCR $= 1/12$).

Setting	INN **training pairs**
Cross-codec	CelebA-HQ images degraded by our 144M-parameter DeepJSCC (BCR $= 0.0052$)
Cross-dataset	CIFAR-10 images degraded by the 0.15M-parameter DeepJSCC (BCR $= 1/12$)
Fully matched	CelebA-HQ images degraded by the 0.15M-parameter DeepJSCC (BCR $= 1/12$)

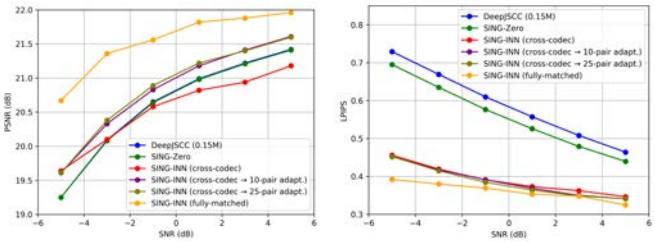
which provides practitioners with a practical roadmap for deploying SING-INN depending on what training data is available. The key insights are:

(i) Generalization across codecs and datasets. Even when the INN is trained on pairs from a fundamentally different codec (cross-codec setting), substantial perceptual gains are retained (e.g., LPIPS improves from 0.6098 to 0.3908 at $\mathrm{SNR} = -1\,\mathrm{dB}$). Likewise, the cross-dataset INN trained on CIFAR-10 pairs (32×32 natural images) transfers effectively to CelebA-HQ (512×512 face images). These results indicate that the INN captures general structural patterns of communication-induced degradation rather than memorizing codec-specific or content-specific artifacts.

(ii) Codec matching is more beneficial than dataset matching. The cross-dataset INN (same codec, different image domain) consistently outperforms the cross-codec INN (same domain, different codec). From a deployment perspective, this implies that generating calibration pairs from the deployed codec, even with generic, readily available images, is both simpler and more effective than reusing an INN trained on in-domain images with a mismatched codec.

(iii) Best performance under matched conditions. When both the codec and image domain match, SING-INN achieves peak performance, confirming its ability to fully exploit matched training data when available.

H. Analysis of Few-Shot Adaptation



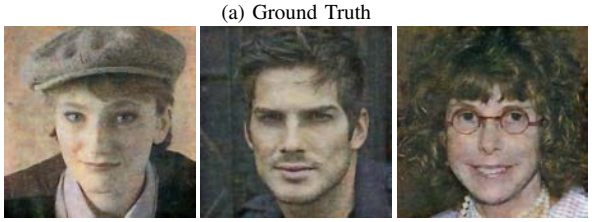
(a) PSNR versus SNR (higher better)

(b) LPIPS versus SNR (lower better)

Fig. 15: Performance comparison for few-shot adaptation of the cross-codec SING-INN using $N \in \{10, 25\}$ target-domain pairs in terms of PSNR and LPIPS.

The performance hierarchy in Section V-G naturally suggests that a pre-trained cross-codec INN can serve as an effective initialisation for rapid adaptation to a new deployment scenario. We validate this by fine-tuning the cross-codec INN using only $N \in \{10, 25\}$ target-domain input–output pairs from the deployed system (i.e., CelebA-HQ images transmitted

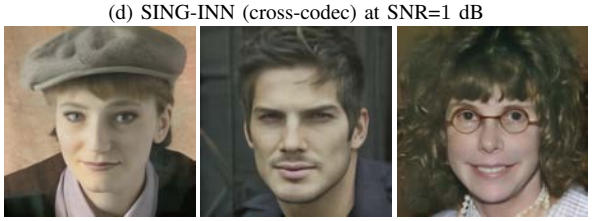
(a) Ground Truth



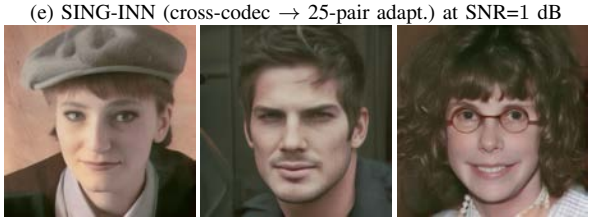
(b) DeepJSCC (0.15M) at SNR=1 dB

(c) SING-Zero at SNR=1 dB

(d) SING-INN (cross-codec) at SNR=1 dB



(e) SING-INN (cross-codec → 25-pair adapt.) at SNR=1 dB



(f) SING-INN (fully-matched) at SNR=1 dB

Fig. 16: Original and reconstructed CelebA-HQ images using DeepJSCC (0.15M) trained on CIFAR10, along with results from SING-Zero, SING-INN (cross-codec), SING-INN (cross-codec → 25-pair adapt.) and SING-INN (fully-matched) under $\rho = 1/12$ and SNR=1 dB.

through the 0.15M-parameter DeepJSCC trained on CIFAR-10 at BCR $= 1/12$). The results are shown in Figure 15, confirming that even minimal fine-tuning yields meaningful improvements, while visual comparisons of the reconstructed images under the same setting at $\mathrm{SNR} = 1\,\mathrm{dB}$ are provided in Figure 16.

With only 10 pairs, PSNR at $\mathrm{SNR} = -1\,\mathrm{dB}$ improves from 20.58 dB (cross-codec baseline) to 20.83 dB, recovering a meaningful fraction of the gap to the fully matched reference (21.56 dB). Gains are more pronounced at higher SNR: at $+5$ dB, PSNR rises from 21.18 dB to 21.61 dB. LPIPS improvements are most visible at moderate-to-high SNR (e.g., $0.3464 \rightarrow 0.3409$ at $+5\,\mathrm{dB}$ with 10 pairs). Increasing N from 10 to 25 yields additional but diminishing returns, primarily in LPIPS at mid-range SNR values ($0.3911 \rightarrow 0.3844$ at -1 dB). This suggests that the adaptation converges quickly due to the lightweight nature of the INN (only 0.32M parameters), and that even a very small calibration set is sufficient to substantially improve performance.

These results establish a practical deployment workflow: one can pre-train the INN in a cross-codec or cross-dataset setting using readily available data, and then rapidly specialise it to a new deployment with only a handful of target-domain pairs. This makes SING-INN well-suited for dynamic real-world environments where a small calibration set can be collected at deployment time (e.g., by transmitting a few known reference images), without requiring large-scale retraining.

I. *Source Faithfulness and Downstream-Task Performance*

PSNR, LPIPS, and FID quantify pixel fidelity, perceptual similarity, and distributional realism, but none of them verifies that a reconstruction stays tied to *its own* source image. We therefore add three label-based evaluations paired with the source in the most challenging bandwidth ratio, $\rho = 0.0013$, which is the regime in which generative ambiguity is the largest. All evaluations use frozen off-the-shelf models, so that no evaluation model is tuned to any method.

Identity consistency. Two independent face-recognition encoders, ArcFace [49] and AdaFace [50], are applied to identically preprocessed full frames without method-dependent detection or alignment. With $\phi(\cdot)$ the encoder, $\mathbf{x}_i$ a source image, and $\hat{\mathbf{x}}_i$ its reconstruction, we report the mean paired cosine similarity

$$C_{\mathrm{pair}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\phi(\hat{\mathbf{x}}_i)^{\top}\phi(\mathbf{x}_i)}{\|\phi(\hat{\mathbf{x}}_i)\|_2\,\|\phi(\mathbf{x}_i)\|_2}, \tag{17}$$

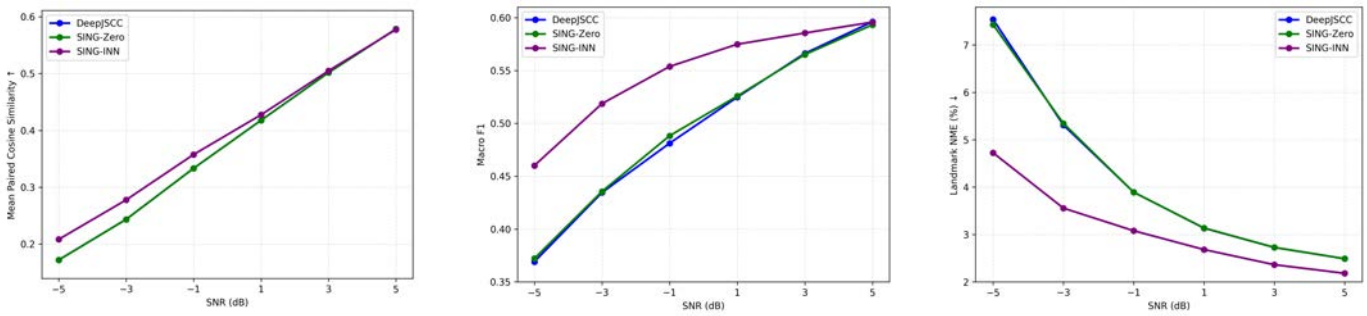
together with exact Top-1 paired-source retrieval, i.e., the fraction of reconstructions whose nearest neighbour in a gallery of all N clean images is its own source. The latter is a per-image consistency test rather than a person-classification score.

Downstream semantic task. A frozen FaRL [51] classifier predicts $K = 40$ binary facial attributes, and we report the macro-averaged accuracy and F_1 over attributes. Macro-F_1 weights every attribute equally and is therefore the more sensitive of the two for attributes with strongly imbalanced label frequencies.

Local structure and downstream alignment. A frozen FaRL 68-point landmark model is run independently on each reconstruction, with its output on the clean image as the paired reference. We report the inter-ocular normalized mean error

$$\mathrm{NME} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \frac{\frac{1}{68}\sum_{j=0}^{67}\|\mathbf{P}_{i,j} - \mathbf{G}_{i,j}\|_2}{\|\mathbf{G}_{i,36} - \mathbf{G}_{i,45}\|_2}, \tag{18}$$

where $\mathbf{G}_{i,j}$ and $\mathbf{P}_{i,j}$ are the reference and reconstructed coordinates of landmark j, landmarks 36 and 45 are the outer eye corners, and $\mathcal{V}$ collects the images detected for



(a) AdaFace paired cosine (higher better). (b) Attribute Macro F1 (higher better). (c) Landmark NME (lower better).

Fig. 17: Source faithfulness and downstream-task performance at the most challenging bandwidth ratio $\rho = 0.0013$: (a) identity consistency, (b) downstream attribute classification, and (c) local-geometry consistency. The remaining ArcFace/AdaFace cosine and Top-1 retrieval curves and the attribute Macro-Accuracy curves are provided in the supplementary material.

all methods. No clean-image landmarks are substituted upon detection failure and no registration is applied, so translation, scale, and local deformation all contribute to the error.

Figure 17 reports the results. At the most severe operating point, $\mathrm{SNR} = -5\,\mathrm{dB}$, SING-INN improves all three families of measures simultaneously: the AdaFace paired cosine rises from 0.172 for both DeepJSCC and SING-Zero to 0.208; attribute Macro-F_1 rises from 0.369 and 0.372 to 0.460; and landmark NME falls from 7.54% and 7.43% to 4.72%, a 37% relative reduction. ArcFace reproduces the same ordering independently, with Top-1 paired retrieval increasing from 0.172 to 0.240 (supplementary material). These are the largest margins observed anywhere in this paper: the faithfulness gain is therefore largest precisely in the extreme regime in which plausible-but-incorrect generation would be the dominant concern.

As the channel improves, the three curves converge, since the DeepJSCC output itself already retains most of the source information, and the residual differences become small and metric-dependent. Landmark NME still favours SING-INN at every tested SNR (2.18% versus 2.49% at $5\,\mathrm{dB}$), whereas ArcFace cosine and attribute Macro-Accuracy are marginally in favour of DeepJSCC at the highest SNRs (e.g., 0.532 versus 0.545 and 0.884 versus 0.886 at $5\,\mathrm{dB}$). These residual differences are of the order of the inter-method spread and reflect the distortion–perception trade-off [52] rather than a systematic change of identity or attributes. SING-Zero tracks DeepJSCC closely throughout, as expected from its conservative zero-shot projection. Overall, the LPIPS and FID gains of SING-INN are also accompanied by a measurably better preservation of the source-specific identity, semantics, and local geometry.

VI. Conclusion

In this paper, we introduced SING, a novel two-stage JSCC scheme for wireless image transmission that formulates the reconstruction of received corrupted images as an inverse problem. Depending on whether partial information about the DeepJSCC system and the channel is available, the degradation introduced by the channel and DeepJSCC can either be approximated as a linear transformation or modeled using an INN. Both the linear approximation and INN modeling enable integration of diffusion models into the reconstruction process. SING delivers image reconstructions with perceptual quality surpassing both DeepJSCC and InverseJSCC, particularly in extreme scenarios with very low BCRs and SNRs. Even in more challenging cases where there is a significant distribution mismatch between the training and testing dataset of DeepJSCC, SING demonstrates remarkable performance. Furthermore, we have demonstrated that SING-INN generalizes effectively across different codecs and image domains, and supports practical few-shot adaptation with as few as 10 target-domain calibration pairs, making it well-suited for dynamic real-world deployment scenarios.

Discussions and Future Directions

While SING achieves strong empirical performance, several limitations represent important directions for future research.

Linear approximation in SING-Zero. Modeling the highly nonlinear DeepJSCC pipeline as a fixed linear operator is an intentional simplification that balances model specificity and zero-shot generality. As shown in Section V-C, learning a data-dependent linear operator yields only a marginal LPIPS gain while reducing PSNR and zero-shot generality. Mean-pooling therefore remains a favorable choice for the strict zero-shot setting, while SING-INN provides the nonlinear alternative when paired observations are available. The smaller gains of SING-Zero under the most extreme conditions reflect this stricter information regime rather than a formal performance ceiling of the two-stage framework. Exploring more expressive zero-shot approximation strategies remains a direction for future work.

Hallucination and faithfulness guarantees. Hallucination, the generation of perceptually plausible but factually incorrect content, is a well-recognized open challenge for any restoration system built upon generative models. This tension is formally characterized by the distortion-perception trade-off [52]. SING mitigates this risk through its measurement-consistency mechanism. Empirically, the consistent maintenance or improvement of PSNR alongside substantial LPIPS and FID gains suggests that generated content is, on average, structurally faithful to the original images. However, average PSNR cannot rule out localized hallucinations, and providing formal guarantees on the faithfulness of diffusion-based reconstruction in highly

nonlinear degradation settings remains an active area of research.

Downstream tasks. Section V-I evaluates two downstream tasks, facial-attribute classification and face alignment, in addition to identity retrieval. Extending this evaluation to other task families (e.g., detection, segmentation, or retrieval in non-facial domains) and to task-aware variants of the restoration objective is another important direction for future work.

Theoretical guarantees. We acknowledge the absence of formal convergence, consistency, or optimality guarantees for the combined INN-guided diffusion process. Providing such guarantees for diffusion-based posterior sampling under highly nonlinear and stochastic degradation models remains an open and challenging problem for the broader research community. Recent preliminary theoretical work for simplified settings [53] provide a promising starting point, but extending them to the compound communication degradation scenario considered here is non-trivial.

REFERENCES

[1] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] S. Vembu, S. Verdu, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 44–54, 1995.

[3] D. Gunduz, E. Erkip, A. Goldsmith, and H. V. Poor, "Source and channel coding for correlated sources over multiuser channels," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 3927–3944, 2009.

[4] D. Gündüz, M. A. Wigger, T.-Y. Tung, P. Zhang, and Y. Xiao, "Joint source-channel coding: Fundamentals and recent progress in practical designs," *Proceedings of IEEE*, 2024.

[5] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "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] D. B. Kurka and D. Gündüz, "Deepjscc-f: Deep joint source-channel coding of images with feedback," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 178–193, 2020.

[7] T.-Y. Tung, D. B. Kurka, M. Jankowski, and D. Gündüz, "DeepJSCC-Q: Constellation constrained deep joint source-channel coding," *IEEE Journal on Selected Areas in Information Theory*, 2022.

[8] T.-Y. Tung and D. Gündüz, "DeepWiVe: Deep-learning-aided wireless video transmission," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2570–2583, 2022.

[9] M. Yang, C. Bian, and H.-S. Kim, "OFDM-guided deep joint source channel coding for wireless multipath fading channels," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 584–599, 2022.

[10] D. B. Kurka and D. Gündüz, "Bandwidth-agile image transmission with deep joint source-channel coding," *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 8081–8095, 2021.

[11] J. Xu, B. Ai, W. Chen, A. Yang, P. Sun, and M. Rodrigues, "Wireless image transmission using deep source channel coding with attention modules," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2021.

[12] J. Xu, T.-Y. Tung, B. Ai, W. Chen, Y. Sun, and D. Gündüz, "Deep joint source-channel coding for semantic communications," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 42–48, 2023.

[13] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[15] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *The Eleventh International Conference on Learning Representations*, 2023.

[16] D. You and P. L. Dragotti, "Indigo+: A unified inn-guided probabilistic diffusion algorithm for blind and non-blind image restoration," *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[17] H. Wu, Y. Shao, E. Ozfatura, K. Mikolajczyk, and D. Gündüz, "Transformer-aided wireless image transmission with channel feedback," *IEEE Transactions on Wireless Communications*, 2024.

[18] H. Wu, Y. Shao, K. Mikolajczyk, and D. Gündüz, "Channel-adaptive wireless image transmission with OFDM," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2400–2404, 2022.

[19] H. Wu, Y. Shao, C. Bian, K. Mikolajczyk, and D. Gündüz, "Deep joint source-channel coding for adaptive image transmission over MIMO channels," *IEEE Transactions on Wireless Communications*, 2024.

[20] K. Zhou, G. Zhang, Y. Cai, Q. Hu, G. Yu, and A. L. Swindlehurst, "Feature allocation for semantic communication with space-time importance awareness," *arXiv preprint arXiv:2401.14614*, 2024.

[21] X. Zhou, D. Huang, Z. Qi, L. Zhang, and T. Jiang, "A multi-scale spatial-temporal network for wireless video transmission," *arXiv preprint arXiv:2411.09936*, 2024.

[22] X. Wang, D. Ye, C. Feng, H. H. Yang, X. Chen, and T. Q. Quek, "Trustworthy image semantic communication with genai: Explainablity, controllability, and efficiency," *IEEE Wireless Communications*, vol. 32, no. 2, pp. 68–75, 2025.

[23] T. Marchioro, N. Laurenti, and D. Gündüz, "Adversarial networks for secure wireless communications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8748–8752.

[24] E. Erdemir, P. L. Dragotti, and D. Gündüz, "Privacy-aware communication over a wiretap channel with generative networks," in *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2989–2993.

[25] S. Guo, A. Zhang, Y. Wang, C. Feng, and T. Q. Quek, "Semantic-enabled 6g communication: A task-oriented and

privacy-preserving perspective," *IEEE Network*, 2025.
[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[28] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *International Conference on Machine Learning*, 2019, pp. 1182–1192.
[29] X. Niu, X. Wang, D. Gündüz, B. Bai, W. Chen, and G. Zhou, "A hybrid wireless image transmission scheme with diffusion," in *IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2023, pp. 86–90.
[30] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International conference on machine learning*, 2017, pp. 537–546.
[31] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 2437–2445.
[32] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2645–2657, 2023.
[33] S. F. Yilmaz, X. Niu, B. Bai, W. Han, L. Deng, and D. Gündüz, "High perceptual quality wireless image delivery with denoising diffusion models," in *IEEE Conference on Computer Communications Workshops*, 2024, pp. 1–5.
[34] G. Daras, J. Dean, A. Jalal, and A. G. Dimakis, "Intermediate layer optimization for inverse problems using deep generative models," in *International Conference on Machine Learning*, 2021.
[35] S. Tang, Q. Yang, D. Gündüz, and Z. Zhang, "Evolving semantic communication with generative modelling," in *2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2024, pp. 1–6.
[36] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," in *Int'l Conf. on Computer Vision (ICCV)*, vol. 1, 2021, p. 2.
[37] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *The Eleventh International Conference on Learning Representations*, 2023.
[38] T. Wu, Z. Chen, D. He, L. Qian, Y. Xu, M. Tao, and W. Zhang, "CDDM: Channel denoising diffusion models for wireless semantic communications," *IEEE Transactions on Wireless Communications*, 2024.
[39] L. Guo, W. Chen, Y. Sun, B. Ai, N. Pappas, and T. Quek, "Diffusion-driven semantic communication for generative models with bandwidth constraints," *arXiv preprint arXiv:2407.18468*, 2024.
[40] M. Zhang, H. Wu, G. Zhu, R. Jin, X. Chen, and D. Gündüz, "Semantics-guided diffusion for deep joint source-channel coding in wireless image transmission," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2025.
[41] S. Wang, J. Dai, K. Tan, X. Qin, K. Niu, and P. Zhang, "Diffcom: Channel received signal is a natural condition to guide diffusion posterior sampling," *IEEE Journal on Selected Areas in Communications*, 2025.
[42] J. Chen, D. You, D. Gündüz, and P. L. Dragotti, "Commin: Semantic image communications as an inverse problem with INN-guided diffusion models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6675–6679.
[43] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 913–11 924, 2020.
[44] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.
[45] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
[46] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk99zCeAb
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
[49] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
[50] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2022, pp. 18 729–18 738.
[51] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 18 676–18 688.
[52] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.
[53] X. Xu and Y. Chi, "Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction," *Advances in Neural Information Processing Systems*, vol. 37, pp. 36 148–36 184, 2024.