SPIFF: Selective Preservation of Image Fidelity for Bandwidth-constrained Heterogeneous Networks

Marco Palena
CNIT
Parma, Italy

Jose A. Ayala-Romero
NEC Laboratories Europe
Madrid, Spain

Andres Garcia-Saavedra
NEC Laboratories Europe
Madrid, Spain

Carla Fabiana Chiasserini
Politecnico di Torino and CNIT
Torino, Italy

***Abstract*—Transmitting rich visual data in resource-constrained environments like Non-Terrestrial Networks (NTNs) poses a significant challenge. While current Semantic Communication (SC) approaches reduce bandwidth consumption, they often lack flexibility and/or compromise the perceptual fidelity of critical details. This paper first analyzes the fundamental trade-offs that exist between perceptual fidelity, semantic fidelity, and bandwidth utilization. It then introduces SPIFF, an SC-Generative AI framework that, by supporting selective fidelity, enables fine-grained control over the above trade-offs while meeting delay requirements. SPIFF features a lightweight, semantic-aware encoder performing semantic segmentation and applying a novel patch preservation strategy that retains perceptually significant regions while adapting lower-relevance areas compression to bandwidth availability. SPIFF also offloads high-complexity reconstruction tasks to a Generative AI-enabled decoder at the receiver, thus addressing asymmetric computation requirements. To support adaptation under dynamic conditions, while meeting system and application constraints, we equip SPIFF with a learning-based decision engine that is able to cope with the system non-linearities and effectively tune SPIFF's configuration online. We evaluate SPIFF by implementing a full encoder-decoder pipeline. Results show that SPIFF fulfills perceptual reconstruction quality in scenarios where SC fails, and improves over state-of-the-art solutions both bandwidth savings (by up to 21%) and perceptual fidelity (by up to 13%).**

***Index Terms*—Non-terrestrial networks, Bandwidth utilization, Goal-oriented computing, Semantic communications**

I. Introduction

The ever-increasing demand for transmitting rich visual data poses a significant challenge for communication systems operating under severe resource constraints. For systems with limited bandwidth, power, or onboard computing—particularly at the transmitter, such as remote IoT sensors, autonomous drones, or satellites—the sheer volume of raw pixel data creates a critical bottleneck, hindering applications from remote monitoring to scientific exploration. To address this fundamental challenge, Semantic Communication (SC) is emerging as a transformative paradigm [1]. By leveraging Generative AI (GAI), SC shifts the focus from transmitting a perfect replica of the source data to conveying its essential semantic meaning, promising dramatic reductions in bandwidth consumption by discarding redundant information.

This work is funded by Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation program Grant Agreement No. 101192521 (MultiX), 101139266 (6G-INTENSE), 101139270 (ORIGAMI), and 101139232 (6G-GOALS).

This challenge is exceptionally well-exemplified by Non-Terrestrial Networks (NTNs). With extensive Low-Earth Orbit (LEO) deployments by entities like SpaceX launching over 7,000 satellites since 2019 [2], [3], NTNs are becoming vital for global connectivity. However, they face a fundamental conflict: they operate under strict spectrum and efficiency limits, with recent tests showing mobile download speeds of only 17 Mbps [4]. This limitation is severely strained by immense data demands, such as Earth observation missions generating petabytes of data cumulatively and terabytes daily [5], [6]. This stark combination of physical constraints and massive data requirements makes NTNs a critical and highly demanding application domain for advanced SC technologies.

The problem. While promising in principle, current GAI-enabled SC solutions have fundamental limitations that make them ill-suited for deployment in demanding, general-purpose environments. First, they are often tailor-made for specific applications with predictable, static scenes, such as highway surveillance [7], preventing them from handling diverse, unrestricted content. Second, in their quest for aggressive compression, these solutions substantially compromise *perceptual fidelity* (PF)—the visual accuracy of the reconstructed data [8], [9]. This is a critical flaw where content streams contain specific elements that cannot tolerate such fidelity degradation.

To illustrate this, consider the example in Fig. 1. A conventional SC system[1] can reconstruct the original image (a) as shown in (c). While conveying the core meaning, i.e., high *semantic fidelity* (SF), it suffers a catastrophic PF loss, rendering important details unrecognizable[2]. These details constitute the *high-relevance* segments (b). This issue is particularly acute in scenarios that handle diverse content. As quantified in Fig. 2, general-purpose visual data often has a high density of perceptually relevant information. For instance, the median image in COCO-Stuff (a dataset with content targeting object recognition) has 40% high-relevance content, compared to just 10% for the specialized MaSTr1325 dataset (with content related to maritime obstacle detection). This observation makes a one-size-fits-all compression strategy ineffective.

The solution. Consequently, a robust SC system for resource-constrained networks such as NTNs must support *selective fidelity*: preserving critical elements while aggressively

[1] I.e., SC methods that maximize semantic fidelity and compression rate.
[2] PF and SF scores, defined in Sec. II, range between 0 (worst) and 1 (best).

(a) Original (b) Relevance Map (c) Standard SC (d) SPIFF (e) SPIFF

Fig. 1. Demonstration of the fidelity-compression trade-off in GAI-enabled Semantic Communication. (a) Original image. (b) Image segmented into high (yellow) and low-relevance (purple) regions. (c) A standard SC method yields high compression but poor perceptual fidelity (PF). (d-e) Our proposed method, SPIFF, selectively preserves high-relevance regions. For (c-e), we report fraction of reconstructed pixels (%), bandwidth savings (%), and PF and Semantic Fidelity (SF) for high- and low-relevance regions. Fidelity scores are normalized to a 0-1 scale, where higher is better. Full technical details are in Sec. IV.

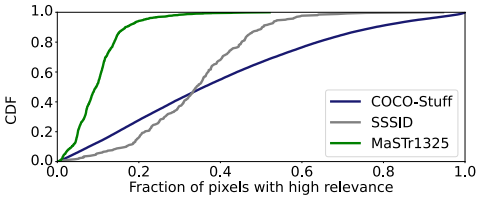


Fig. 2. Empirical CDF of the pixel relevance ratio for three distinct datasets, introduced in Table I. See Sec. II-B for details on the datasets.

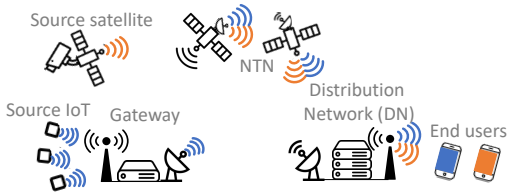


Fig. 3. Reference NTN scenario, with two data flows in different colors: an IoT resource-constrained transmitter and a satellite. The GW performs several operations on the images: i) semantic segmentation, ii) patch generation, and iii) image preparation and compression. Then the images are sent through a satellite network to a computationally capable distribution network, where image regeneration is executed. The regenerated image is finally delivered to the user. A similar scenario holds when a satellite is the image source.

compressing non-relevant content. To this end, we propose SPIFF (Selective Preservation of Image Fidelity Framework), a novel SC framework for images that enables a fine-grained trade-off between the PF of specific image regions, overall semantic integrity, and transmission bandwidth.

Unlike conventional methods that apply a uniform compression policy, our framework provides dynamic, content-aware control, allowing it to operate anywhere along a continuous spectrum bounded by two extremes: transmitting the original raw data (0% savings) and the highly compressed output of a conventional SC system (Fig. 1(c)). The outputs in Fig. 1(d) and (e) are merely two examples of the numerous intermediate operating points SPIFF can generate on demand to meet specific system constraints and task fidelity requirements.

Challenges and our contributions. The practical deployment of such a system presents two significant operational challenges: (1) *Asymmetric Computational Constraints*, where the transmitter (e.g., a satellite) is resource-constrained while the receiver (a ground station) is not; and (2) *Online Configuration Optimization*, the need to dynamically tune system parameters in response to variable network conditions and content characteristics, without a priori knowledge of the complex relationship between configuration and performance.

This paper addresses these challenges with two core contributions. First, we design the SPIFF pipeline, which supports selective fidelity while explicitly handling asymmetric computation. Its lightweight encoder identifies low-relevance semantic regions and applies a novel *patch preservation* technique, strategically removing only a fraction of patches from these regions, before transmission. This offloads the demanding generative inpainting task to the more powerful receiver. Second, to enable online adaptation, we introduce the *BITS (Bit-efficient Image Transmission via Satellite)* formulation and a learning-based decision-making engine that efficiently optimizes SPIFF's configuration in real-time.

In summary, our main contributions are as follows:

- A novel SC framework, SPIFF, enabling selective, dynamically adaptive fidelity for efficient image transmission.
- A lightweight, content-aware encoder architecture featuring a novel patch preservation mechanism, designed for resource-constrained transmitters.
- A GAI-decoder at the receiver, reconstructing the full image by properly exploiting the transmitted features.
- An optimization formulation and a learning-based solution engine that allows real-time adaptation of our framework in unpredictable NTN environments.
- An implementation of SPIFF's full encoder-decoder pipeline and a thorough experimental evaluation of its performance. Results show that SPIFF meets perceptual quality targets in scenarios where SC fails, and improves both bandwidth savings (by up to 21%) and perceptual fidelity (by up to 13%) w.r.t. state-of-the-art solutions.

II. Reference Scenario and Analysis

This section details our reference scenario, formalizes the concept of *perceptual relevance*, and presents a quantitative analysis of the fidelity-bandwidth trade-off that establishes the empirical foundation for our work.

A. Reference Scenario and Evaluation Metrics

We consider the reference scenario depicted in Fig. 3, where resource-constrained devices at the network edge generate and transmit image data through an NTN. These devices may be

TABLE I
PERCEPTUAL RELEVANCE RANKING OF THE SEMANTIC CATEGORIES IN THREE DIFFERENT DATASETS

Relevance level	Semantic Categories
Task 1: Object detection (COCO-Stuff dataset)	
Low	water, ground, sky, plant
Medium	solid, structural, building, textile, furniture, window, floor, ceiling, wall, rawmaterial
High	sports, accessory, animal, outdoor, vehicle, person, indoor, appliance, electronic, furniture, food, kitchen
Task 2: Semantic Segmentation of Satellite Imagery (SSSID dataset)	
Low	water_body, vegetation, flooded, trees, grass, background
Medium	industrial_site, sports_complex, water_tank, power_lines, construction_site, trampoline, garbage_bins, satellite_antenna, window, street_light, chimney, solar_panels, swimming_pool, secondary_structure, roof
High	boat, parking_area, road, vehicle
Task 3: Maritime Obstacle Detection (MaSTr1325 dataset)	
Low	water, sky
Medium	obstacle
High	unknown

terrestrial nodes (e.g., IoT sensors) or non-terrestrial assets (e.g., imaging satellites). The data is relayed via LEO satellites to a ground station and then to a computationally capable server in the distribution network (DN). At the destination, images are used for visualization or downstream machine perception tasks. This model captures numerous use cases such as precision agriculture and maritime navigation.

Operating in this scenario imposes three primary challenges: (*i*) strict bandwidth constraints due to spectrum scarcity, (*ii*) the need to preserve the PF of critical image segments for downstream tasks, and (*iii*) the requirement for bounded end-to-end delay. The first two challenges create a fundamental trade-off between data volume and image quality, which we evaluate using two distinct metrics.

Perceptual Fidelity (PF) assesses subjective visual similarity. We quantify PF using the Learned Perceptual Image Patch Similarity (LPIPS) score [10], which ranges from 0 (identical) to 1 (dissimilar). We report our results using $\mathbf{1}-$LPIPS, so that higher values signify higher perceptual fidelity.

Semantic Fidelity (SF) measures the preservation of high-level meaning. We use CLIPScore [11], which ranges from -1 to 1 (perfect alignment). We normalize this to a $[0, 1]$ scale to match that of the PF score.

B. Perceptual Relevance and Fidelity-Bandwidth Trade-offs

Preserving uniform PF across an entire image is often unnecessary and wasteful [12], [13]. We thus formalize the concept of perceptual relevance:

Definition (Category-wise Perceptual Relevance). The measure of how important it is to preserve the original visual appearance of a given semantic category for the task at hand.

Table I exemplifies the perceptual relevance ranking for three tasks. This concept is critical because, as shown in Fig. 2 with three public datasets aimed at different applications, different tasks have diverse proportions of high/medium/low-relevant pixels, making a selective approach essential. This motivates the central tenet of our work: *selectively preserving higher relevance regions while generatively reconstructing lower relevance ones can dramatically reduce bandwidth with minimal impact on task performance.*

To quantify the effects of this selective approach, we emulate the transmission pipeline of our SPIFF framework, as detailed in Sec. IV. The core idea is to split the image to transmit into regions with different perceptual relevance. First, high-relevance segments are preserved within a cropped bounding box. Second, from the remaining lower-relevance regions, we use a *patch preservation* technique to sample random patches, prioritizing those from medium-relevance over low-relevance areas. These selected patches are then assembled into a compact "patch grid." The transmitter configuration is controlled by two key parameters: (*i*) the size of the high-relevance bounding box, and (*ii*) the fraction of patches preserved to create the patch grid. These two parameters dictate the final data volume and quality of the reconstruction.

Both components—the cropped high-relevance image and the patch grid (corresponding to medium-low relevance regions)—are transmitted to the receiver. There, a generative inpainting model uses both pieces of information to reconstruct the full, original image. Specifically, we use *Stable Diffusion 2*, a Latent Diffusion Model (LDM) that reconstructs content by iteratively refining a noisy signal [14]. The number of these iterations, or *denoising steps*, directly controls the computational effort required at the receiver.

Our analysis of this process yields three key insights. *First*, by using an optimized configuration for a single image, we can trace a clear and controllable trade-off curve, as shown in Fig. 4. This demonstrates that there exists a relationship between the configuration parameters (namely, the high-relevance bounding box and the lower-relevance patches), the resulting PF and SF, and the bandwidth that can be saved with respect to transmitting an image compressed using a standard image codec (namely, JPEG).

Second, the approach inherently accommodates asymmetric computation, typical of many network environments exhibiting a high level of heterogeneity. The encoder performs comparatively lightweight tasks (segmentation and patch sampling), while the decoder handles the more computationally intensive generative reconstruction. Fig. 5, obtained using COCO-Stuff images, provides critical support for the efficiency of this design. It shows that reconstruction quality is driven by the richness of the transmitted information (the preserved patches), not the raw computational effort at the receiver (denoising steps). This confirms that our lightweight encoder effectively controls the final quality, while the decoder's computational load still remains manageable as few denoising steps suffice.

Third, and most critically, this trade-off is not universal. Fig. 6 visualizes this complexity by plotting the performance of a wide range of configurations (i.e., different bounding box sizes and patch preservation fractions) for several images from the COCO-Stuff dataset. Each point in the plot represents a specific configuration outcome in terms of bandwidth savings and PF. For every image, the set of optimal configurations forms a Pareto frontier of this point cloud. The crucial observation is that this optimal frontier is highly distinct for each image. This establishes the final, critical challenge: *the optimal transmission strategy for an arbitrary image, which*

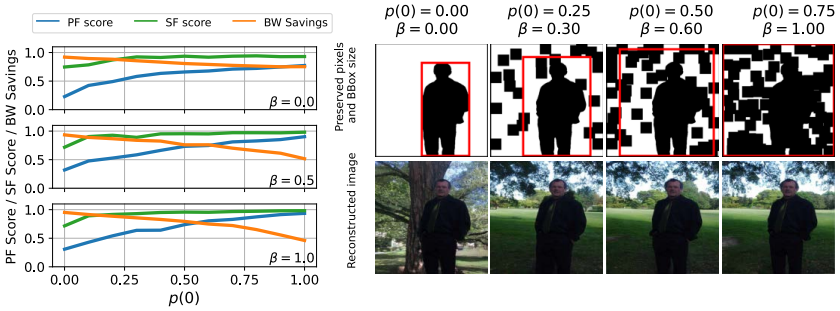


Fig. 4. Trade-off between PF and SF, and bandwidth savings (%), for a given image as the fraction of preserved pixels p and the size of the bounding box β varies. Bandwidth savings account for the high-relevance bounding box and lower relevance patches and are normalized w.r.t. the original image. Visual examples show the quality at different preservation levels, also highlighting the different sizes of the bounding box (in red).

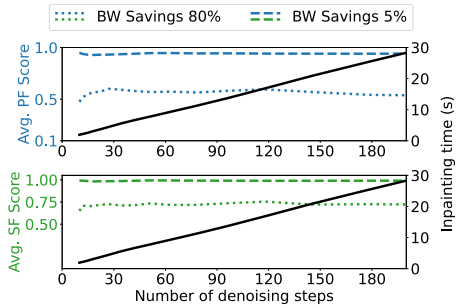


Fig. 5. PF score, SF score, and GPU inpainting time vs. the number of LDM denoising steps, and for 80% and 5% bandwidth savings. The results show that quality is largely insensitive to the number of steps, validating our architectural choices.

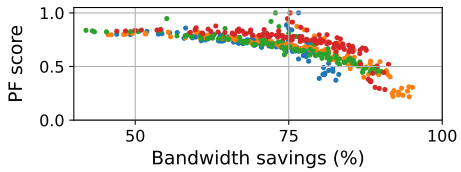


Fig. 6. The bandwidth savings-PF trade-off is unique to each image, as shown for four different images (represented by colors) from the COCO-Stuff dataset.

lies on this unknown frontier, cannot be predetermined. This reality necessitates an intelligent, online learning approach to discover these optimal configurations in real-time.

III. Related Work

Our work mainly relates to two research areas: semantic communication and deep learning-based image compression.

Semantic communication. SC is expected to be a critical technology for next-generation networks [15], and it is already driving advancements in a broad range of applications [16]. Generative methods have recently been applied to SC. Notably, generative adversarial networks (GANs) have been adapted for tasks like image compression [17], variational autoencoders [18], and normalizing flows [17]. However, some of these works incorrectly mix the semantics with stylistic elements, limiting SC full potential. More recently, the emergence of diffusion models represents a significant paradigm shift in generative modeling, achieving state-of-the-art performance in diverse domains such as image [19], audio [20], and video generation [21]. These models operate by progressively refining a standard Gaussian noise distribution via an iterative denoising process to produce novel content. This methodology provides superior training stability compared to GANs [22]. Moreover, diffusion models have recently been applied to SC as a new benchmark, generating high-fidelity, semantically coherent scenes that outperform prior methods [23].

Deep learning-based image compression. Some works rely on joint source channel coding (JSCC), which combines source coding and channel coding into a single, integrated process [24], [25]. The objective is to achieve good performance in very challenging channel conditions such as very low SNR or small bandwidth. The approach in [24] encodes image pixels directly as noise-resistant channel symbols, a strategy that surpasses standard separation-based digital communication at any SNR. The attention-based JSCC framework in [25] is notable for its ability to function effectively across different SNRs during transmission. However, a critical drawback of the aforementioned approaches is their lack of semantic awareness, i.e., the interpretation of the image content is missing.

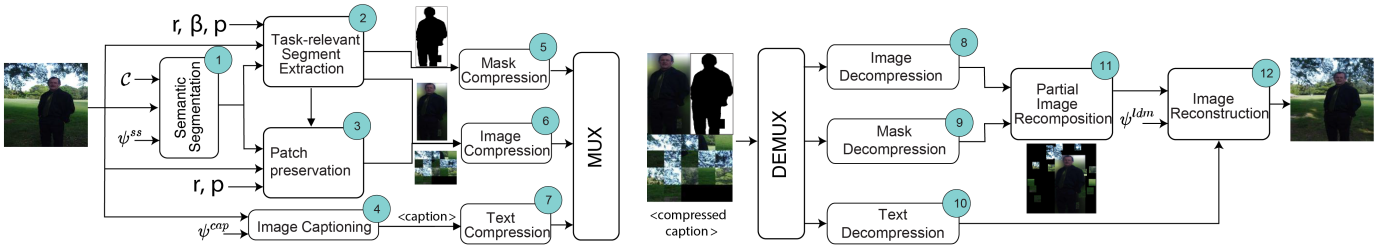
Other works propose deep learning-based codecs to solve the rate-distortion optimization problem [26], [27]. By pixel-by-pixel processing, these methods aim at perfect data recovery, pushing the boundaries of Shannon's compression theory. The downside is that aggressively compressed images are visually poor, suffering from artifacts like blocking, ringing, and blurriness [28]. This degradation also harms the performance of computer vision tasks such as classification and detection [29]. Consequently, simply minimizing per-pixel errors can be a wasteful strategy that incurs unnecessary data overhead.

Some studies leverage semantic similarity as the reconstruction criterion, i.e., they tolerate some pixel-level distortion and evaluate the usefulness of the reconstructed image. For example, [30] applies semantic similarity in low-bitrate facial image compression by filtering task-irrelevant information. Nevertheless, these works only consider face recognition and are not generalizable to other tasks. The work in [31] adapts a traditional hybrid coding framework for semantic tasks by using reinforcement learning to optimize bit allocation. Using deep Q-learning to set quantization parameters, it achieves strong results on classification, detection, and segmentation. In contrast, [32] presents a detection-focused compression scheme that creates a semantically structured bitstream where each part explicitly represents a specific object; this is then generalized to multiple tasks in [33].

Limitation of prior work. None of the above methods can meet the semantic and perception performance targets jointly with a low bit rate. In fact, different image regions may have different semantic or perceptual importance and thus should be adaptively compressed as we propose in this work.

IV. The SPIFF Semantic Communication System

This section details the architecture and operational workflow of our proposed SC framework, SPIFF. Our goal is a generative system that intelligently partitions an image by its perceptual relevance, ensuring high-fidelity transmission for critical content while efficiently reconstructing less critical



(a) The SPIFF Semantic Encoder Architecture

(b) The SPIFF Semantic Decoder Architecture

Fig. 7. Detailed schematics of the encoder and decoder components of the SPIFF framework.

Fig. 8. Visual walkthrough of the SPIFF pipeline. (a) Original image, (b) Semantic segmentation map, (c) High-relevance region mask, (d) Patches sampled from background, (e) Patch grid from sampled patches, (f) Combined data for reconstruction, (g) Final reconstructed image.

regions at the receiver to save bandwidth. Fig. 7 depicts the end-to-end architecture, with the numbered steps in the schematics corresponding to the descriptions below.

A. System Parameters and Configuration

The behavior of the SPIFF framework is governed by a set of parameters that define the task and control the system's operation. For a given application, these are:

- $R \in \mathbb{N}$: The number of discrete perceptual relevance ranks (e.g., $R=3$ for High, Medium, Low).
- $\mathcal{C}$: The set of all task-relevant semantic categories (e.g., 'person', 'sky').
- $r : \mathcal{C} \to \{0, \ldots, R-1\}$: The *relevance function* that maps each semantic category to a perceptual relevance rank. The highest rank, $R-1$, signifies the most critical content.
- $\beta \in [0, 1]$: The *bounding box parameter* that controls the size of the area preserved around the highest-relevance segments, with 0 signifying that the bounding box coincides with the segment size and 1 that the bounding box is as large as the entire image.
- $p : \{0, \ldots, R-1\} \to [0, 1]$: The *patch preservation function* that maps each perceptual relevance rank to a fraction of patches to preserve. By definition, $\sum_{l=0}^{R-1} p(l)=1$.
- $\psi^{ss}, \psi^{cap}, \psi^{ldm}$: Sets of hyperparameters for the underlying AI models: Semantic Segmentation, optional Image Captioning, and Latent Diffusion Model (LDM) for reconstruction, respectively.

While these parameters define the entire operational space, our BITS formulation (Sec. V) focuses on optimizing a specific subset of them in real-time. This subset forms the dynamic configuration vector $\mathbf{x}_i$, which we define in Sec. IV-C.

B. Framework Architecture

SPIFF consists of a lightweight semantic encoder at the transmitter and a generative semantic decoder at the receiver. A visual walkthrough of the pipeline is shown in Fig. 8.

1) Transmitter (Semantic Encoder): As detailed in Fig. 7a, the encoder analyzes the input image and transforms it into a compact representation through steps 1–7.

- **Semantic Segmentation (①):** The encoder first decomposes the image into its constituent semantic parts, identifying categories from the set $\mathcal{C}$. This is performed by a model like Grounded Segment Anything [34], governed by hyperparameters ψ^{ss}, producing a segmentation map as in Fig. 8b.
- **Task-relevant Segment Extraction (②):** Using the relevance function $r(c)$, this block identifies all segments with the highest rank (i.e., s.t. $r(c)=R-1$). It then isolates this content within a bounding box with size controlled by the parameter β. The resulting high-relevance crop and its corresponding mask (Fig. 8c) are passed forward for compression using standard techniques.
- **Patch Preservation (③):** For all lower-relevance regions located outside the bounding box from the previous step, this block applies the patch preservation function $p()$ to determine the fraction of patches to sample for each category's corresponding rank (Fig. 8d). Such patches and their coordinate metadata are consolidated into a "patch grid" (Fig. 8e).
- **Image Captioning (④):** In parallel, an optional captioning model, controlled by hyperparameters ψ^{cap}, can generate a text description of the image.
- **Component Compression and Multiplexing:** Finally, the various data components—the patch grid and the high-relevance crop (⑥), the associated mask (⑤), and the optional text caption (⑦)—are compressed and multiplexed into a single bitstream.

2) Receiver (Semantic Decoder): As shown in Fig. 7b, the receiver reconstructs the image from the incoming bitstream:

- **Demultiplexing and Decompression (⑧, ⑨, ⑩):** The receiver separates and decompresses the data components.
- **Partial Image Recomposition (⑪):** The high-relevance crop and patch grid are recomposed into a single, partially masked image using the coordinate metadata (Fig. 8f).
- **Generative Reconstruction (⑫):** The decoder uses the recomposed image and text caption to guide a generative model (e.g., Stable Diffusion 2 [14]), controlled by hyperparameters ψ^{ldm}, to inpaint the missing regions and produce the final image (Fig. 8g).

C. System Configuration Vector

For real-time adaptation, our BITS formulation focuses on the two primary parameters that control the fidelity-bandwidth trade-off at runtime. We therefore define the dynamic configuration vector $\mathbf{x}_i$ for each image I_i as $\mathbf{x}_i = (\beta_i, p_i)$, where (abusing the notation) β_i is the parameter controlling the *size of the high-relevance bounding box* and p_i is the *patch preservation function* that dictates the sampling fraction for lower-relevance regions, for image I_i. The goal of our BITS formulation, detailed in the next section, is to learn a policy that selects the optimal configuration vector $\mathbf{x}_i$ to adapt to changing network conditions and content characteristics.

V. Problem Formulation

Our aim is to design an intelligent decision-making framework that dynamically selects the optimal transmission configuration for each image in the NTN scenario introduced in Sec. II. The goal is to minimize bandwidth consumption while meeting application-specific constraints on perceptual fidelity, semantic fidelity, and end-to-end latency. To this end, we formally define the system model and problem at hand, named **Bit-efficient Image Transmission via Satellite (BITS)**.

We refer to the system in Fig. 3 and, for simplicity, focus on a single image, I_i, generated and transmitted by a given source to a final destination via an NTN. We consider the most general case where the source is a ground node and images are processed by a gateway (GW) using the SPIFF encoder before satellite transmission. Notice that this scenario can be easily specified to the case where satellite acts as the source and (possibly) the GW. Also, we recall that the image is then reconstructed at a node in the distribution network (DN).

In the above most general scenario, the entire system state is captured by $\Omega_t = (\omega_t^{\text{GW}}, \omega_t^{\text{NTN}}, \omega_t^{\text{DN}})$, representing the state in terms of communication and computation resources of the access link and the GW, the NTN, and the DN at time t (resp.). At the GW, the image I_i is processed according to a configuration vector $\mathbf{x}_i$, which encapsulates all decisions made by SPIFF. Given I_i, the configuration $\mathbf{x}_i$ then determines the value of the following features:

- **Size:** The size of the processed image, $\sum_{l=0}^{R-1} b_{i,l}(\mathbf{x}_i, I_i)$, measured in bits. This value is content-dependent and unknown a priori for a given configuration $\mathbf{x}_i$.
- **Fidelity:** The Perceptual Fidelity (PF) and Semantic Fidelity (SF) of the reconstructed image. For a given perceptual relevance rank l, these are denoted by the functions $f_l^{\text{per}}(\mathbf{x}_i)$ and $f_l^{\text{sem}}(\mathbf{x}_i)$, respectively.
- **End-to-End Delay:** The total latency $d_i(\mathbf{x}_i, I_i, \Omega_t)$ required to deliver and reconstruct the image I_i. This delay is the sum of several components:

$$d_i(\cdot) = d_{\text{GW}}^{\text{comm}} + d_{\text{GW}}^{\text{comp}} + d_{\text{NTN}}^{\text{comm}} + d_{\text{DN}}^{\text{comp}} + d_{\text{DN}}^{\text{comm}} \tag{1}$$

where the terms represent: the communication time from source to GW, computation time at the GW, NTN transmission time, reconstruction (inpainting) time at the DN server, and the DN communication time to deliver the reconstructed image at the destination. Each component depends on the size of the original image, the overall size of the processed image—hence, the selected configuration vector $\mathbf{x}_i$—and the state (including bandwidth availability) of the corresponding system in Ω_t.

Then we formulate the BITS problem for the image I_i as:

BITS (Bit-efficient Image Transmission via Satellite)

$$\min_{\mathbf{x}_i} \quad \sum_{l=0}^{R-1} b_{i,l}(\mathbf{x}_i, I_i) + \Big(\lambda_l^{\text{per}} [f_{l,\min}^{\text{per}} - f_l^{\text{per}}(\mathbf{x}_i)]^+ + \lambda_l^{\text{sem}} [f_{l,\min}^{\text{sem}} - f_l^{\text{sem}}(\mathbf{x}_i)]^+ \Big) \tag{2a}$$

$$\text{s.t.} \quad d_i(\mathbf{x}_i, I_i, \Omega_t) \le d^{\max} \tag{2b}$$

where $[z]^+ = \max(0, z)$. The objective (2a) seeks to minimize the total number of transmitted bits for the image I_i (i.e., the sum of the bits transmitted for each perceptual relevance rank), while penalizing any deviation from the minimum required fidelity targets, $f_{l,\min}^{\text{per}}$ and $f_{l,\min}^{\text{sem}}$. Such a formulation thus makes these fidelity requirements soft constraints ensuring problem feasibility, while the associated penalty weights, λ_l^{per} and λ_l^{sem}, make the terms in the objective function vary over a comparable range of values while allowing for application-specific prioritization. The constraint (2b) instead imposes a hard deadline $d^{\max}$ on the end-to-end delay. Solving this problem is highly challenging for the following reasons:

(i) High variability in the fidelity values. As shown in Fig. 6, although there exists a correlation between the minimum PF score and the bandwidth savings (i.e., % of preserved pixels), the measurements exhibit a high degree of variability. This is due to the randomness introduced by the diffusion model as well as the specific characteristics of each image. This variability, which we observed also in the case of the SF score, makes it hard to accurately predict the value of such metrics for the images at hand.

(ii) Hard constraints under uncertainty. Constraint (2b) deals with the end-to-end delay of the system. Even when the system state (Ω_t) is fully observable, in practice the transmission delay may suffer random variations due to the inherent stochasticity of the system.

(iii) Unknown and complex objective function. The problem objective includes $2{\cdot}R{+}1$ terms and a non-linear operator ((2a)), which may be hard to approximate, even by traditional function approximators (e.g., fully connected neural networks).

(iv) Variable objective function. The target balance between the PF and SF scores may change over time to adapt to different application requirements (λ's values in (2a)).

These challenges preclude traditional optimization methods and motivate our adoption of a learning-based approach, which we detail in the subsequent sections.

VI. Learning-based SPIFF Optimization

This section describes our learning-based SPIFF engine, which addresses the above challenges and solves the BITS problem. The solution framework architecture is depicted in Fig. 9. The learning engine takes as input the state s_i, which characterizes the image I_i, and outputs the configuration $\mathbf{x}_i$.

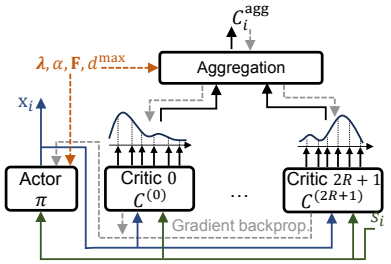


Fig. 9. Architecture of the learning-based SPIFF solution engine.

Given that we solve the problem image by image and assume that images are not correlated over time, the problem can be seen as a contextual bandit problem with instantaneous constraints. Indeed, to address the challenges *(i)* and *(ii)* described in Sec. V (namely, the variability and stochastic nature of the quantities in the BITS problem), the solution engine should learn the distribution of the fidelity scores and of the latency, and their quantiles, so that it can provide a configuration policy that meets both the soft and the hard constraints with very high probability. In the case of the soft *minimum* fidelity constraints, the goal is to approximate a *low* quantile; for the hard *maximum* latency constraint, a *high* quantile has to be approximated. We thus propose an actor-multi-critic architecture with distributional critics, in which the critics approximate the distributions of the objective function and constraints in the BITS problem, while the actor learns from these approximations the optimal configuration policy.

Further, to address challenge *(iii)*, the proposed solution uses several critics to approximate the distribution of each of the terms in (2a) separately, making the aggregated approximation easier to accomplish. Finally, our learning architecture also addresses challenge *(iv)*, as it enables learning the policy as a function of the problem hyperparameters (e.g., the λ's), without the need for retraining when the requirements change.

Below, we detail how to train a distributional critic (Sec. VI-A) and learn the configuration policy (Sec. VI-B).

A. Distributional Critic Training

This section introduces the methodology to approximate the distribution of a target function using quantile regression, which will then be used by the distributional critics.

The value of the quantile function q_τ for a distribution X is defined as $q_\tau{=}F_X^{-1}(\tau)$, where $\tau{\in}[0,1]$ represents a quantile and $F_X(x)$ denotes the cumulative distribution function. To learn the distribution, we employ a neural network (critic) trained to minimize the quantile regression loss—an asymmetric convex function that differentially penalizes underestimation and overestimation errors:

$$\mathcal{J}^\tau(\hat{q}_\tau) := \mathbb{E}_{x\sim X}\left[\rho_\tau(x-\hat{q}_\tau)\right]. \tag{3}$$

In (3), $\hat{q}_\tau$ is the estimated value of q_τ, and $\rho_\tau(u){:=}u\cdot(\tau{-}\delta_{[u<0]})\ \forall u{\in}\mathbb{R}$, where $\delta[z]$ is the indicator function.

The critic approximates N quantile values, $q_{\tau_1},\dots,q_{\tau_N}$, by minimizing the following objective via gradient descent: $\sum_{i=1}^{N}\mathcal{J}^{\tau_i}(\hat{q}_{\tau_i})$. To mitigate the non-smoothness of the quantile regression loss at $u{=}0$, which can hinder the performance of function approximators such as neural networks, we incorporate the quantile Huber loss [35]. This loss has a quadratic form within the interval $[-\kappa,\kappa]$, transitioning to the standard quantile loss outside this range:

$$J_\kappa(u) := \begin{cases} \frac{1}{2}u^2 & \text{if } |u|\le\kappa \\ \kappa\cdot(|u|-\frac{1}{2}\kappa) & \text{otherwise.}\end{cases} \tag{4}$$

The asymmetric quantile Huber loss is then defined as: $\rho_\tau^\kappa(u) := |\tau-\delta_{[u<0]}|\frac{J_\kappa(u)}{\kappa}$. Replacing $\rho_\tau(u)$ in (3) with $\rho_\tau^\kappa(u)$ yields the quantile Huber loss we use, which converges to the quantile regression loss as κ approaches zero.

B. Learning the Optimal Configuration Policy

We now detail the architecture of the overall SPIFF learning framework, which comprises an actor-multi-critic scheme [36]. The actor network, $\pi(s|\eta,\boldsymbol{\lambda})$, outputs a continuous action (i.e., $\mathbf{x}$) for a given state s and $\boldsymbol{\lambda}{=}\{\lambda_l^{\text{per}},\lambda_l^{\text{sem}},\lambda^{\text{del}}\}_{l=0}^{R-1}$ value. We define the state s as a vector characterizing an image, with R dimensions indicating the percentage of pixels belonging to each perceptual relevance rank. The actor is a neural network defined by the parameters (weights) η.

The distributional critics are denoted by $C^{(j)}(s,\mathbf{x}|\theta^{(j)})$ for $j{=}0,\dots,2{\cdot}R+1$, where $\theta^{(j)}$ are the parameters (weights) of critic j. They provide estimations of each term in the objective function and constraints. The aggregated cost function for a given image, C^{agg}, integrates information from all critics:

$$\begin{aligned} C^{\text{agg}}(s,\mathbf{x},|\,\theta) :=& \\ \bar{C}^{(0)}(s,\mathbf{x}\,|\,\theta^{(0)}) &+ \sum_{j=1}^{2\cdot R}\boldsymbol{\lambda}_j[\mathbf{F}_{j,\min} - Q^\alpha(C^{(j)}(s,\mathbf{x}\,|\,\theta^{(j)}))]^+ \\ &+ \lambda^{\text{del}}[Q^{1-\alpha}(C^{(2\cdot R+1)}(s,\mathbf{x}\,|\,\theta^{(2\cdot R+1)})) - d^{\max}]^+ \end{aligned} \tag{5}$$

where $Q^\alpha(X)$ represents the quantile function value of distribution X at quantile α, $\mathbf{F}_{\min}{=}\{f_{l,\min}^{\text{per}},f_{l,\min}^{\text{sem}}\}_{l=0}^{R-1}$ and $\boldsymbol{\lambda}$ are a compact representation of the minimum PF and SF scores, and of their associated weights, for all perceptual relevance ranks, λ^{del} is a penalty weight for the delay constraint, $\theta{=}\{\theta^{(0)},\dots,\theta^{(2\cdot R+1)}\}$ denotes the collective critic parameters, and $\bar{C}^{(0)}(\cdot)$ is the mean of critic 0's output distribution. In detail, Eq. (5) comprises three terms. The first one approximates the average of the used bits. The second term approximates the second part of the objective function in (2a), i.e., the penalty incurred by the violation of the minimum fidelity scores. We recall that we consider a low quantile of the distribution of the fidelity scores, denoted by α. Finally, the last term captures the penalty yielded by a violation of the maximum latency constraint. In this case, we consider a high quantile $(1-\alpha)$ of the distribution of the delay, to meet the constraint with high probability.

The actor's objective function is defined as $V(\pi) = \mathbb{E}_{s\sim\gamma}[C^{\text{agg}}(s,\pi(s\ |\ \eta,\boldsymbol{\lambda})\ |\ \theta)]$, where $\gamma(s)$ represents the stationary distribution of the states s. The update rule, derived via the chain rule applied to the actor objective [37], is:

$$\nabla_\eta V(\pi)\approx\mathbb{E}_{s\sim\gamma}\left[\nabla_a C^{\text{agg}}(s,\mathbf{x}|\theta)\,|_{\mathbf{x}=\pi(s|\eta,\boldsymbol{\lambda})}\ \nabla_\eta\pi(s|\eta,\boldsymbol{\lambda})\right].$$

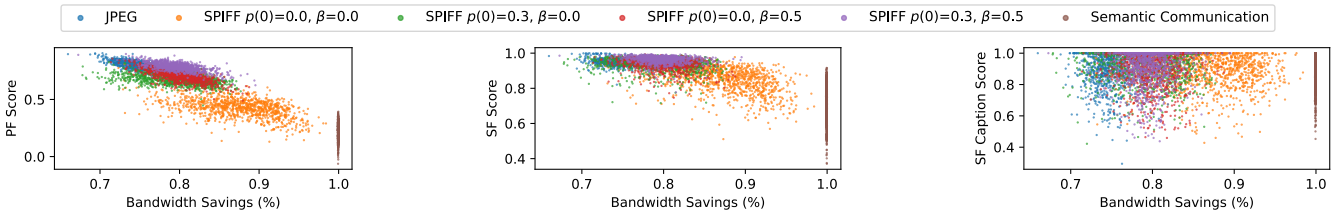


Fig. 10. SPIFF (under various settings of the patches fraction to preserve $p(0)$ and bounding box size β) vs. the JPEG and Semantic Communication benchmarks, for varying bandwidth savings. Each point represents the PF score (left), SF scores (center), and SF score caption (right) computed on an entire image.

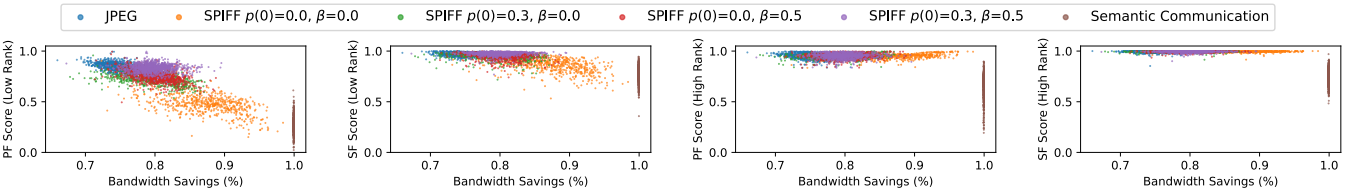


Fig. 11. SPIFF (under different settings of the patches fraction to preserve $p(0)$ and bounding box size β) vs. JPEG and Semantic Communication, for varying bandwidth savings. PF and SF scores for 'Low' rank (left and center-left), and PF and SF scores for 'High' rank (center-right and right).

VII. Experimental Evaluation

This section first introduces the experimental settings used for assessing SPIFF's performance. Then, it analyzes the impact of different system configurations on the fidelity of the reconstructed image and the achievable bandwidth savings, comparing the SPIFF pipeline with traditional image encoding and standard Semantic Communication (SC). Finally, it presents the effectiveness of SPIFF's learning engine in finding optimal configurations against three benchmarks.

Experimental Settings. We implemented the full SPIFF encoder-decoder pipeline (see Fig. 7) using state-of-the-art pretrained models, as detailed in Table II. The semantic encoder is executed on a system with 2 vCPUs (Intel Xeon @ 2.20GHz) and 51GB of RAM, while the decoder runs on an NVIDIA Tesla T4 GPU with 16GB memory. We focus on a person monitoring task, thus we define the task-relevant semantic categories as $\mathcal{C}$={'other', 'person'} and the relevance function as r('other')=0 ('Low' relevance) and r('person')=1 ('High' relevance); hence, we have R=2. Experiments are conducted on a subset of the COCO-Stuff dataset, including 1,062 images where (i) a ground-truth segmentation annotation for 'person' is available, and (ii) the 'person' segment occupies between 20% and 60% of the total image area.

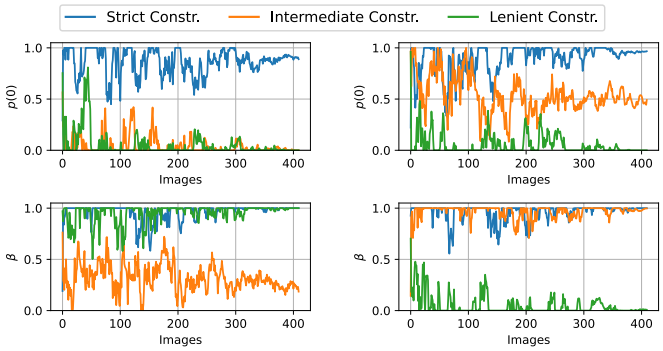
For the implementation of the decision-making engine, we configure all actor and critic neural networks with two hidden layers of 256 units. We update the networks weights using Adam and learning rates of 10^{-4} and 10^{-3} for the actor and critics, respectively. We use Ornstein-Uhlenbeck process with the parameters θ_{noise}=0.15 and σ_{noise}=0.15 for the exploration noise [38]. Finally, we use a reply buffer of $2,000$ samples and a minibatch size of 32 samples.

Without loss of generality, we set a delay constraint $d^{\max}$=20 s, λ_1^{per}=λ_1^{sem}=10, $f_{1,\min}^{\text{per}}$=$f_{1,\min}^{\text{sem}}$=1, and study different combinations of λ_0^{per}, λ_0^{sem}, $f_{0,\min}^{\text{per}}$ and $f_{0,\min}^{\text{sem}}$. To this end, we simulate a satellite communication link according to 3GPP guidelines [39], and evaluate SPIFF against three benchmarks: (i) *Traditional Communication (JPEG):* Images are encoded using the JPEG standard with a quality level set a priori to 45. This level was chosen to make sure this benchmark always meets the established delay constraint. (ii) *Standard Semantic Communication (SC):* The transmitter generates a textual caption of the input image, which is then sent over the channel. The receiver reconstructs the image based on this caption, using an LDM. The captioning and LDM models are the same as those used in the SPIFF pipeline (Table II). (iii) *Neural Contextual Bandit (NCB)*: This corresponds to SPIFF's framework but, instead of using our learning engine, it uses the widely used Neural Contextual Bandit (NCB) approach [36] (with an actor and a single critic).

TABLE II
State-of-the-art ML models used to implement the building blocks of the SPIFF framework

Block	Implementation	Parameters
Semantic Segmentation	MobileViT + DeepLabV3 (small) [40]	∼6.4M parameters, pre-trained on PASCAL VOC
Image Captioning	BLIP with ViT-B/16 backbone [41]	∼213M parameters, pre-trained on COCO, max. length=100
LDM	Stable Diffusion 2 [42]	∼890M parameters, pre-trained on LAION-5B, 50 denoising steps

Configuration Impact Analysis. Fig. 10 illustrates the trade-offs between fidelity (PF and SF scores) and bandwidth savings for various SPIFF configurations, benchmarked against JPEG and standard SC. As expected, SC achieves extreme compression (nearly 100% savings) but at the cost of very poor perceptual fidelity and inconsistent semantic fidelity. Conversely, JPEG provides reasonably good fidelity but is inflexible, offering only a narrow range of bandwidth savings (up to 80%). SPIFF, in contrast, demonstrates the ability to flexibly navigate the entire fidelity-bandwidth spectrum. The results highlight that preserving even a small fraction of low-relevance patches ($p(0)$>0, denoted by green and purple points) is crucial for maintaining high overall fidelity. Configurations that discard all low-relevance patches ($p(0)$=0, orange and red points) achieve much higher bandwidth savings, approaching SC levels but with significantly better and more controllable fidelity. Further, for a given patch preservation fraction, adjusting the bounding box size (β) allows for fine-



(a) Low pixel fraction (L) (b) High pixel fraction (H)

Fig. 12. SPIFF's convergence to optimal $p(0)$ and β configuration values under different PF and SF scores requirements. Results for a low (a) and high (b) pixel fraction taken by the segment labeled as 'person'.

tuning this trade-off. For instance, when $p(0)>0$, increasing β (purple vs. green points) can slightly improve bandwidth savings by reducing the patch grid size while still preserving the most critical content.

To highlight SPIFF's core benefit of *selective fidelity*, Fig. 11 isolates the performance for high- and low-relevance regions separately. Unlike JPEG, which compresses the entire image uniformly, SPIFF successfully maintains very high PF and SF scores for the high-relevance 'person' rank across all configurations. It selectively degrades instead the quality of the low-relevance 'other' rank to maximize bandwidth savings. This confirms that SPIFF can minimize data transmission without compromising the integrity of task-critical image components, a capability that both benchmarks lack.

Learning Engine Evaluation. We now evaluate the ability of the SPIFF learning engine to converge to an optimal configuration under different operational scenarios. Fig. 12 shows the engine's behavior for two distinct image subsets: one where the 'person' segment occupies a high ('H', i..e, 0.45 ± 0.04) pixel fraction and one where it occupies a low ('L', i.e., 0.22 ± 0.01) pixel fraction. We test the engine under three fidelity requirement levels, as defined in Table III. The results show that the engine intelligently adapts its policy based on both content and constraints. For 'L' images (Fig. 12a) and lenient fidelity requirements, the engine quickly learns to maximize bandwidth savings by setting a low patch preservation fraction ($p(0)\rightarrow0$), which mitigates the impact of β. In contrast, when faced with the same 'L' images but under strict fidelity targets, the engine selects a high $p(0)$. Conversely, for 'H' images (Fig. 12b), $\beta\rightarrow0$ only for lenient targets. This strategy maintains the quality of the large, low-relevance background, which is necessary to meet the strict fidelity target, while perfectly preserving the small, high-relevance region. In all cases, the hard delay constraint is satisfied.

Comparison. We finally compare the performance of SPIFF against all the benchmarks as summarized in Table IV. The comparison is conducted across the three scenarios defined by the fidelity requirements in Table III: Strict, Intermediate, and Lenient. We report the empirical probability of satisfying the perceptual (f_0^{per}) and semantic (f_0^{sem}) fidelity constraints for the low-relevance rank (as the high one is always fully preserved), alongside the average bandwidth savings. Under Strict fidelity targets, SPIFF shows a superior balance of performance. It meets the PF constraint in 96% of cases, significantly outperforming the NCB and JPEG benchmarks by 12% and 13%, respectively. SC fails to meet this target. Notably, SPIFF achieves its high fidelity with 74% bandwidth savings, comparable to JPEG's 75% and 21% higher than NCB's. For the Intermediate case, SPIFF proves to be a reliable approach, meeting both constraints with very high probability. In contrast, NCB's performance collapses, satisfying the PF and SF targets only 40% and 37% of the time, respectively. While JPEG also meets the fidelity targets, SPIFF provides slightly better bandwidth savings (79% vs. 75%). For the Lenient targets, all methods except SC succeed. In this case, SPIFF stands out by yielding the highest bandwidth savings (85%), surpassing both NCB (83%) and JPEG (75%).

In summary, SPIFF consistently provides the best trade-off across all scenarios. It effectively adapts its policy to either guarantee high fidelity under strict constraints or to maximize bandwidth savings under more relaxed conditions, proving its effectiveness and versatility compared to the benchmarks.

TABLE III
PERCEPTUAL FIDELITY (PF) AND SEMANTIC FIDELITY (SF) TARGETS

Target	**PF** ($f_{0,\text{min}}^{\text{per}}$)	**SF** ($f_{0,\text{min}}^{\text{sem}}$)	λ_0^{per}	λ_0^{sem}
Strict	0.8	0.9	1.25	5
Intermediate	0.6	0.8	1.2	4.75
Lenient	0.4	0.6	0.1	0.5

TABLE IV
COMPARISON BETWEEN SPIFF AND OTHER BENCHMARKS

Target	**Scheme**	$\mathbf{P}(\mathbf{f_0^{per}}\geq\mathbf{f_{0,\text{min}}^{per}})$	$\mathbf{P}(\mathbf{f_0^{sem}}\geq\mathbf{f_{0,\text{min}}^{sem}})$	**BW Savings (%)**
Strict	SPIFF	0.96	1.00	74%
	NCB	0.84	0.99	53%
	JPEG	0.83	1.00	75%
	SC	0.00	0.02	99%
Interm.	SPIFF	0.98	0.94	79%
	NCB	0.40	0.37	83%
	JPEG	1.00	1.00	75%
	SC	0.00	0.38	99%
Lenient	SPIFF	1.00	1.00	85%
	NCB	1.00	1.00	83%
	JPEG	1.00	1.00	75%
	SC	0.05	0.98	99%

VIII. CONCLUSIONS

We tackled the efficient transmission of visual data under strict resources constraints, as it is the case in non-terrestrial networks, and proposed SPIFF, a novel framework that overcomes the lack of adaptability of current approaches by introducing the concept of selective fidelity. SPIFF incorporates a lightweight semantic encoder that leverages a novel patch preservation strategy, and a generative decoder. For robustness in highly dynamic and unpredictable scenarios, SPIFF also integrates a learning-based decision engine capable of online configuration tuning, which trades off perceptual fidelity, semantic fidelity, and bandwidth utilization at best, while fulfilling delay constraints. Experimental results, obtained using a complete SPIFF encoder-decoder implementation, show that SPIFF succeeds in meeting fidelity targets where semantic communication fails, and improves over state-of-the-art solutions bandwidth savings (by up to 21%) and perceptual fidelity (by up to 13%).

REFERENCES

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

[2] Earth Observation Portal, "Starlink satellite mission," https://www.eoportal.org/satellite-missions/starlink, 2024, accessed: July 2025.

[3] Spaceflight Now, "Live coverage: SpaceX Falcon 9 Rocket to Launch 23 Starlink satellites from Cape Canaveral," https://spaceflightnow.com/2023/12/06/live-coverage-spacex-falcon-9-rocket-to-launch-23-starlink-satellites-from-cape-canaveral/, 2023, accessed: July 2025.

[4] B. Guo, Z. Xiong, B. Wang, T. Q. S. Quek, and Z. Han, "Semantic communication-aware end-to-end routing in large-scale LEO satellite networks," in *IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*, 2024, pp. 137–142.

[5] H. Ramapriyan, "The role and evolution of NASA's Earth science data systems," in *IEEE EDS/CAS Chapter Meeting*, no. GSFC-E-DAA-TN24713, 2015.

[6] V. C. Gomes, G. R. Queiroz, and K. R. Ferreira, "An overview of platforms for big earth observation data management and analysis," *Remote Sensing*, vol. 12, no. 8, p. 1253, 2020.

[7] W. Yang, Z. Xiong, Y. Yuan, W. Jiang, T. Quek, and M. Debbah, "Agent-driven generative semantic communication with cross-modality and prediction," *IEEE Trans. on Wireless Communications*, 01 2024.

[8] R. Cheng, Y. Sun, D. Niyato, L. Zhang, L. Zhang, and M. A. Imran, "A wireless ai-generated content (aigc) provisioning framework empowered by semantic communication," *IEEE Trans. on Mobile Computing*, vol. 24, no. 3, pp. 2137–2150, 2025.

[9] Y. Lin, Z. Gao, H. Du, D. Niyato, J. Kang, A. Jamalipour, and X. S. Shen, "A unified framework for integrating semantic communication and AI-generated content in Metaverse," *Netw. Mag. of Global Internetwkg.*, vol. 38, no. 4, p. 174–181, 2024. [Online]. Available: https://doi.org/10.1109/MNET.2023.3321539

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF CVPR*, 2018.

[11] J. Hessel, A. Holtzman, M. Forbes, R. Bras, and C. Yejin, "CLIPScore: A reference-free evaluation metric for image captioning," 2021, pp. 7514–7528.

[12] Y. Wang, P. H. Chan, and V. Donzella, "Semantic-Aware Video Compression for Automotive Cameras," *IEEE Trans. on Intelligent Vehicles*, vol. 8, no. 6, pp. 3712–3722, Jun. 2023.

[13] V. Sivaraman, P. Karimi, V. Venkatapathy, M. Khani, S. Fouladi, M. Alizadeh, F. Durand, and V. Sze, "Gemino: practical and robust neural compression for video conferencing," in *USENIX Symposium on Networked Systems Design and Implementation*, ser. NSDI'24. USA: USENIX Association, 2024.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF CVPR*, June 2022, pp. 10 684–10 695.

[15] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.

[16] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. on Wireless Communications*, vol. 22, no. 9, pp. 6227–6240, 2023.

[17] T. Han, J. Tang, Q. Yang, Y. Duan, Z. Zhang, and Z. Shi, "Generative model based highly efficient semantic communication approach for image transmission," in *IEEE ICASSP*, 2023.

[18] A. H. Estiri, M. R. Sabramooz, A. Banaei, A. H. Dehghan, B. Jamialahmadi, and M. J. Siavoshani, "A variational auto-encoder approach for image transmission in wireless channel," *arXiv preprint arXiv:2010.03967*, 2020.

[19] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[20] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.

[21] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," *arXiv preprint arXiv:2205.15868*, 2022.

[22] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[23] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," *arXiv preprint arXiv:2207.00050*, 2022.

[24] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. on Cognitive Comm. and Netw.*, vol. 5, no. 3, pp. 567–579, 2019.

[25] 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 Trans. on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2315–2328, 2021.

[26] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *IEEE/CVF CVPR*, 2018, pp. 4394–4402.

[27] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. J. Hwang, J. Shor, and G. Toderici, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," in *IEEE/CVF CVPR*, 2018, pp. 4385–4393.

[28] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *IEEE/CVF ICCV*, 2019, pp. 221–231.

[29] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *Iet Biometrics*, vol. 7, no. 1, pp. 81–89, 2018.

[30] Z. Chen and T. He, "Learning based facial image compression with semantic fidelity metric," *Neurocomputing*, vol. 338, pp. 16–25, 2019.

[31] X. Li, J. Shi, and Z. Chen, "Task-driven semantic coding via reinforcement learning," *IEEE Trans. on Image Processing*, vol. 30, pp. 6307–6320, 2021.

[32] T. He, S. Sun, Z. Guo, and Z. Chen, "Beyond coding: Detection-driven image compression with semantically structured bit-stream," in *IEEE Picture Coding Symposium (PCS)*, 2019.

[33] S. Sun, T. He, and Z. Chen, "Semantic structured image coding framework for multiple intelligent applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3631–3642, 2020.

[34] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded SAM: Assembling open-world models for diverse visual tasks," 2024.

[35] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[36] J. A. Ayala-Romero, A. Garcia-Saavedra, and X. Costa-Perez, "Risk-aware continuous control with neural contextual bandits," in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 20 930–20 938.

[37] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International conference on machine learning*. Pmlr, 2014, pp. 387–395.

[38] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[39] 3GPP, "Solutions for NR to support non-terrestrial networks (NTN)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.821, 2023, version 16.2.0.

[40] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," 2022. [Online]. Available: https://arxiv.org/abs/2110.02178

[41] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF CVPR*, 2022, pp. 10 684–10 695.