

Article

Energy-Efficient Autoencoder-Based Compact Image Payload Transmission over Noisy Indoor Industrial VLC Links

Alejandro Arratia Pavat ¹, Pablo Palacios Játiva ¹ , María Camila Reyes ¹, Muhammad Ijaz ² , Cesar Azurdia Meza ³ , David Zabala-Blanco ⁴  and Iván Sanchez ^{5,*} 

¹ Facultad de Ingeniería y Ciencias, Escuela de Informática y Telecomunicaciones, Universidad Diego Portales, Santiago 8370109, Chile; pablo.palacios@mail.udp.cl (P.P.J.)

² Department of Engineering, Faculty of Science and Engineering, Manchester Metropolitan University, Manchester M1 5GD, UK

³ Department of Electrical Engineering, Universidad de Chile, Santiago 8370451, Chile

⁴ Faculty of Engineering Sciences, Universidad Católica Del Maule, Talca 3466706, Chile

⁵ Faculty of Engineering and Applied Sciences, Networking and Telecommunications Engineering, Universidad de Las Américas (UDLA), Quito 170503, Ecuador

* Correspondence: ivan.sanchez.salazar@udla.edu.ec

Abstract

Visible light communication (VLC) can reduce radio-frequency (RF) congestion in indoor industrial monitoring, but a short transmitted payload does not by itself prove that visual or task-relevant information has been preserved. This study therefore frames the proposed method as an autoencoder-based compact latent-payload transmission scheme and explicitly distinguishes it from channel-aware joint source-channel coding (JSCC). Raw red–green–blue (RGB), lossless Huffman, and autoencoder latent payloads are first compared under the same VLC model using bit error rate (BER), calculated/model-derived VLC transmission energy, reconstruction quality, and task utility. The 128-component latent representation contains 4096 bits, corresponding to a 294-fold payload-size reduction relative to an uncompressed 224×224 , 24-bit RGB image; this ratio is used only as a raw-payload reference and not as a general codec-compression claim. An independent industrial-domain audit is conducted on 30 Northeastern University (NEU) steel-surface images using four-fold out-of-fold evaluation, peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and a fixed defect-class proxy. An NEU architecture-and-resolution comparison shows that changing from Compact-NEU to the full 224×224 model increases PSNR from 14.12 dB to 15.61 dB at the same 4096-bit bottleneck, while a regularized full model gives the highest SSIM of 0.471. Because input resolution and network capacity both change in this comparison, the result is interpreted as evidence that the architecture/resolution setting contributes to the industrial-domain gap, not as a strict isolation of capacity alone. Finally, an end-to-end JSCC-VLC baseline with 4096 nonnegative optical channel uses is trained through a differentiable intensity channel. It obtains 23.82 dB/0.568 SSIM in the clean case and 22.18 dB/0.519 SSIM at 5 dB SNR, showing more channel-aware behavior and more graceful degradation than the separated serialized-latent pipeline. Overall, the results support the modeled energy and active-time benefits of compact latent payloads while showing that robust industrial visual transmission requires architecture/resolution controls, practical codec baselines, and channel-aware JSCC comparisons.

Keywords: visible light communication; compact latent payload; convolutional autoencoder; joint source–channel coding; industrial image transmission; energy efficiency



Received: 3 May 2026

Revised: 7 July 2026

Accepted: 8 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Industry 4.0 has increased the demand for wireless links that can support sensing, monitoring, and machine coordination under tight reliability and energy constraints. Factories, warehouses, and automated indoor facilities concentrate robots, sensors, cameras, and control devices in electromagnetically dense spaces, where conventional radio-frequency (RF) systems compete for spectrum and may suffer from interference, congestion, or electromagnetic compatibility restrictions. These conditions motivate the search for complementary short-range technologies that can deliver predictable connectivity without increasing the RF burden [1].

Visible light communication (VLC) is one of the most promising options for such scenarios [2,3]. By modulating LED luminaires already installed for illumination, VLC can provide data transmission over an unlicensed optical spectrum while preserving spatial confinement and intrinsic immunity to electromagnetic interference [4]. Foundational and applied studies have shown that VLC can support indoor links with attractive security, reuse, and coexistence properties, especially in environments where RF emissions are undesirable or heavily constrained [1,5,6]. However, the same environments that motivate VLC also create important design challenges. Indoor industrial links remain sensitive to line-of-sight geometry, distance-dependent attenuation, ambient light, optical reflections, intersymbol interference, and partial blockage caused by machinery or moving elements.

These limitations become more critical when the payload is visual. Industrial inspection, remote supervision, and scene-aware monitoring often require the transmission of images rather than scalar sensor values. A conventional 224×224 RGB image contains more than one million bits before channel coding, which translates into substantial transmission energy and channel occupancy. Classical syntactic schemes can reduce this burden while preserving bit-level fidelity, but they still depend on exact recovery of the encoded stream. In noisy optical links, small bit errors may propagate through compressed payloads and degrade the reconstructed image disproportionately, which makes the trade-off between compression and robustness especially difficult for industrial VLC [7].

Semantic communication changes this design objective [8]. Instead of maximizing bitwise fidelity, it seeks to preserve the information that is relevant for interpretation, perception, or a downstream task. Recent surveys have framed semantic communication as a natural extension of conventional communication theory toward utility-aware transmission [9,10]. For image signals, deep neural networks, especially convolutional autoencoders and related end-to-end models, can learn latent representations that retain salient visual structure with a much smaller payload than direct pixel transmission. Recent work also shows that learned image representations can remain useful under severe bitrate reductions and noisy-channel conditions [11–14].

The application of learned, utility-aware visual representations to VLC is promising and still developing. Existing studies indicate that semantic models can maintain perceptual coherence and tolerate optical impairments better than strictly syntactic approaches [15]. Nevertheless, the literature remains fragmented across three partially disconnected lines of work: industrial VLC studies that focus on physical-layer feasibility, semantic image compression studies that do not model the optical channel explicitly, and end-to-end semantic communication studies developed mainly for RF links. As a result, there is still limited evidence about how learned compact image-payload transmission behaves in a realistic industrial VLC setting when it is compared directly against raw transmission and classical lossless compression under the same channel assumptions.

This paper addresses that gap through a comparative study of three image-payload representations over a common indoor industrial VLC channel: raw RGB transmission, lossless Huffman coding, and a convolutional autoencoder with a 128-component latent

vector. The channel model includes distance variation, background illumination, one optical interferer, intersymbol interference, stochastic blockage, and receiver noise. The impairment set is optical-link oriented and focuses on the channel effects most relevant to the evaluated VLC scenario rather than on a complete electromagnetic-compatibility model of a factory. The objective is deliberately focused: we ask how payload length changes transmission energy and how reconstruction fidelity and a downstream industrial proxy change under payload errors. The learned latent representation is therefore assessed with reconstruction and task metrics in addition to BER and SNR, rather than being treated as automatically equivalent to end-to-end semantic or JSCC transmission.

The contributions are sixfold. First, the three separated payload representations are evaluated under the same framing and VLC assumptions, while actual transmission and processing energy are reported without baseline-normalized “equivalent information” scaling. Second, visual fidelity is quantified using PSNR and SSIM, including image-level bootstrap confidence intervals and paired Wilcoxon tests. Third, a separate industrial-domain audit uses steel-surface defect images and a fixed histogram-of-oriented-gradients support-vector machine (HOG-SVM) as a defect-class proxy to test whether reconstructed images retain information useful for a downstream task. Fourth, payload-error robustness is defined through the slope of reconstruction quality versus logarithmic BER and is interpreted jointly with the absolute quality level, so stable curves are evaluated in the context of their reconstruction fidelity and task utility. Fifth, an architecture-capacity ablation quantifies how much the compact NEU audit network affects the industrial-domain conclusions. Sixth, an end-to-end JSCC-VLC baseline is added to distinguish the proposed separated serialized-latent pipeline from a truly channel-aware learned transmission system.

The comparison is framed as a controlled system study and as a rate-context evaluation rather than as a complete codec benchmark. A rate-matched JPEG reference is included in the industrial audit to contextualize the 4096-bit latent payload. The analysis also distinguishes channel BER from decoder failure: variable-length Huffman desynchronization affects reconstructed-image quality and frame recovery, not the physical definition of BER itself. The resulting interpretation is measured: the current separated autoencoder pipeline provides large reductions in payload size and optical active time, while the industrial evidence and JSCC comparison highlight the need for stronger preservation of fine defects, channel-aware training, and downstream task cues. Task-aware losses, larger industrial training sets, forward-error correction, and hardware validation are natural next steps toward deployment-oriented evaluation.

2. State of the Art

The recent convergence of optical wireless communication and learned signal representation has created a new research space in which VLC, image compression, and semantic communication intersect. The literature already provides strong evidence that each of these areas is individually mature enough to justify integration. What remains less developed is a unified understanding of how they interact under realistic industrial constraints.

2.1. Industrial VLC for Indoor Environments

VLC has been widely proposed as a complement to radio-frequency systems in indoor industrial spaces because it exploits existing illumination infrastructure and offers immunity to electromagnetic interference. Reviews, standards-oriented papers, and channel studies show that industrial deployments can benefit from directional coverage, high spatial reuse, and physical security, but performance is tightly conditioned by optical geometry, receiver field of view, multipath reflections, background light, and dimming constraints [1,5,6,16,17].

In warehouses and production floors, metallic surfaces, moving machinery, and heterogeneous lighting create fluctuations in received power and BER, which means that realistic evaluation must include more than a simple line-of-sight channel model.

2.2. Classical Syntactic Compression for Visual Payloads

When visual data must be transmitted efficiently, conventional systems usually rely on syntactic compression. Lossless approaches such as Huffman coding remain attractive when exact reconstruction is mandatory because they reduce redundancy without introducing distortion [7]. However, these methods still encode the signal at the bit level and therefore remain vulnerable to channel errors. Their efficiency also depends on statistical redundancy rather than semantic relevance, so they cannot prioritize visually meaningful content when the channel becomes unreliable. For industrial VLC, this limitation is important because even modest optical degradation may invalidate an otherwise compact bitstream.

2.3. Semantic Communication for Images

Semantic communication has shifted attention from exact symbol recovery to task-relevant content preservation. Surveys by Xin et al. and Wheeler and Natarajan identify semantic communication as a key research direction for next-generation networks and highlight the role of shared knowledge, context, and neural representations in improving communication efficiency [9,10]. For visual data, deep models such as convolutional autoencoders and learned source-channel coding architectures can compress images into low-dimensional latent spaces that are jointly optimized for reconstruction quality and robustness. Foundational neural transmission studies have also shown that deep joint source-channel coding (JSCC) yields graceful degradation under channel mismatch and low signal-to-noise ratio (SNR) operation, which is closely aligned with the objectives of semantic image delivery [18,19]. Recent image-oriented studies report that semantic-aware compression and end-to-end neural transmission preserve perceptual coherence even at ultra-low bitrates [11–13]. In parallel, reviews of deep learning at the physical layer show that end-to-end models can learn channel-aware modulation, coding, and recovery strategies that outperform rigid analytical pipelines in complex environments [14].

2.4. Semantic Communication over VLC and Remaining Gap

The application of semantic models to VLC is more recent. Chen et al. showed that semantic image transmission over VLC can maintain visual coherence under severe optical degradation, and their results provide early evidence that learned latent representations are compatible with practical optical links [15]. Even so, prior work has not yet established a comprehensive benchmark for industrial indoor VLC. The current literature rarely compares semantic transmission against both raw and lossless syntactic baselines under the same physically realistic channel model, and it seldom combines optical metrics with an explicit evaluation of processing energy. Consequently, an important gap remains: there is still no integrated assessment of the trade-off among payload duration, energy consumption, reconstruction fidelity, and downstream task utility for noisy industrial VLC image links. The present work is designed precisely to fill that gap.

To make this gap more explicit, Table 1 summarizes how representative previous works differ from the present study. Foundational VLC and optical-wireless studies established the channel, modulation, and deployment constraints of indoor optical links, but they did not address semantic image reconstruction or task-oriented visual payloads [5,6,16,17]. Deep JSCC and semantic-communication studies demonstrated graceful degradation and compact learned representations for images, but most of them were developed for generic wireless or abstract channels rather than industrial VLC links [11–13,18,19]. The semantic

VLC work of Chen et al. is the closest optical reference; however, it does not benchmark raw, lossless, and semantic transmission under a unified industrial channel that includes attenuation, background light, optical interference, intersymbol interference (ISI), blockage, receiver noise, and processing energy [15].

Table 1. Positioning of the present work with respect to representative previous studies.

Reference Group	Main Scope and Metrics	Difference from the Present Work
Foundational and industrial VLC [1,5,6,16,17]	Optical channel modeling, indoor VLC feasibility, modulation, dimming, link budget, geometry, and deployment constraints	These works establish the VLC physical-layer basis but do not evaluate semantic image payloads or processing-energy trade-offs.
Deep JSCC and semantic image transmission [11–13,18,19]	Learned image representation, graceful degradation, reconstruction quality, robustness, bitrate reduction, and semantic relevance	These studies motivate learned visual transmission but generally do not include a realistic VLC channel with optical interference, blockage, and receiver noise.
Semantic VLC [15]	Semantic image transmission over VLC using deep generative models and visual robustness under optical degradation	This is the closest prior optical study, but it does not provide a unified comparison against raw and lossless baselines with explicit communication and energy metrics.
Present study	Common VLC link evaluation plus an industrial-domain audit using payload size, BER, actual energy, PSNR, SSIM, quality-degradation slopes, and a downstream defect-class proxy	The study separates communication efficiency from visual/task utility and reports conditions under which a compact latent payload fails to preserve industrial details.

The semantic-aware image compression architecture of Wang et al. [12] is the closest source-representation reference. This comparison does not claim superiority from payload size alone. Instead, the 4096-bit semantic payload is evaluated together with PSNR, SSIM, a rate-matched JPEG reference, and a downstream industrial proxy. This distinction is essential because a low transmitted bit count can reduce optical active time while still discarding fine task-relevant information.

3. Materials and Methods

3.1. Dataset Preparation

The original VLC link and energy campaign used preprocessed ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) natural images [20]. Each image was resized while preserving aspect ratio until the shorter side reached 256 pixels, followed by a centered 224 × 224 RGB crop. This natural-image campaign is used only for link-level and prototype-energy characterization. The 224 × 224 semantic weights for that campaign were trained on the available preprocessed ILSVRC2012 natural-image pool and then fixed before VLC transmission runs. Because the random image selection was not documented as an independent held-out reconstruction test split, those natural-image simulations are used for link-level and energy characterization rather than for PSNR, SSIM, qualitative, or task-effectiveness claims. No claim of industrial task generalization or statistical reconstruction quality is derived from that campaign.

To answer the industrial-domain concern directly, a separate audit was performed with 30 publicly available steel-surface images drawn from the six NEU defect categories: crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches [21]. The images were standardized to 128 × 128 RGB. A four-fold stratified protocol was used, so every image was evaluated out-of-fold exactly once, and no held-out image was used to update the corresponding autoencoder or task classifier. Depending on the fold, 22 or 23 images were used for training and 7 or 8 for testing.

Across the industrial folds, the audit used the same five-stage hidden-layer backbone, Squeeze-and-Excitation (SE) attention, 128-component latent interface, and fold-level training budget. Compact channel widths (8, 16, 32, 64, and 64) were used to limit over-

parameterization on the small industrial subset, and each fold was trained for eight epochs with two deterministic augmentation repetitions per training image. This audit is intentionally framed as a focused diagnostic evaluation rather than a production defect-inspection benchmark. Its role is to determine whether the compact latent representation preserves visual fidelity and downstream utility when measured on industrial textures.

For downstream evaluation, a support vector machine (SVM) classifier using histogram of oriented gradients (HOG) features was trained only on the training images of each fold, with deterministic flips and small rotations, and then frozen before testing [22]. Defect-class accuracy is used as a transparent proxy for task preservation; it is not presented as a state-of-the-art detector. In addition, a JPEG image was generated for each held-out image by choosing the quality factor whose complete file size was closest to 4096 bits [23]. This rate-matched reference contextualizes the semantic payload without implying identical codec structure or resilience.

3.2. System Overview

The complete system follows the same end-to-end workflow for the three evaluated schemes. A preprocessed RGB image is first converted into a mode-dependent payload: uncompressed bytes in the baseline mode, a lossless entropy-coded bitstream in the Huffman mode, or a compact convolutional autoencoder (CAE) latent vector in the semantic mode. The payload is then encapsulated into a common frame structure, modulated with on-off keying non-return-to-zero (OOK-NRZ), propagated through the indoor VLC channel, detected by a photodiode receiver, and finally decoded back into an image representation for evaluation. The common frame includes a preamble, header, and cyclic redundancy check (CRC). Only the source-coding and reconstruction blocks vary across modes; the optical link and receiver chain remain identical. Figure 1 summarizes this shared workflow.

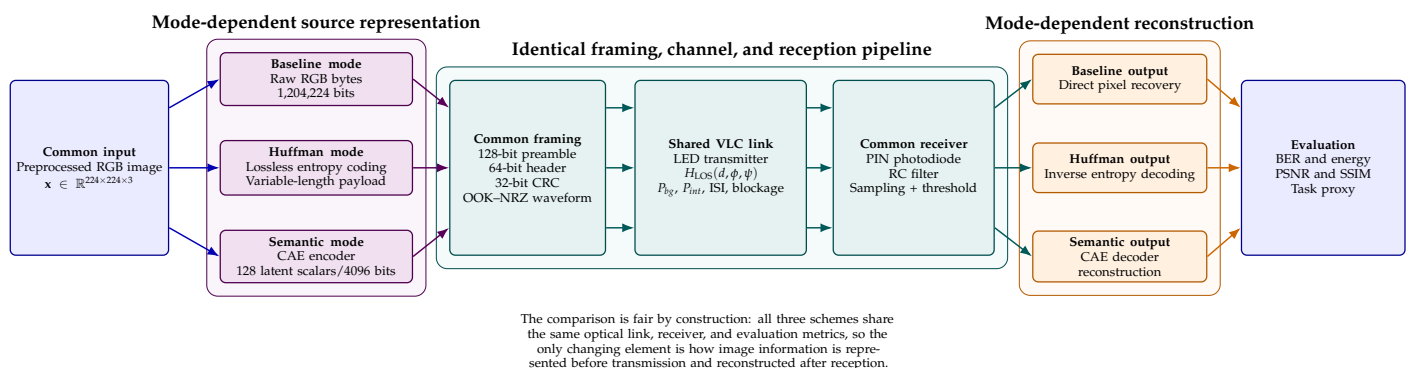


Figure 1. Enhanced end-to-end comparison framework used throughout the study. The three transmission modes start from the same preprocessed image and traverse the same framing, optical VLC link, receiver chain, and evaluation stage.

3.3. Transmitter (TX), Receiver (RX), and VLC Theoretical Model

Figure 2 provides a visual summary of the line-of-sight (LOS) VLC scenario considered in this work, including the light-emitting diode (LED) luminaire, the photodiode receiver, and the main propagation geometry used in the analytical model. In this model, P_{int} denotes one neighboring LED luminaire acting as a non-cooperative optical interferer, whereas P_{bg} denotes non-data-bearing ambient/background illumination that contributes to receiver noise. For readability, the main notation and numerical values introduced in this subsection are summarized in Table 2.

Table 2. Core physical-layer, framing, and semantic-model parameters used in the simulation campaign.

Parameter	Value
Average optical power P_t	0.7 W
Electrical bandwidth B	5 MHz
Bit rate R_b	1 Mb/s
Sampling frequency F_s	20 MHz
Receiver active area A	10^{-4} m^2
Responsivity R_{resp}	0.5 A/W
Refractive index n_{refr}	1.5
Semi-power angle $\phi_{1/2}$	60°
Incidence and irradiance angles	$\psi = 10^\circ, \phi = 10^\circ$
Field of view Ψ_{FOV}	$20^\circ, 30^\circ, 40^\circ, 60^\circ, 80^\circ$
Background optical power P_{bg}	0.2, 0.5, 1, 2 mW
Interfering source power P_{int}	$0.3P_t$ nominal
Interfering geometric factor H_{factor}	0.5 lumped coupling
Shot-noise variance σ_{shot}^2	$2qR_{\text{resp}}P_{\text{tot}}B$
Thermal-noise variance σ_{th}^2	$4k_B T_{\text{rx}} B / R_L$
ISI time constant τ_{ISI}	$2 \times 10^{-7} \text{ s}$
Blockage probability p_{blk}	0.1
Blockage attenuation	15 dB
Distance sweep d	1 to 8 m in 0.5 m steps
Frames per channel setting	20
Image resolution	$224 \times 224 \times 3$
Frame overhead	128-bit preamble + 64-bit header + 32-bit CRC
Baseline payload N_{base}	1,204,224 bits
Semantic latent dimension n_ℓ	128
Semantic payload N_{sem}	4096 bits
Latent scalar format	Institute of Electrical and Electronics Engineers 754 (IEEE-754)
Semantic training data	single precision (32 bits/component)
Optimizer and loss	Preprocessed ILSVRC2012 images
Training schedule	Adam with mean squared error
	20 epochs, batch size 32, learning rate 10^{-3}

The optical transmitter was modeled as a Lambertian LED source using intensity modulation/direct detection (IM/DD) with OOK–NRZ signaling. The transmitted average optical power was set to $P_t = 0.7 \text{ W}$, the electrical bandwidth to $B = 5 \text{ MHz}$, the bit rate to $R_b = 1 \text{ Mb/s}$, and the sampling frequency to $F_s = 20 \text{ MHz}$, which corresponds to an oversampling factor of 20 samples per symbol. At the receiver, a positive–intrinsic–negative (PIN) photodiode with optical concentrator was assumed, with active area $A = 10^{-4} \text{ m}^2$, responsivity $R_{\text{resp}} = 0.5 \text{ A/W}$, refractive index $n_{\text{refr}} = 1.5$, and optical filter gain $T_s = 1$ within the tested field of view.

The line-of-sight channel gain followed a standard Lambertian VLC formulation. The Lambertian order m is obtained from the LED semi-power angle $\phi_{1/2}$ as [24]

$$m = -\frac{\ln 2}{\ln(\cos \phi_{1/2})}. \quad (1)$$

In this work, $\phi_{1/2} = 60^\circ$, which yields $m = 1$. The LOS optical gain H_{LOS} is then written using the TX–RX distance d , irradiance angle ϕ , incidence angle ψ , optical-filter gain $T_s(\psi)$, and optical-concentrator gain $g(\psi)$ as [24]

$$H_{\text{LOS}} = \frac{(m+1)A}{2\pi d^2} \cos^m(\phi) T_s(\psi) g(\psi) \cos(\psi). \quad (2)$$

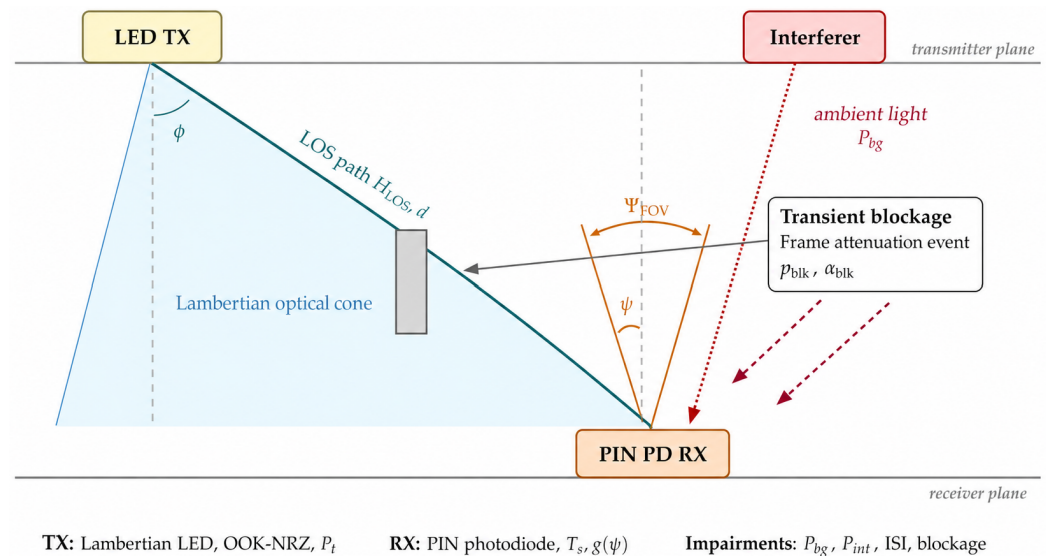


Figure 2. VLC geometry used as the basis of the analytical link model. The intended transmitter is one LED luminaire used for image communication, while P_{int} represents one additional neighboring LED acting as an optical interferer. The background term P_{bg} denotes ambient illumination and is treated as a noise-producing optical power rather than as a second interferer. The schematic also isolates intersymbol interference and a transient blockage event on the LOS path.

The optical-concentrator gain $g(\psi)$ is defined for the receiver field of view Ψ_{FOV} as [24]

$$g(\psi) = \begin{cases} \frac{n_{\text{refr}}^2}{\sin^2 \Psi_{FOV}}, & 0 \leq \psi \leq \Psi_{FOV}, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The received LOS power $P_{rx,LOS}$ is therefore

$$P_{rx,LOS} = H_{LOS}P_t. \quad (4)$$

The theoretical model separates ambient illumination from optical interference. The simulated layout contains two LED luminaires: one desired LED transmitter that carries the image payload and one neighboring non-cooperative LED that acts as the optical interferer. Thus, not all LEDs are simultaneous transmitters for the evaluated image stream; the second LED only stresses the receiver as an external optical source. Background optical power was evaluated at $P_{bg} \in \{0.2, 0.5, 1, 2\}$ mW and was treated as an ambient-noise contribution rather than as a modulated interferer.

The interfering LED was modeled as a received-equivalent optical stress term with nominal power $P_{int} = 0.3P_t$ and a lumped geometric coupling coefficient $H_{\text{factor}} = 0.5$. This coefficient is not a universal VLC constant; it represents the relative attenuation of a neighboring off-axis luminaire at the photodiode after angular displacement, partial field-of-view overlap, fixture placement, and non-aligned incidence are summarized in a single factor. Thus, $H_{\text{factor}}P_{int}$ should be interpreted as the effective received interference power used to stress the common receiver chain. The half-coupling value provides a mid-severity condition in which the interfering LED is weaker than an aligned LOS transmitter but still non-negligible, and it is applied identically to the three transmission modes. The total incident optical power $P_{\text{tot}}(t)$ as a function of time t , with time-dependent desired LOS and interfering optical powers $P_{rx,LOS}(t)$ and $P_{int}(t)$, becomes

$$P_{\text{tot}}(t) \approx P_{rx,LOS}(t) + H_{\text{factor}}P_{int}(t) + P_{bg}. \quad (5)$$

After photoelectric conversion, the received signal current $i_{\text{sig}}(t)$ is

$$i_{\text{sig}}(t) = R_{\text{resp}} P_{\text{tot}}(t). \quad (6)$$

For discrete sample index k , the current samples at the receiver were corrupted by independent shot and thermal current-noise components. Let $\mathcal{N}(\mu, \sigma^2)$ denote a Gaussian distribution with mean μ and variance σ^2 . Following standard IM/DD VLC receiver models [5,6], the shot-noise sample $n_{\text{shot}}[k]$ and its variance σ_{shot}^2 were represented using electron charge q as

$$n_{\text{shot}}[k] \sim \mathcal{N}(0, \sigma_{\text{shot}}^2), \quad \sigma_{\text{shot}}^2 = 2qR_{\text{resp}}P_{\text{tot}}B. \quad (7)$$

This term is linked to the average optical power incident on the photodiode: the desired LED, the interfering LED, and the ambient/background optical power all contribute to photon shot noise through P_{tot} . Consequently, P_{bg} is not modeled as a coherent interference waveform; it is a DC optical background that increases the receiver noise variance.

The thermal-noise sample $n_{\text{th}}[k]$ and its variance σ_{th}^2 were modeled explicitly as additive zero-mean Gaussian current noise using Boltzmann's constant k_B , receiver noise temperature T_{rx} , and equivalent receiver load resistance R_L :

$$n_{\text{th}}[k] \sim \mathcal{N}(0, \sigma_{\text{th}}^2), \quad \sigma_{\text{th}}^2 = \frac{4k_B T_{\text{rx}} B}{R_L}. \quad (8)$$

The noisy receiver current $i_{\text{rx}}[k]$ before filtering and detection is therefore

$$i_{\text{rx}}[k] = i_{\text{sig}}[k] + n_{\text{shot}}[k] + n_{\text{th}}[k]. \quad (9)$$

Intersymbol interference was introduced through a first-order discrete resistance-capacitance (RC) filter with time constant $\tau_{\text{ISI}} = 2 \times 10^{-7}$ s. Let $x[k]$ denote the noisy current sample before filtering, $y[k]$ the filtered sample, and α the discrete RC smoothing coefficient:

$$y[k] = \alpha y[k-1] + (1 - \alpha)x[k]. \quad (10)$$

Here $\alpha = \exp[-1/(F_s \tau_{\text{ISI}})]$ for sampling frequency F_s .

To emulate intermittent obstruction by moving workers or objects, stochastic blockage was modeled as a frame-level attenuation of the optical gain. The effective gain H_{eff} is defined using blockage probability p_{blk} and blockage attenuation factor α_{blk} as

$$H_{\text{eff}} = \begin{cases} H_{\text{LOS}}, & \text{with probability } 1 - p_{\text{blk}}, \\ \alpha_{\text{blk}} H_{\text{LOS}}, & \text{with probability } p_{\text{blk}}, \end{cases} \quad (11)$$

with $p_{\text{blk}} = 0.1$ and $\alpha_{\text{blk}} = 10^{-15/10}$, corresponding to a 15 dB attenuation event.

3.4. Simulation Design and Parameter Configuration

The numerical campaign combined a full-factorial sweep of the main VLC stressors with a shared end-to-end transmission pipeline for the three coding modes. The link distance was varied from 1 to 8 m in 0.5 m steps, the receiver field of view covered narrow to wide settings, and the background illumination ranged from dim to strongly contaminated conditions. For each channel configuration, 20 independent frames were simulated in order to average the effects of shot noise, thermal noise, ISI, optical interference, and stochastic blockage.

Table 2 summarizes the physical-layer, framing, and semantic-model parameters used throughout the study. Presenting these values together makes the simulation design explicit and facilitates the interpretation of the subsequent comparative results.

3.5. Frame Structure and Execution Protocol

Building on the full-factorial configuration sweep described above, each simulation run treated one image as one transmitted frame. For every channel realization defined by $(d, \Psi_{\text{FOV}}, P_{bg})$, 20 independent frame transmissions were executed so that noise, optical interference, intersymbol interference, and stochastic blockage were re-sampled across trials. This frame-based protocol is important because it moves the comparison from a purely analytical description of the VLC channel to a repeated end-to-end transmission experiment under controlled but random conditions.

All transmission modes shared the same frame envelope in order to preserve fairness at the communication level. Each frame contained a 128-bit preamble, a 64-bit header, and a 32-bit CRC, which gives a fixed overhead N_{oh} of

$$N_{\text{oh}} = 128 + 64 + 32 = 224 \text{ bits.} \quad (12)$$

For mode-dependent payload size N_{payload} , the total transmitted frame size N_{frame} was therefore

$$N_{\text{frame}} = N_{\text{oh}} + N_{\text{payload}}, \quad (13)$$

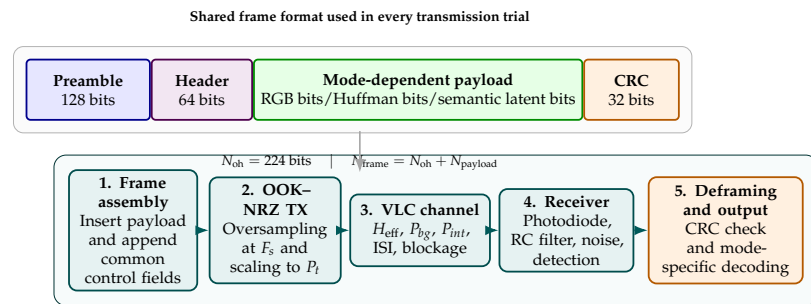
where N_{payload} can represent direct RGB bits in baseline mode, a variable-length Huffman bitstream in Huffman mode, or a 4096-bit latent representation in semantic mode. Keeping the preamble, control information, and integrity field identical across schemes ensures that any observed difference is attributable to the payload representation and the decoding strategy rather than to unequal synchronization or signaling assumptions.

Figure 3 summarizes this shared frame structure and the execution chain applied after payload generation. The upper part of the diagram highlights the fixed framing overhead and the single mode-dependent field, whereas the lower part shows the common physical-layer and receiver-side processing pipeline.

As summarized in Figure 3, once the frame was assembled, the same physical-layer execution chain was applied to all modes:

1. Generation of an oversampled OOK–NRZ waveform at sampling frequency F_s from the complete binary frame;
2. Scaling of the waveform to satisfy the transmitted optical power constraint P_t ;
3. Application of the effective VLC gain H_{eff} , including possible blockage-induced attenuation;
4. Addition of background illumination and optical interference;
5. Conversion of received optical power into current through the photodiode responsivity;
6. RC filtering and additive noise injection to emulate receiver impairments and ISI;
7. Symbol-rate sampling and threshold-based detection to recover the received bitstream;
8. Removal of the framing fields and delivery of the recovered payload to the mode-specific decoder.

This ordered protocol clarifies the role of framing in the study. The baseline mode reshapes the recovered payload directly into pixels, the Huffman mode attempts exact lossless reconstruction from a synchronization-sensitive bitstream, and the semantic mode deserializes the latent vector before neural decoding. Consequently, the comparative results reflect not only differences in payload size but also the distinct ways in which channel errors propagate through each representation under identical VLC conditions.



Only the payload field changes across baseline, Huffman, and semantic transmission; framing, optical propagation, detection, and payload delivery remain common.

Figure 3. Shared frame composition and execution protocol. The preamble, header, and CRC define a constant overhead for all schemes, while the payload field changes according to the evaluated transmission mode.

3.6. Transmission Modes

3.6.1. Baseline Mode

In the baseline mode, the preprocessed RGB image was vectorized and transmitted directly without compression or semantic extraction. Since each color component was represented with 8 bits, the baseline payload size N_{base} was

$$N_{base} = 224 \times 224 \times 3 \times 8 = 1,204,224 \text{ bits.} \quad (14)$$

This mode represents a purely syntactic transmission strategy that preserves the complete pixel stream, maximizes channel occupancy, and establishes the reference case for energy consumption, goodput, and robustness.

3.6.2. Huffman Mode

In the Huffman mode, the RGB values were vectorized into 8-bit symbols, the empirical byte histogram was computed, and an optimal Huffman dictionary was constructed for each image. The resulting variable-length code minimized the average code length while preserving exact recoverability. To focus on the effect of payload compression, the dictionary was assumed to be available at the receiver through prior signaling or shared storage. Consequently, only the compressed payload was transmitted. Under a fixed bit rate, this shorter payload reduced frame duration and transmission energy relative to the baseline mode. However, because Huffman decoding depends on correct synchronization, the mode remained highly sensitive to channel-induced bit errors.

3.6.3. Semantic Mode

The semantic mode replaced direct pixel transmission with a convolutional autoencoder with Squeeze-and-Excitation channel-attention blocks [25]. During operation, each preprocessed image was normalized to the interval $[0, 1]$, encoded into a latent vector, serialized into a bitstream, transmitted over the VLC link, deserialized at the receiver, and decoded back into a reconstructed image.

For the original 224×224 VLC link-level campaign, the encoder comprised five convolutional stages with progressive spatial downsampling and increasing channel depth. The spatial resolution evolved from 224×224 to 112×112 , 56×56 , 28×28 , 14×14 , and finally 7×7 . At each stage, SE attention recalibrated the feature channels by emphasizing informative responses and suppressing less relevant ones. The final $512 \times 7 \times 7$ tensor was flattened and linearly projected into a latent space of dimension $n_\ell = 128$. The decoder mirrored this structure by expanding the latent vector back to $512 \times 7 \times 7$ and then applying successive upsampling and deconvolutional stages until the original image resolution was restored. The separate 128×128 industrial audit used the compact fold-trained variant

described in the dataset subsection, preserving the 128-component latent interface rather than reusing the exact 224×224 channel configuration.

In the original VLC link-level campaign, each latent component was serialized using the Institute of Electrical and Electronics Engineers 754 (IEEE-754) single-precision floating-point format, which corresponds to 32 bits per component. The semantic payload size N_{sem} was therefore

$$N_{\text{sem}} = 128 \times 32 = 4096 \text{ bits}, \quad (15)$$

which yields the payload-size reduction factor ρ_{payload} relative to uncompressed 24-bit RGB:

$$\rho_{\text{payload}} = \frac{1,204,224}{4096} \approx 294. \quad (16)$$

This value is a payload-size ratio against uncompressed RGB, not a compression ratio against JPEG, JPEG2000, or another practical image codec. The reduction in payload size lowers transmission duration and spectral occupancy, but its usefulness must be assessed jointly with reconstruction and task metrics. The 224×224 semantic model used in the original link-level campaign was trained on the available preprocessed ILSVRC2012 natural-image pool using the Adam optimizer [26] and mean squared error loss, with 20 epochs, a batch size of 32, and an initial learning rate of 10^{-3} . The best-performing weights were then fixed before the natural-image VLC simulations. Because the randomly selected natural-image VLC frames were not documented as an independent held-out reconstruction test split, that campaign is used for link-level and energy analysis rather than for independent PSNR, SSIM, qualitative, or task-preservation claims.

For the industrial audit, each latent component was serialized with a per-component affine mapping to an unsigned 32-bit representation calibrated only on the training fold. This preserves the 4096-bit payload while bounding decoded values and avoiding not-a-number (NaN) or infinity events that can arise when arbitrary channel errors alter IEEE-754 exponent fields.

3.7. Architecture-and-Resolution Comparison for the NEU Audit

An architecture-and-resolution comparison was added in Table 3 to clarify the effect of using a compact industrial-audit network rather than the full natural-image architecture. The compact NEU audit model used a 128×128 input with channel widths (8, 16, 32, 64, 64), whereas the original natural-image link-level campaign used a wider 224×224 five-stage encoder–decoder whose final feature tensor was $512 \times 7 \times 7$. Since both the input resolution and the network width/depth change between Compact-NEU and Full-NEU, this experiment is not described as strictly isolating model capacity. It is instead interpreted as a combined architecture-and-resolution comparison under a fixed latent-payload interface.

The comparison evaluates three encoder–decoder variants on the same 30-image NEU manifest and four out-of-fold splits while keeping the transmitted interface fixed at $n_\ell = 128$ latent components, or 4096 serialized bits. Compact-NEU is the compact audit network. Full-NEU uses the 224×224 five-stage channel progression (32, 64, 128, 256, 512) used in the natural-image model. Full-NEU-Reg keeps the same full architecture but adds stronger regularization through larger weight decay, a 5% intermediate-stage dropout, and fold-local data augmentation. Thus, the payload size, manifest, and fold protocol are held constant, while input resolution and architecture/capacity change together.

Table 3. Architecture-and-resolution comparison for the NEU audit. All variants keep the same 128-component/4096-bit latent payload, while Compact-NEU and Full-NEU differ in both input resolution and network capacity.

Model	Input	Latent	Channel Widths	Trainable Parameters
Compact-NEU	128×128	128	8–16–32–64–64	487,307
Full-NEU	224×224	128	32–64–128–256–512	12,083,699
Full-NEU-Reg	224×224	128	32–64–128–256–512	12,083,699

For each fold, the autoencoder is trained only on the training portion of the NEU split and evaluated on the held-out images from that fold. This architecture-and-resolution comparison reports clean-reconstruction PSNR and SSIM on the identical manifest images after fold-wise training. The image remains the independent statistical unit, and aggregate means and standard deviations are computed across the 30 held-out reconstructions from the four folds.

This comparison is interpreted conservatively. If Full-NEU or Full-NEU-Reg improves fidelity over Compact-NEU, then part of the industrial fidelity limitation may be associated with the combined change in architecture and input resolution. If the full-resolution variants remain close to the compact model, the result would suggest that other factors, such as the fixed 4096-bit bottleneck, reconstruction-only training, and limited industrial-domain data, remain plausible contributing limitations. In either case, low industrial fidelity is not attributed solely to domain shift or loss design without acknowledging the architecture/resolution change.

3.8. End-to-End JSCC-VLC Baseline

To distinguish compact latent-payload transmission from true channel-aware semantic communication, the study includes an end-to-end JSCC-VLC baseline. This baseline is not the proposed separated pipeline. Instead, it is a methodological control in which the encoder, optical channel layer, and decoder are trained jointly.

For an input image $\mathbf{x}$ resized to 128×128 , the JSCC encoder $f_\theta(\cdot)$ uses four stride-two convolutional stages and produces a tensor of size $64 \times 8 \times 8$, corresponding to $K = 4096$ real-valued channel uses. To respect the nonnegative intensity constraint of IM/DD VLC, the transmitted optical-symbol tensor is

$$\mathbf{s} = \sigma(f_\theta(\mathbf{x})), \quad 0 \leq s_k \leq 1, \quad (17)$$

where $\sigma(\cdot)$ is the logistic function. A differentiable scalar optical-intensity channel is then applied as

$$\mathbf{y} = \text{clip}(\mathbf{s} + \mathbf{n}, 0, 1), \quad n_k \sim \mathcal{N}\left(0, \frac{P_s}{10^{\text{SNR}/10}}\right), \quad (18)$$

where $P_s = \mathbb{E}[s_k^2]$ is the average symbol power within the mini-batch. During training, the SNR is sampled uniformly from 8 to 20 dB, so the decoder $g_\phi(\cdot)$ learns to reconstruct from noisy optical intensities rather than from a separately serialized bitstream. During evaluation, the same model is tested in the clean case and at SNR values of 20, 10, and 5 dB.

The JSCC-VLC baseline uses the same four NEU out-of-fold splits and the same 30 manifest images as the industrial audit. The comparison is controlled by channel uses rather than file bits: the separated latent mode transmits 4096 OOK bits, whereas JSCC-VLC uses 4096 real-valued nonnegative optical channel uses. Therefore, the JSCC result is not interpreted as a direct codec-compression comparison, but as a channel-aware baseline showing what changes when the representation is optimized jointly with the channel. The training loss is the same reconstruction loss family used in the ablation, combining an

ℓ_1 image term and a gradient-consistency term. PSNR, SSIM, and the fixed HOG–SVM defect-class proxy are computed on the held-out fold images.

3.9. MATLAB–Python Integration

The optical-link simulation was executed in MATLAB 2025B, whereas the semantic encoder and decoder were implemented in Python 3.14.3. During the natural-image link-level campaign, MATLAB invoked the external Python module to encode an image into a 128-dimensional latent vector and stored the result in a .npy file. The latent vector was then converted to bits using IEEE-754 single-precision packing. At the receiver, the detected bitstream was converted back into a latent vector and sent again to Python for decoding. This workflow maintained strict consistency between the trained semantic model and the physical-layer VLC simulator. The industrial audit used the bounded affine-mapped unsigned representation stated above to keep corrupted latent values finite and fold-calibrated.

The experiments were run on a workstation equipped with an Intel Core i5-9600K central processing unit (CPU), 32 GB random-access memory (RAM), and an NVIDIA RTX 3070 graphics processing unit (GPU). GPU acceleration was used for semantic training and inference, whereas the VLC channel simulation and receiver processing remained CPU-based. Encoding and decoding times were measured with MATLAB timing functions and later used to estimate computational energy.

The reported semantic processing energy of 0.85 J per image should therefore be interpreted as an implementation-specific prototype value. It corresponds to batch-one inference on 224×224 images in the workstation-based MATLAB–Python workflow and includes invocation, serialization, and encoder/decoder runtime. For deployment platform $\mathcal{H}$, input height H , input width W , inference batch size b , active processing power $P_{\mathcal{H}}$, fixed overhead time $t_0(\mathcal{H})$, and per-pixel processing-time coefficient $\kappa(\mathcal{H})$, a first-order per-image processing-energy model $E_{\text{proc}}^{(\text{img})}$ is

$$E_{\text{proc}}^{(\text{img})}(H, W, b, \mathcal{H}) \approx P_{\mathcal{H}} \left[\frac{t_0(\mathcal{H})}{b} + \kappa(\mathcal{H})HW \right]. \quad (19)$$

For the fixed convolutional encoder–decoder used here, the dominant workload scales approximately with the number of pixels; thus, increasing both image dimensions by a factor of two would increase the dominant convolutional cost by roughly four times before any architectural redesign. Larger batches can reduce the amortized t_0/b term and improve accelerator utilization, but they also increase buffering, memory demand, and end-to-end latency, which may be undesirable for real-time monitoring. As a deployment-oriented estimate, the fixed MATLAB–Python overhead can be removed, and the processing energy can be approximated by $E_{\text{proc}}^{(\text{img})} \approx P_{\mathcal{H}}t_{\text{inf}}$, where t_{inf} is the optimized per-image encoder–decoder inference time on the embedded platform. For a representative edge-AI power envelope, NVIDIA Jetson Orin Nano power modes include 7 W and 15 W operation, with higher-power configurations also available in the Orin Nano family [27]. Adding the semantic optical-transmission energy of 0.01 J, an inference time of 20–100 ms gives approximately 0.15–0.71 J at 7 W, 0.31–1.51 J at 15 W, and 0.51–2.51 J at 25 W. These values are estimates, not measurements on the target hardware.

3.10. Evaluation Metrics

Performance was first characterized at the link level. For the number of erroneous detected bits N_{err} and transmitted detected bits N_{tot} , the bit error rate was

$$\text{BER} = \frac{N_{\text{err}}}{N_{\text{tot}}}, \quad (20)$$

Using total per-image latency L_{total} , optical frame-transmission duration L_{TX} , receiver and link-processing latency L_{proc} , encoder and decoder runtimes t_{enc} and t_{dec} , electrical-to-optical conversion efficiency η_{EO} , CPU power P_{CPU} , transmission energy E_{TX} , processing energy E_{CPU} , and total energy E_{total} , end-to-end latency and energy were computed as

$$L_{\text{total}} = L_{\text{TX}} + L_{\text{proc}} + t_{\text{enc}} + t_{\text{dec}}, \quad (21)$$

$$E_{\text{TX}} = \frac{P_t}{\eta_{\text{EO}}} L_{\text{TX}}, \quad E_{\text{CPU}} = (L_{\text{proc}} + t_{\text{enc}} + t_{\text{dec}}) P_{\text{CPU}}, \quad E_{\text{total}} = E_{\text{TX}} + E_{\text{CPU}}. \quad (22)$$

A baseline-normalized equivalent-energy metric is avoided because multiplying by the raw-payload ratio would implicitly assume that all modes deliver the same quantity of useful information. Energy and visual/task quality are therefore reported separately.

Image fidelity was quantified for every held-out industrial reconstruction using PSNR and SSIM [28]. For an $M \times N$ image with maximum normalized value one, reference pixel x_{ijc} , reconstructed pixel $\hat{x}_{ijc}$ at row i , column j , and RGB channel c , and mean squared error (MSE),

$$\text{PSNR} = 10 \log_{10} \left(\frac{1}{\text{MSE}} \right), \quad \text{MSE} = \frac{1}{3MN} \sum_{c=1}^3 \sum_{i=1}^M \sum_{j=1}^N (x_{ijc} - \hat{x}_{ijc})^2. \quad (23)$$

SSIM was computed over local luminance, contrast, and structure terms and averaged over the image. For N_{img} evaluated images with true label y_i , predicted label $\hat{y}_i$, and indicator function $\mathbb{I}(\cdot)$, defect-class accuracy A_{task} was

$$A_{\text{task}} = \frac{1}{N_{\text{img}}} \sum_{i=1}^{N_{\text{img}}} \mathbb{I}(\hat{y}_i = y_i). \quad (24)$$

To quantify degradation rather than relying on a qualitative figure, a robustness slope β_Q was defined from mean metric values $\bar{Q}$ for metric $Q \in \{\text{PSNR}, \text{SSIM}, A_{\text{task}}\}$ over the nonzero BER points:

$$\beta_Q = \frac{d\bar{Q}}{d \log_{10}(\text{BER})}. \quad (25)$$

Thus, β_Q denotes the change in mean metric per decade of BER. A small $|\beta_Q|$ is therefore reported together with the absolute mean quality, since a decoder output dominated by low-frequency structure can also exhibit a nearly zero slope.

The nonzero-BER audit used three independent corruption realizations at 10^{-4} , 10^{-3} , and 10^{-2} ; trials were first averaged within each held-out image. Means and standard deviations were then computed across the 30 independent out-of-fold images. Ninety-five percent confidence intervals were obtained with 5000 image-level bootstrap resamples, and paired semantic-versus-baseline and semantic-versus-Huffman comparisons used two-sided Wilcoxon signed-rank tests. Payload errors were injected after successful preamble/header acquisition. The CRC was used as an error flag, but no retransmission was applied because the objective was to measure distortion after corruption. Huffman decoding was performed on a best-effort basis using the shared per-image dictionary; any incomplete decoded pixel vector was zero-filled to its expected length. These policies are now stated explicitly to make the visual comparison reproducible.

4. Results

4.1. Link-Level Payload and Energy Results

Table 4 reports the link-level averages without using cross-mode SNR differences or a baseline-normalized equivalent-energy metric. Payload duration changes frame energy,

but it should not by itself be interpreted as changing the instantaneous received SNR under the same optical power and channel. Similarly, Huffman desynchronization is treated as a reconstruction failure mechanism rather than as the cause of physical BER.

Table 4. Link-level and prototype-energy summary. Total energy is reported directly, without equivalent-information normalization.

Mode	Average BER	Payload [bit]	E _{CPU} [J]	E _{TX} [J]	E _{total} [J]
Baseline	0.093	1,204,224	0.00	2.81	2.81
Huffman	0.108	variable	0.38	2.73	3.11
Semantic	0.094	4096	0.85	0.01	0.86

The 4096-bit latent interface therefore reduces payload size by a factor of approximately 294 relative to direct 224×224 RGB transmission and reduces the model-derived optical transmission energy from 2.81 J to 0.01 J. After the prototype processing cost is included, total energy is 0.86 J for semantic transmission, 2.81 J for baseline, and 3.11 J for Huffman. This communication-resource result is not interpreted as a 294-fold efficiency advantage over practical image codecs; it is interpreted together with the rate-matched JPEG reference and the visual and task metrics reported below.

4.2. Industrial Clean-Payload Fidelity and Rate-Matched Context

Table 5 reports the out-of-fold industrial audit before channel errors. Raw RGB and Huffman are lossless. Because comparison against uncompressed RGB alone is insufficient for judging compression performance, the semantic and JPEG rows have nearly the same payload size, enabling a direct rate-context comparison with a practical classical codec. This table reports the optimized clean-payload industrial audit, including the 18.59 dB separated-latent result. The separate architecture-and-resolution comparison in Section 4.3 then evaluates the compact and full-resolution models under a unified retraining protocol.

Table 5. Clean-payload industrial audit over 30 out-of-fold NEU images. Values are mean $\pm$ standard deviation across images.

Representation	Payload [bit]	PSNR [dB]	SSIM	A _{task}
Raw RGB	393,216	exact	1.000	0.733
Huffman	313,884 $\pm$ 37,823	exact	1.000	0.733
JPEG near 4 kbit	4119 $\pm$ 76	30.77 $\pm$ 6.62	0.699 $\pm$ 0.169	0.167
Semantic, 4096 bits	4096	18.59 $\pm$ 3.94	0.452 $\pm$ 0.212	0.100

The rate-matched JPEG reference obtains higher PSNR and SSIM than the compact semantic audit model. The separated latent output achieves 0.10 defect-class accuracy, compared with 0.733 for the uncorrupted lossless representations. These results explicitly limit the claim: the compact latent interface is promising for reducing transmitted payload and optical active time, but the present reconstruction-only model is not superior to a rate-matched JPEG reference in fidelity or task preservation. Defect-relevant representation learning therefore remains an important target for improvement.

4.3. Architecture-and-Resolution Comparison for the Industrial Audit

Table 6 shows that the compact NEU audit network and the original full 224×224 architecture differ both in input resolution and in trainable-parameter count. The comparison therefore should not be interpreted as a strict single-factor isolation of model capacity. Instead, it keeps the latent payload fixed and compares Compact-NEU, Full-NEU, and Full-NEU-Reg on the same fold splits to assess whether the combined architecture/resolution setting affects industrial-domain fidelity.

The numerical results are reported in Table 6, and the mean SSIM trend is visualized in Figure 4. This experiment is a separate fold-wise retraining analysis designed to compare the compact and full-resolution architectures under a common manifest and split protocol; it complements, rather than replaces, the optimized clean-payload industrial audit in Table 5. Under this retraining protocol, Full-NEU improves mean PSNR from 14.12 dB for Compact-NEU to 15.61 dB, a gain of 1.49 dB at the same 4096-bit payload. Full-NEU-Reg yields 14.48 dB PSNR and the highest mean SSIM, 0.471, compared with 0.464 for Compact-NEU and 0.443 for Full-NEU. These results indicate that the combined change in input resolution and network architecture affects the reported industrial fidelity. However, because resolution and capacity are not independently ablated, the gain cannot be assigned exclusively to model capacity. The larger remaining gap to the approximately rate-matched JPEG reference (30.77 dB and 0.699 SSIM) is consistent with the fixed 4096-bit bottleneck and reconstruction-only training objective being important contributing limitations, but those factors are not claimed to be independently proven as dominant causes.

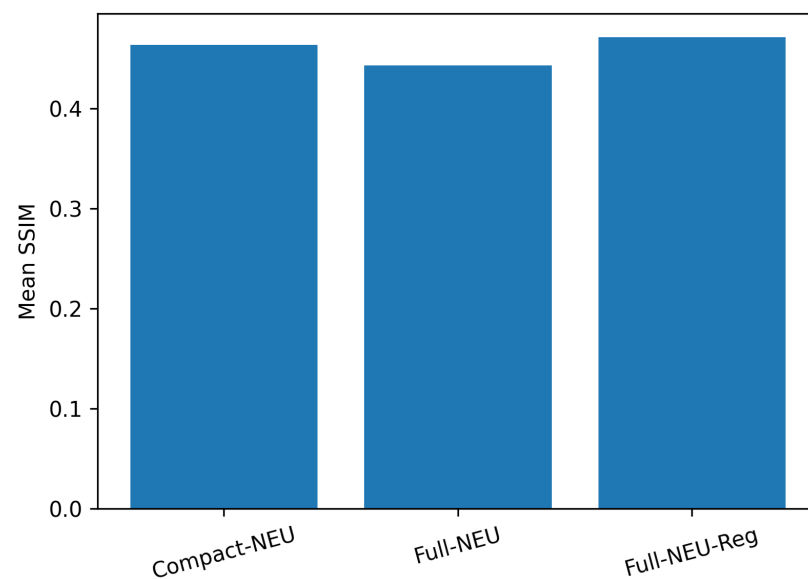


Figure 4. Mean SSIM for the three NEU variants evaluated on the same industrial manifest with a fixed 4096-bit payload. Compact-NEU and the full models differ in both input resolution and network capacity; therefore, the comparison is interpreted as an architecture-and-resolution control rather than as a strict capacity-only ablation.

Table 6. Clean reconstruction results for the NEU architecture-and-resolution comparison with a fixed 128-component/4096-bit latent payload. Values are mean $\pm$ standard deviation over the 30 held-out images from the four folds.

Model	Input	Params	Payload Bits	PSNR [dB]	SSIM
Compact-NEU	128 $\times$ 128	487,307	4096	14.12 $\pm$ 4.25	0.464 $\pm$ 0.224
Full-NEU	224 $\times$ 224	12,083,699	4096	15.61 $\pm$ 3.59	0.443 $\pm$ 0.252
Full-NEU-Reg	224 $\times$ 224	12,083,699	4096	14.48 $\pm$ 3.73	0.471 $\pm$ 0.231

4.4. End-to-End JSCC-VLC Baseline Results

Table 7 reports the end-to-end JSCC-VLC baseline. In the clean case, JSCC-VLC reaches 23.82 dB PSNR and 0.568 SSIM with 4096 nonnegative optical channel uses. At 20 dB SNR, the result is nearly unchanged, with 23.73 dB and 0.566 SSIM. At 10 dB and 5 dB, quality decreases gradually to 23.13 dB/0.548 SSIM and 22.18 dB/0.519 SSIM, respectively. Figure 5 shows the corresponding SSIM trend.

These results clarify an important methodological distinction. The proposed serialized latent method is a compact separated source/channel pipeline: the autoencoder is trained without the optical channel, and the latent values are converted into bits for OOK transmission. JSCC-VLC, in contrast, is trained through the noisy intensity channel and therefore learns a representation that is explicitly channel-aware. With channel-use budgets of the same order, the JSCC-VLC baseline provides higher reconstruction quality than the separated compact latent reconstruction reported in the industrial audit, and it degrades gracefully as SNR decreases. However, its fixed proxy task accuracy remains 0.167, which is chance level for the six-class NEU subset, so this baseline should not be interpreted as solving defect-aware semantic recognition. This supports the conclusion that the separated latent pipeline should not be interpreted as equivalent to JSCC. The analysis therefore describes the main contribution as compact autoencoder-based payload transmission and treats JSCC-VLC as a channel-aware reference baseline.

Table 7. End-to-end JSCC-VLC baseline on the NEU industrial manifest. The model uses 4096 nonnegative optical channel uses and is trained through a differentiable intensity channel with SNR sampled uniformly from 8 to 20 dB. Values are mean $\pm$ standard deviation over the 30 held-out images.

Evaluation SNR	Channel Uses	PSNR [dB]	SSIM	A_{task}
Clean	4096	23.82 ± 4.50	0.568 ± 0.199	0.167
20 dB	4096	23.73 ± 4.39	0.566 ± 0.198	0.167
10 dB	4096	23.13 ± 3.98	0.548 ± 0.189	0.167
5 dB	4096	22.18 ± 3.47	0.519 ± 0.175	0.167

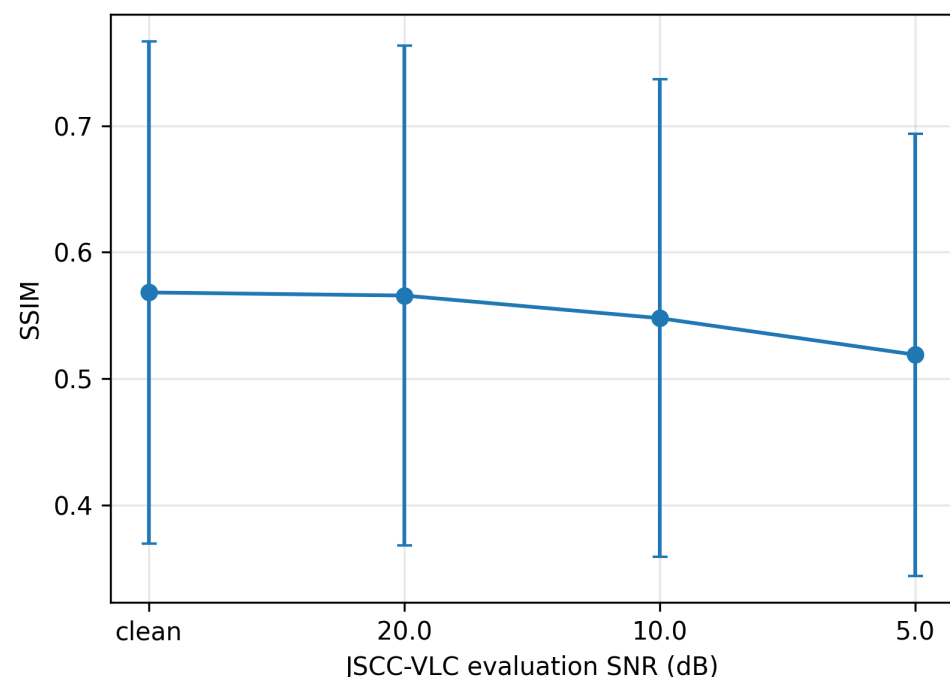


Figure 5. SSIM of the end-to-end JSCC-VLC baseline versus evaluation SNR. The baseline is trained through a differentiable nonnegative optical-intensity channel and therefore represents a channel-aware reference, not a separated source/channel payload.

4.5. Reconstruction Quality and Task Utility Under Payload Errors

Table 8 provides PSNR, SSIM, and downstream accuracy for every tested BER. Figure 6 visualizes the SSIM trends, while Figure 7 reports the task proxy.

Table 8. Industrial reconstruction and task metrics versus post-detection BER. Values are image-level mean $\pm$ standard deviation after averaging the three corruption trials at nonzero BER.

BER	Mode	PSNR [dB]	SSIM	A_{task}
0	Baseline	exact	1.000	0.733
0	Huffman	exact	1.000	0.733
0	Semantic	18.59 ± 3.90	0.452 ± 0.213	0.100
10^{-4}	Baseline	45.06 ± 0.97	0.995 ± 0.002	0.700
10^{-4}	Huffman	30.09 ± 6.04	0.768 ± 0.152	0.544
10^{-4}	Semantic	18.59 ± 3.90	0.452 ± 0.213	0.100
10^{-3}	Baseline	34.82 ± 0.29	0.948 ± 0.019	0.678
10^{-3}	Huffman	23.16 ± 5.27	0.455 ± 0.208	0.533
10^{-3}	Semantic	18.60 ± 3.90	0.452 ± 0.213	0.089
10^{-2}	Baseline	24.73 ± 0.09	0.617 ± 0.114	0.522
10^{-2}	Huffman	19.19 ± 5.26	0.218 ± 0.177	0.311
10^{-2}	Semantic	18.58 ± 3.88	0.452 ± 0.213	0.100

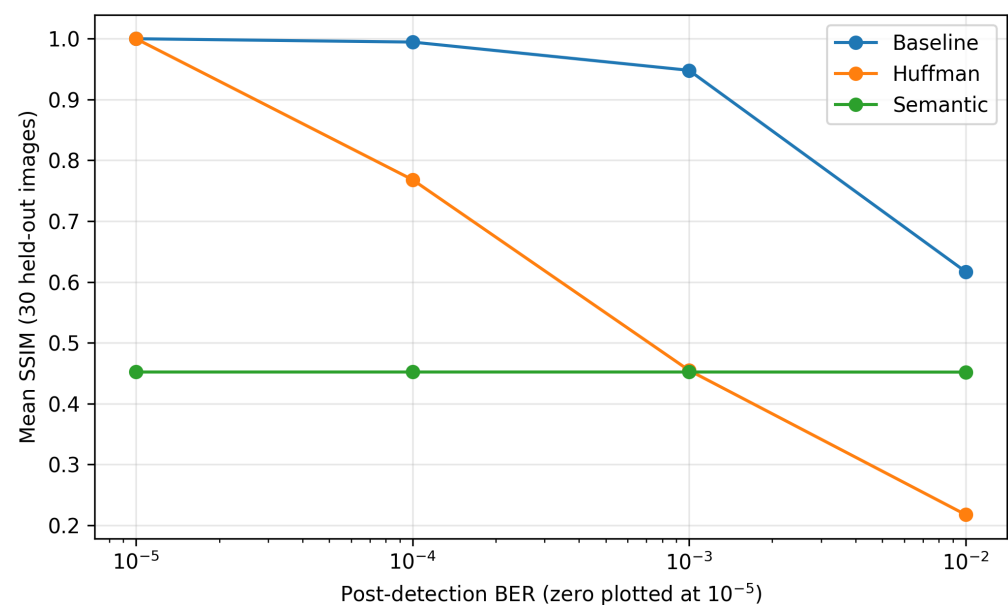


Figure 6. Mean SSIM versus post-detection BER for 30 out-of-fold industrial images. The semantic curve is nearly flat because its clean reconstruction emphasizes low-frequency structure; the flat slope is therefore interpreted together with absolute quality and task metrics.

The paired tests support the same interpretation. At BER 10^{-3} , semantic PSNR and SSIM are both lower than baseline with $p < 2 \times 10^{-9}$. Semantic and Huffman SSIM are statistically indistinguishable at that point ($p = 0.968$), and both values remain around 0.45, which indicates modest absolute fidelity in this audit. At BER 10^{-2} , semantic SSIM is lower than baseline ($p = 0.0197$) but higher than the severely desynchronized Huffman output ($p < 4 \times 10^{-9}$). Semantic task accuracy is lower than baseline at all tested BER values ($p \leq 0.0014$). Full bootstrap intervals and paired-test outputs are provided as Supplementary Files.

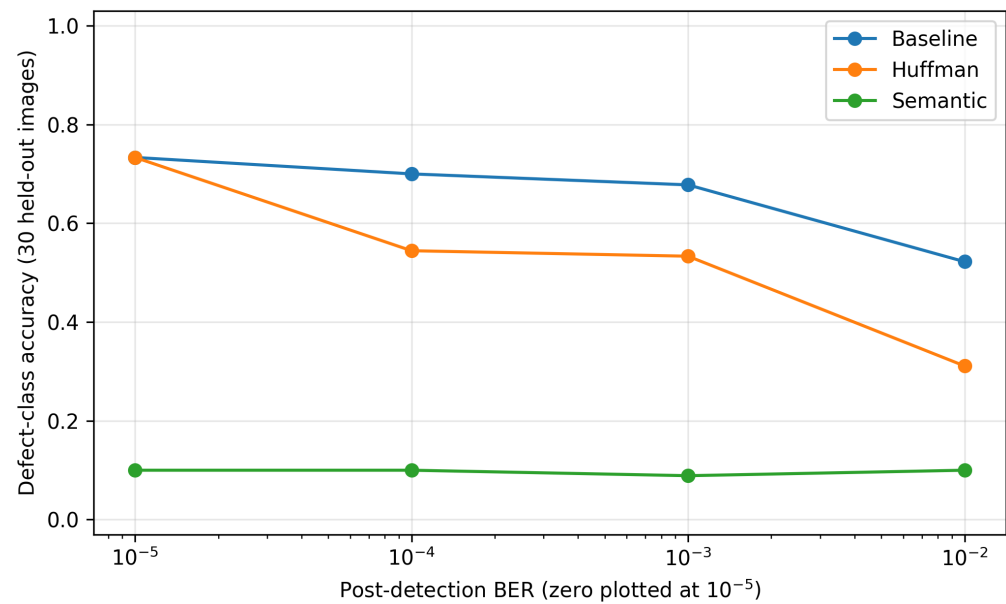


Figure 7. Defect-class proxy accuracy versus post-detection BER. Baseline retains the highest task accuracy throughout the tested range; the semantic reconstruction indicates a task-utility gap relative to the nominal six-class reference.

The three modes consequently exhibit different error-propagation regimes. Raw RGB has the highest clean fidelity and remains the strongest option in absolute PSNR, SSIM, and task accuracy throughout the tested range, although isolated bit errors become visible as spatially dispersed color perturbations. Huffman is lossless when the payload is error-free; once a codeword boundary is corrupted, a single detected-bit error can affect a long sequence of decoded symbols, producing burst-like spatial artifacts and a rapid SSIM decrease. The semantic path is lossy before transmission because of the 128-component bottleneck, but its reconstructed image changes comparatively little as BER increases. This behavior provides stability of the dominant low-frequency structure, while finer industrial defect information remains harder to preserve with the current model. In particular, at BER 10^{-2} , the semantic mode exceeds Huffman in SSIM because Huffman has become strongly desynchronized, yet the baseline still retains higher SSIM and substantially higher task accuracy. The result therefore identifies a useful distinction between channel-error tolerance and semantic utility: a representation can be less sensitive to additional bit errors while still benefiting from richer task-relevant detail for the intended application.

4.6. Robustness Slopes and Correct Interpretation

Table 9 reports the fitted quality-degradation slopes and the corresponding nonzero-BER mean SSIM values.

Table 9. Quality-degradation slopes per decade of BER, fitted over 10^{-4} – 10^{-2} . Absolute mean SSIM over the same nonzero-BER points is included to prevent slope-only interpretation.

Mode	β_{PSNR}	β_{SSIM}	β_{task}	Mean SSIM
Baseline	−10.162	−0.1887	−0.0889	0.853
Huffman	−5.448	−0.2753	−0.1167	0.480
Semantic	−0.0069	−0.00008	0.0000	0.452

The near-zero semantic slopes are informative but should be interpreted together with the absolute metric levels. They indicate that the semantic decoder changes comparatively little as BER increases, but the slope must be interpreted together with the quality intercept and the downstream task score. Baseline has the steepest PSNR decline because it starts

from exact reconstruction and therefore has the largest available dynamic range; its negative slope is therefore best understood together with its high absolute quality. Huffman has the steepest SSIM and task-accuracy deterioration because variable-length desynchronization amplifies a detected-bit error into a longer sequence of source-decoding errors. The semantic path has a near-zero slope because the decoder preserves a similar coarse, low-frequency representation across the tested BER values. At the same time, its mean SSIM of 0.452 and task accuracy of 0.10 indicate that fine texture and defect-relevant detail would benefit from task-aware model refinement. Robustness is therefore characterized here by both the degradation slope and the absolute quality level, rather than by the slope alone.

4.7. Qualitative Error-Propagation Comparison

Figure 8 compares qualitative error propagation for the same representative held-out image. Raw RGB and Huffman panels show visible BER-dependent artifacts. In the semantic row, the decoded image remains visually stable across the tested BER values, consistent with the near-zero semantic degradation slopes. All quantitative metrics are computed from the original per-BER RGB reconstructions.

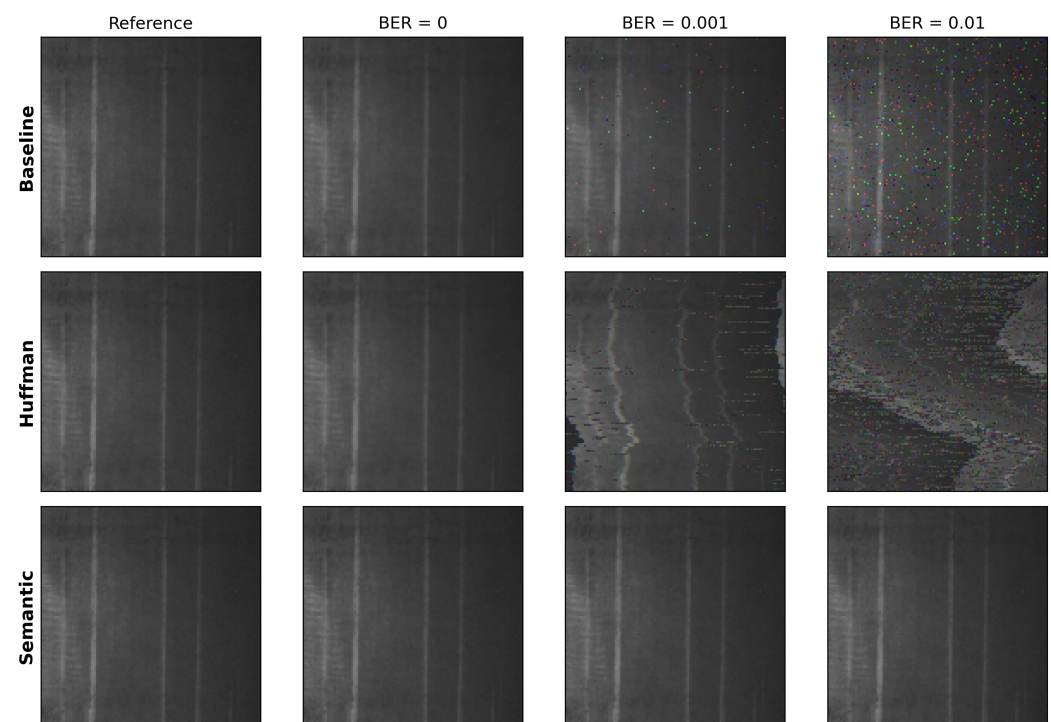


Figure 8. Qualitative error-propagation example. Raw RGB and Huffman show BER-dependent artifacts, whereas the semantic row remains visually stable across BER. Metrics use the original per-BER RGB reconstructions.

5. Discussion

The results reveal a three-way trade-off among payload size, channel-error propagation, and retained visual utility. From the communication perspective, the 4096-bit latent payload is approximately 294 times smaller than an uncompressed 224×224 , 24-bit RGB payload. The corresponding model-derived optical transmission energy decreases from 2.81 J to 0.01 J, while the 0.85 J processing term is measured from the workstation prototype. Combining these two quantities gives a prototype-level energy estimate of 0.86 J per image. This represents an approximate 69% reduction relative to raw RGB and a 72% reduction relative to Huffman in the reported prototype. These gains are meaningful for VLC links with constrained optical airtime, but they quantify resource consumption rather than the equivalence of the conveyed information.

The rate-matched comparison shows why payload reduction must be interpreted together with fidelity and why the 294-fold raw-payload ratio cannot by itself establish a transmission advantage over practical codecs. JPEG and the separated latent representation both use approximately 4 kbit, but JPEG achieves 30.77 dB PSNR and 0.699 SSIM, compared with 14.12 dB and 0.464 for Compact-NEU under the architecture-and-resolution comparison. Moving to the full 224×224 architecture improves the separated latent result to 15.61 dB, and the regularized full model gives the highest SSIM of 0.471, but neither closes the gap to JPEG. Since the bottleneck size and reconstruction objective were not independently ablated, this remaining gap is interpreted cautiously: it is consistent with bottleneck pressure, reconstruction-only training, and limited industrial-domain data being important contributing limitations, but not as definitive proof that any one of them is dominant. The JSCC-VLC baseline further clarifies the issue: when the representation is trained through a differentiable optical channel, the clean result improves to 23.82 dB and 0.568 SSIM and remains at 22.18 dB and 0.519 SSIM at 5 dB SNR. Thus, the observed quality gap is not explained by bit budget alone. It also depends on whether the representation is trained as a separated source code or as a channel-aware JSCC mapping. The downstream proxy reinforces the need for task-aware objectives: clean lossless images obtain 0.733 accuracy, whereas the separated latent reconstruction reaches 0.10 and JSCC-VLC reaches 0.167 in this small audit. Therefore, compact latent interfaces are promising for reducing transmitted payload, but future industrial systems should combine domain-aware representation learning, explicit channel-aware training, and task-relevant loss terms.

The BER sweep further distinguishes the modes. Baseline transmission retains the best absolute image and task quality over the complete tested interval; its gradual deterioration is dominated by spatially isolated pixel errors. Huffman is highly efficient when error-free, while bit errors can strongly affect reconstruction quality because corrupted variable-length boundaries propagate through many decoded symbols. At BER 10^{-2} , its mean SSIM falls to 0.218, below the semantic value of 0.452. The semantic representation, in contrast, remains visually similar across BER values. Figure 8 illustrates this visually stable semantic reconstruction, and the numerical comparisons remain those reported in Tables 8 and 9. This stability is consistent with the near-zero quality slopes and indicates that the compact latent reconstruction is comparatively insensitive to the tested random bit flips. This result is best interpreted as coarse-output stability rather than full task-level robustness, because the clean reconstruction emphasizes low-frequency structure and the task score indicates room for improvement. Thus, the current semantic path favors a stable coarse reconstruction over high-frequency fidelity.

These observations also clarify two physical-layer issues. First, Huffman desynchronization changes the impact of a given detected-bit pattern on the reconstructed image; it leaves the physical BER definition unchanged. Second, reducing payload duration decreases frame energy and optical channel occupancy, while instantaneous received SNR remains governed by transmit power and channel gain. Energy, BER, reconstruction quality, and task utility are consequently reported as separate quantities rather than combined into a single equivalent-efficiency measure. The embedded-energy estimate further shows that the semantic mode is not automatically energy-competitive on every edge node. Relative to the 2.81 J raw-RGB reference, and using the same 0.01 J semantic transmission energy, the break-even inference time satisfies $t_{\text{inf}} < (2.81 - 0.01) / P_H$. This corresponds to approximately 400 ms at 7 W, 187 ms at 15 W, and 112 ms at 25 W. Thus, semantic transmission remains energy-competitive only if the deployed encoder–decoder runs below the corresponding latency threshold; otherwise, processing can dominate the communication saving.

For an industrial semantic VLC system, the results point to a specific redesign path. A natural development path is to use a substantially larger industrial corpus and combine

pixel reconstruction with edge-, texture-, anomaly-, and task-aware losses. The latent representation should be serialized using importance-aware quantization, and high-significance components should receive stronger forward-error correction or unequal error protection. Joint source-channel training could further expose the encoder and decoder to the error patterns expected from the VLC link. The encoder–decoder should also be exported to an embedded inference runtime and measured directly on the intended edge processor, because the workstation MATLAB–Python timing is only a prototype value. Evaluation should then include full industrial benchmarks, detector-level precision, recall, F1 score, mean average precision, and measurements on a physical LED/photodiode platform.

The industrial audit contains 30 out-of-fold images and a compact cross-validated model. It is useful for identifying the principal trade-offs of the evaluated compact model and can be viewed as a focused diagnostic evaluation rather than a deployment benchmark. The natural-image physical-layer sweep and the industrial audit therefore answer complementary questions: the former characterizes the link and energy behavior and is not treated as a leakage-free held-out reconstruction benchmark, whereas the latter evaluates whether the reconstructed information remains visually and operationally useful under an explicit out-of-fold protocol. Combining both levels provides a more defensible assessment than using BER or payload size alone.

6. Conclusions

This work evaluates raw RGB, Huffman, autoencoder-based compact latent payloads, and an end-to-end JSCC-VLC baseline over indoor industrial VLC-oriented image transmission scenarios while explicitly separating communication efficiency from visual and task utility. The 4096-bit latent vector provides a 294-fold payload-size reduction relative to uncompressed 224×224 RGB and is used here as a payload-size reference rather than as a general compression-ratio claim or as evidence of superiority over practical image codecs. It reduces the model-derived VLC transmission-energy component to 0.01 J. When combined with the measured workstation prototype processing term, the corresponding prototype-level energy estimate is 0.86 J per image, compared with 2.81 J for the baseline and 3.11 J for Huffman.

The independent industrial-domain audit further clarifies the current operating point of the reconstruction-only model. Under the architecture-and-resolution comparison, Compact-NEU obtains 14.12 dB PSNR and 0.464 SSIM on 30 out-of-fold steel-surface images, while Full-NEU improves PSNR to 15.61 dB and Full-NEU-Reg gives the highest SSIM at 0.471. A JPEG file with approximately the same 4-kbit payload obtains 30.77 dB and 0.699 SSIM. Because Compact-NEU and Full-NEU differ in both input resolution and network architecture/capacity, this comparison indicates a combined architecture/resolution effect rather than a capacity-only effect. The separated latent quality slope is nearly zero as BER increases, and the qualitative comparison shows a visually stable reconstruction across BER columns, while fine texture and local defect detail would benefit from further task-aware refinement. The remaining gap to JPEG is therefore interpreted cautiously as being consistent with bottleneck pressure and reconstruction-only training as important and plausible contributing limitations, not as an independent proof that these factors are dominant.

Accordingly, the present evidence supports a potentially energy-efficient and comparatively error-stable latent payload interface under the measured prototype and estimated embedded-latency conditions, while also identifying the main ingredients needed for improved defect-aware industrial semantic transmission. Future work should employ full industrial datasets, task-aware and channel-aware training, importance-based latent protection, embedded edge measurements, and physical VLC experiments to progress toward robust industrial visual-monitoring systems.

Supplementary Materials: The industrial audit uses publicly available NEU steel-surface defect images cited in the manuscript. The exact image manifest, out-of-fold split metadata, scripts, and generated metric tables are included in the supplementary package. The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/photonics13070660/s1>.

Author Contributions: Conceptualization, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; methodology, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; software, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; validation, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; formal analysis, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; investigation, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; data curation, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; writing—original draft preparation, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S.; writing—review and editing, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B., M.C.R. and I.S.; visualization, A.A.P., P.P.J., M.I., C.A.M., D.Z.-B. and I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ANID/FONDECYT Iniciación No. 11240799, Universidad de Las Américas under Project 563.B.XVI.25, FONDECYT Regular (Grant No. 1261732/2026), Pontificia Universidad Católica del Ecuador under Project PEP QINV0485-IINV528020300, ANID AC3E CIA250006, Dicyt/USACH/code 062413SG, ANID/FIU 137139, and the Department of Networking and Telecommunication Engineering, Universidad de las Américas.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: These metadata identify which industrial images are used for training and testing in each fold. The original ILSVRC2012 data remain subject to their source access conditions.

Acknowledgments: This research was supported by Universidad Diego Portales, Manchester Metropolitan University and the Department of Networking and Telecommunication Engineering at Universidad de las Américas.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BER	Bit error rate
CAE	Convolutional autoencoder
CNN	Convolutional neural network
CPU	Central processing unit
CRC	Cyclic redundancy check
CSV	Comma-separated values
FOV	Field of view
GPU	Graphics processing unit
HOG	Histogram of oriented gradients
HOG-SVM	Histogram-of-oriented-gradients support-vector machine
IEEE	Institute of Electrical and Electronics Engineers
IEEE-754	IEEE Standard for Floating-Point Arithmetic
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IM/DD	Intensity modulation/direct detection
ISI	Intersymbol interference
JPEG	Joint Photographic Experts Group
JPEG2000	JPEG 2000 image-coding standard
JSCC	Joint source-channel coding
LED	Light-emitting diode
LOS	Line of sight
MSE	Mean squared error

NEU	Northeastern University
OOK–NRZ	On–off keying non–return–to–zero
PIN	Positive–intrinsic–negative
PSNR	Peak signal–to–noise ratio
RAM	Random–access memory
RC	Resistance–capacitance
RF	Radio frequency
RGB	Red–green–blue
RX	Receiver
SE	Squeeze–and–Excitation
SNR	Signal–to–noise ratio
SSIM	Structural similarity index measure
SVM	Support vector machine
TX	Transmitter
VLC	Visible light communication

References

- Almadani, Y.; Plets, D.; Bastiaens, S.; Joseph, W.; Ijaz, M.; Ghassemlooy, Z.; Rajbhandari, S. Visible Light Communications for Industrial Applications—Challenges and Potentials. *Electronics* **2020**, *9*, 2157. [\[CrossRef\]](#)
- Játiva, P.P.; Azurdia-Meza, C.A.; Cañizares, M.R.; Sánchez, I.; Iturralde, D. On the performance of visible light communications in underground mines. In *Proceedings of the 2020 IEEE Latin-American Conference on Communications (LATINCOM)*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
- Játiva, P.P.; Azurdia-Meza, C.A.; Cañizares, M.R.; Zabala-Blanco, D.; Soto, I. BER performance of OFDM-based visible light communication systems. In *Proceedings of the 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
- Palacios Játiva, P.; Sánchez, I.; Soto, I.; Azurdia-Meza, C.A.; Zabala-Blanco, D.; Ijaz, M.; Dehghan Firoozabadi, A.; Plets, D. A novel and adaptive angle diversity-based receiver for 6g underground mining vlc systems. *Entropy* **2022**, *24*, 1507. [\[CrossRef\]](#)
- Komine, T.; Nakagawa, M. Fundamental Analysis for Visible-Light Communication System Using LED Lights. *IEEE Trans. Consum. Electron.* **2004**, *50*, 100–107. [\[CrossRef\]](#)
- Ghassemlooy, Z.; Popoola, W.O.; Rajbhandari, S. *Optical Wireless Communications: System and Channel Modelling with MATLAB*; CRC Press: Boca Raton, FL, USA, 2012.
- Muthukumar, M.; Sathiamoorthy, M.; Kannan, R. An Efficient Data Compression Technique Using Huffman Coding. *Int. J. Comput. Appl.* **2015**, *975*, 1–6. [\[CrossRef\]](#)
- Ma, K.; Abumarshoud, H.; Hua, S.; Imran, M.A.; Sun, Y. Design and Demonstration of a Visible Light Based Semantic Communication System. *IEEE Trans. Veh. Technol.* **2026**, 1–12. [\[CrossRef\]](#)
- Xin, G.; Fan, P.; Letaief, K.B. Semantic Communication: A Survey of Its Theoretical Development. *Entropy* **2024**, *26*, 102. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wheeler, D.; Natarajan, B. Engineering Semantic Communication: A Survey. *IEEE Access* **2023**, *11*, 13965–13995. [\[CrossRef\]](#)
- Hu, L.; Yu, L.; Qin, Z. Deep Learning-Based Semantic Communication System for Wireless Image Transmission. *IEEE Wirel. Commun. Lett.* **2025**, *14*, 2391–2395. [\[CrossRef\]](#)
- Wang, H.; Cao, Z.; Zhang, H. Semantic-Aware Image Compression Architecture for Semantic Communication. In *Proceedings of the 2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
- Zeng, C.; Zhang, L. An Improved Low-Bit-Rate Image Compression Framework Based on Semantic-Aware Model and Neighborhood Attention. *IEEE Access* **2025**, *13*, 139490–139505. [\[CrossRef\]](#)
- Islam, N.; Shin, S. Deep Learning in Physical Layer: Review on Data-Driven End-to-End Communication Systems and Their Enabling Semantic Applications. *IEEE Open J. Commun. Soc.* **2024**, *5*, 4207–4240. [\[CrossRef\]](#)
- Chen, W.; Ren, T.; Yuan, T.; Han, D.; Yang, C.; Luo, M.; Ju, C.; Zhang, M.; Wang, D. Semantic-Empowered Visible Light Communication for Image Transmission Based on Deep Convolutional Generative Adversarial Network. *Opt. Express* **2024**, *32*, 32564–32584. [\[CrossRef\]](#) [\[PubMed\]](#)
- Elgala, H.; Mesleh, R.; Haas, H. Indoor Optical Wireless Communication: Potential and State-of-the-Art. *IEEE Commun. Mag.* **2011**, *49*, 56–62. [\[CrossRef\]](#)
- Rajagopal, S.; Roberts, R.D.; Lim, S.K. IEEE 802.15.7 Visible Light Communication: Modulation Schemes and Dimming Support. *IEEE Commun. Mag.* **2012**, *50*, 72–82. [\[CrossRef\]](#)

18. Bourtsoulatzé, E.; Kurka, D.B.; Gündüz, D. Deep Joint Source-Channel Coding for Wireless Image Transmission. *IEEE Trans. Cogn. Commun. Netw.* **2019**, *5*, 567–579. [[CrossRef](#)]
19. Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.H. Deep Learning Enabled Semantic Communication Systems. *IEEE Trans. Signal Process.* **2021**, *69*, 2663–2675. [[CrossRef](#)]
20. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
21. Song, K.; Yan, Y. A Noise Robust Method Based on Completed Local Binary Patterns for Hot-Rolled Steel Strip Surface Defects. *Appl. Surf. Sci.* **2013**, *285*, 858–864. [[CrossRef](#)]
22. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893. [[CrossRef](#)]
23. Wallace, G.K. The JPEG Still Picture Compression Standard. *IEEE Trans. Consum. Electron.* **1992**, *38*, xviii–xxxiv. [[CrossRef](#)]
24. Candia, D.A.; Játiva, P.P.; Azurdia Meza, C.; Sánchez, I.; Ijaz, M. Performance analysis of the particle swarm optimization algorithm in a VLC system for localization in hospital environments. *Appl. Sci.* **2024**, *14*, 2514. [[CrossRef](#)]
25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141. [[CrossRef](#)]
26. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 7–9 May 2015.
27. NVIDIA Corporation. Jetson Orin Nano Series, Jetson Orin NX Series, and Jetson AGX Orin Series: Platform Power and Performance. NVIDIA Jetson Linux Developer Guide. 2025. Available online: <https://docs.nvidia.com/jetson/archives/r36.5/DeveloperGuide/SD/PlatformPowerAndPerformance/JetsonOrinNanoSeriesJetsonOrinNxSeriesAndJetsonAgxOrinSeries.html> (accessed on 28 June 2026).
28. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.