RECOVER SEMANTICS, NOT CODEWORDS: AN ASYMMETRIC PARADIGM FOR VECTOR-QUANTIZED JSCC

Anonymous authors
Paper under double-blind review

ABSTRACT

Vector-quantized semantic communication commonly follows a recover-then-reconstruct paradigm: the receiver detects a transmitted symbol, retrieves its source codeword, and then reconstructs the image. This design resolves channel ambiguity through a discrete decision, although a codeword error need not imply poor reconstruction. We argue that faithful source reconstruction does not require exact codeword recovery. Our insight is to jointly learn which source representations become neighbors in the channel and how observations within those neighborhoods are reconstructed. We introduce **AsymVQ-JSCC**, which retains a discrete source vocabulary while coupling source-aware irregular QAM with continuous semantic recovery. The transmitter learns constellation coordinates from source relationships and refines their geometry through the image reconstruction objective. Building on this geometry, a continuous receiver maps raw observations into semantic latents, learning denoising jointly with the decoder. Bypassing hard detection and codebook lookup allows observations between constellation points to support reconstruction through latents beyond the source vocabulary. Each spatial token occupies one channel use, independently of vocabulary size. Across four image datasets, bandwidth ratios, and SNRs, AsymVQ-JSCC achieves competitive PSNR and perceptual quality with graceful degradation under both AWGN and Rayleigh fading. In a fixed-checkpoint intervention on Kodak, projecting receiver latents to the nearest source codeword, while leaving the transmitted symbols, channel observations, and model weights unchanged, reduces PSNR by 1.46–2.57 dB and increases LPIPS by 0.063–0.077 across the tested SNRs. These results provide direct evidence that reconstruction benefits from continuous receiver representations beyond the source vocabulary.

1 INTRODUCTION

Deep joint source–channel coding (JSCC) offers an alternative to separate source and channel coding (Shannon, 1948) for wireless image transmission. Directly optimizing reconstruction over a differentiable channel (Choi et al., 2019) exploits source statistics at finite blocklength (Fresia et al., 2010; Dai et al., 2022), avoids manually designed coding interfaces, and enables graceful degradation as channel quality changes (Bourtsoulatze et al., 2019). To improve reconstruction under limited resources (Kurka & Gündüz, 2020; 2021), the community has advanced source representations (Zhang et al., 2023b; 2025a), encoder–decoder capacity (Yang et al., 2023; 2024) and adaptation (Zhang et al., 2024), and perceptual or generative modeling (Yang et al., 2025a), strengthening both the transmitted representations and the models that reconstruct them.

Vector quantization (VQ) offers two advantages for visual communication. First, its learned codebook provides a compact vocabulary of visual patterns (Van Den Oord et al., 2017) that, with a capable decoder and perceptual training, supports visually faithful reconstruction under strong compression. Second, the vocabulary is finite, discrete, and enumerable (Nemati et al., 2023): indexing its entries and associating them with channel symbols (Fu et al., 2023; Huh et al., 2025) connects learned source representations to digital transmission (Huang et al., 2025). VQ thus combines source representation capacity with a structured communication alphabet.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A common use of this interface is symmetric: the transmitter maps a source codeword to a channel symbol; the receiver detects a symbol, recovers its index, and retrieves a codeword from the same vocabulary before decoding (Hu et al., 2023; Meng et al., 2025). The transmission vocabulary thereby also restricts the decoder input, although reconstruction does not require this coupling. The resulting recover-then-reconstruct (RTR) interface, together with source-independent constellation design, introduces three limitations.

1. Codeword-constrained reconstruction. Receiver-side codebook lookup makes a discrete codeword decision a prerequisite for decoding, even under end-to-end reconstruction training. Yet codeword identity is an intermediate representation: token errors can have different reconstruction costs, and a continuous latent estimate may reconstruct better than either candidate codeword. Restricting each received latent to the transmission vocabulary therefore excludes potentially useful decoder inputs.

2. Source-agnostic physical geometry. When constellation geometry is chosen independently of source representations, physically neighboring symbols need not correspond to similar source content. Channel noise can then confuse codewords with very different reconstruction consequences. Prior work improves robustness through coding and modulation (Tung et al., 2022; Bo et al., 2024; Xie et al., 2023), channel protection (Guo et al., 2024; Liu et al., 2024), and channel-aware codebooks (Meng et al., 2025). Additional protection consumes channel uses at fixed payload and modulation order; adapting the source codebook imposes channel-related demands on source quantization. These tradeoffs motivate learning the physical coordinates associated with source codewords, so the neighborhoods exposed to channel noise can adapt to source relationships and the reconstruction objective.

3. Observation collapse before reconstruction. Hard detection replaces a continuous observation with a discrete index. Observations with different confidence and positions within a decision region become the same codeword, whereas a small boundary-crossing perturbation abruptly changes the decoder input. Neither a stronger decoder nor a better constellation can recover the discarded observation information while the detect-and-lookup interface remains.

We argue that the source codebook should define the transmission vocabulary without restricting the receiver to that set. An observation assigned to an incorrect token can still support faithful reconstruction. Our central insight is to learn which source representations become neighbors in channel space, accounting for the reconstruction costs of symbol confusion, together with how their continuous noisy neighborhoods are decoded. Figure 1 contrasts this joint design with RTR in geometry, observation processing, and decoder constraints.

We introduce **AsymVQ-JSCC**, with discrete transmission and continuous reception. Its transmitter retains the source-VQ codebook, mapping codewords to learned, power-constrained two-dimensional coordinates. This irregular QAM replaces a prescribed grid with source-aware neighborhoods. Joint reconstruction training of the mapping, source codec, and receiver links geometry to the reconstruction consequences of channel ambiguity.

The continuous receiver maps raw observations directly into decoder latents, bypassing hard detection and codebook lookup. Jointly with the decoder, it learns semantic conversion, denoising, and reconstruction, preserving observation information and recovering useful off-codebook latents from observations between constellation points. The irregular mapping organizes the physical representation; the receiver learns to reconstruct from its noisy neighborhoods.

Each spatial token uses one channel symbol. At a fixed encoder downsampling factor, spatial token density determines the channel bandwidth ratio: enlarging the source codebook increases constellation cardinality but leaves the symbol count unchanged.

Our contributions are summarized as follows:

- **Reconstruction beyond codeword identity.** We separate the vocabulary needed for discrete transmission from the latent space used for reconstruction. An affine reconstruction analysis characterizes the excess conditional reconstruction error caused by restricting receiver estimates to source codewords, providing a rationale for continuous recovery.
- **Jointly learned geometry and continuous semantic recovery.** AsymVQ-JSCC jointly optimizes source-aware irregular QAM, a continuous receiver, and the source codec while

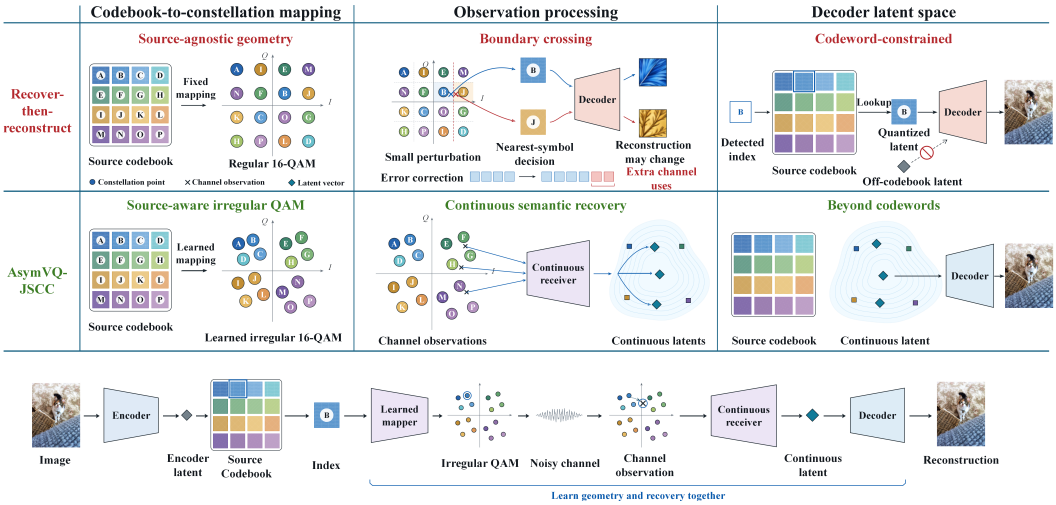

Figure 1: **Recover semantics, not codewords.** RTR (top) versus AsymVQ-JSCC (middle). **Left:** Illustrated RTR uses fixed regular QAM; learned irregular QAM adapts neighborhoods to source relationships and reconstruction costs. **Middle:** Boundary crossings change hard decisions and potentially reconstruction; error correction adds overhead. Our receiver learns semantic recovery and denoising from noisy observations. **Right:** Codeword lookup versus continuous, off-codebook latents. **Bottom:** Discrete source quantization with jointly learned geometry and continuous recovery for reconstruction, bypassing receiver-side hard detection and lookup. Matching colors and letters identify codeword–symbol pairs; crosses denote observations and diamonds latent vectors. The 16-point constellations are schematic.

retaining the source-VQ vocabulary. The learned geometry organizes physically confusable source representations, and the receiver and decoder learn semantic conversion, denoising, and reconstruction from raw observations through continuous latents beyond the source vocabulary, without hard detection or codebook lookup.

- **Empirical evaluation of the proposed design.** Across four image datasets, bandwidth ratios, and SNRs, AsymVQ-JSCC achieves competitive PSNR and perceptual quality and degrades gracefully under both AWGN and Rayleigh fading. In a fixed-checkpoint intervention on Kodak, enforcing source-codebook membership at the receiver reduces PSNR by 1.46–2.57 dB and increases LPIPS by 0.063–0.077 across the tested SNRs, while keeping the transmitter, channel observations, and model weights unchanged. Sensitivity studies further examine constellation cardinality, mapper parameterization, and PAPR.

2 Related Work

2.1 Learned Joint Source–Channel Coding

Classical JSCC (Shannon, 1948; Fresia et al., 2010) jointly optimizes source representations and channel protection (Hamzaoui et al., 2005) under finite budgets (Nosratinia et al., 2003), including noisy vector quantization (Farvardin, 1990). Neural methods learn this mapping end to end: NECST jointly learns compression and error correction over binary channels (Choi et al., 2019); DeepJSCC maps images to continuous symbols with graceful degradation under channel variation (Bourtsoulatze et al., 2019). Extensions address feedback (Kurka & Gündüz, 2020), bandwidth adaptation (Kurka & Gündüz, 2021; Zhang et al., 2023b), nonlinear transforms (Dai et al., 2022), probabilistic latents (Saidutta et al., 2021), and other modalities (Xie et al., 2021; Weng & Qin, 2021).

Transformer architectures (Yang et al., 2023; 2024), hierarchical representations (Zhang et al., 2025a), channel conditioning (Zhang et al., 2024), and generative priors (Yang et al., 2025a; Zhang et al., 2025b) improve rate adaptation (Huang et al., 2025), generalization, and perceptual

quality. Beyond these advances in models, architectures, and priors, we examine whether reconstruction requires recovering each transmitted source token as a discrete codeword.

2.2 Digital and Vector-Quantized Semantic Communication

Finite-alphabet JSCC (Liu et al., 2024) employs scalar quantization (Guo et al., 2024), learned constellations (Tung et al., 2022; Wang et al., 2022), probabilistic coding and modulation (Bo et al., 2022; 2024), or hybrid semantic–digital transmission (Evgenidis et al., 2024). DeepJSCC-Q (Tung et al., 2022) already combines predefined or learned finite constellations with neural reconstruction from noisy observations. JCM (Bo et al., 2024) uses probabilistic coding and modulation; Digital-SC (Guo et al., 2024) and D^2-JSCC (Huang et al., 2025) introduce explicit digital quantization, coding, or adaptive rate allocation.

VQ-based systems connect semantic features to symbols or channel indices (Xie et al., 2023; Huh et al., 2025) through discrete codebooks (Fu et al., 2023; Nemati et al., 2023). Robustness improves through masked-codebook learning (Hu et al., 2023), modulation-specific codebooks, channel-aware geometry (Meng et al., 2025), or multiple codebooks (Shin et al., 2025). These approaches design or recover robust discrete representations. AsymVQ-JSCC retains a separate high-dimensional source-VQ vocabulary, learns its physical geometry, and directly estimates decoder-compatible latents from observations without requiring codebook membership or correct token recovery.

2.3 Discrete Representation Learning and Generative Reconstruction

VQ-VAE (Van Den Oord et al., 2017) inspired hierarchical codebooks (Razavi et al., 2019), adversarial reconstruction (Esser et al., 2021), residual quantization (Lee et al., 2022), improved codebook utilization (Zheng & Vedaldi, 2023), and simplified finite quantizers (Mentzer et al., 2024). Learned compression combines entropy models (Yi et al., 2025), hierarchical priors (Ballé et al., 2018), autoregressive context (Minnen et al., 2018), and perceptual objectives (Mentzer et al., 2020) for compact, high-quality representations (Cheng et al., 2020). Variational (Vahdat & Kautz, 2020) and diffusion decoders (Ho et al., 2020; Rombach et al., 2022; Zhang et al., 2023a) also support realistic reconstruction from continuous latent estimates when observations are incomplete or corrupted.

These methods typically study compression or generation without a noisy physical channel, whereas codebook-based communication commonly preserves discreteness at both endpoints. AsymVQ-JSCC connects discrete transmission to continuous, reconstruction-trained reception by removing receiver-side codebook membership as a constraint, beyond increasing decoder capacity alone.

3 Method

AsymVQ-JSCC first initializes source-aware irregular QAM from a pretrained VQ codec, then learns continuous semantic recovery while jointly refining the codec and constellation (Figure 2).

3.1 Overview and Problem Formulation

An RGB image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ is encoded as $\mathbf{Z} = E(\mathbf{x}) \in \mathbb{R}^{N\times d}$, where $N = HW/f^2$ for downsampling factor f. The source codebook $\mathcal{C} = \{\mathbf{e}_i\}_{i=1}^K \subset \mathbb{R}^d$ and physical constellation $\mathcal{S} = \{\mathbf{s}_i\}_{i=1}^K \subset \mathbb{R}^2$ share indices; each point gives one symbol's in-phase and quadrature components. Spatial token t is quantized as

$$i_t = \arg\min_i \|\mathbf{z}_t - \mathbf{e}_i\|_2^2, \qquad \mathbf{z}_t^q = \mathbf{e}_{i_t}, \qquad \mathbf{s}_{i_t} = M(\mathbf{e}_{i_t}), \tag{1}$$

where M maps codewords to physical symbols. The transmitted sequence $\mathbf{S} = [\mathbf{s}_{i_t}]_{t=1}^N$ produces observations $\mathbf{Y} = \mathcal{H}(\mathbf{S})$ under AWGN or Rayleigh fading with additive noise. For AWGN at SNR γ in dB,

$$\mathbf{y}_t = \mathbf{s}_{i_t} + \mathbf{n}_t, \qquad \mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_2), \qquad \sigma^2 = \frac{P}{2\,10^{\gamma/10}}, \tag{2}$$

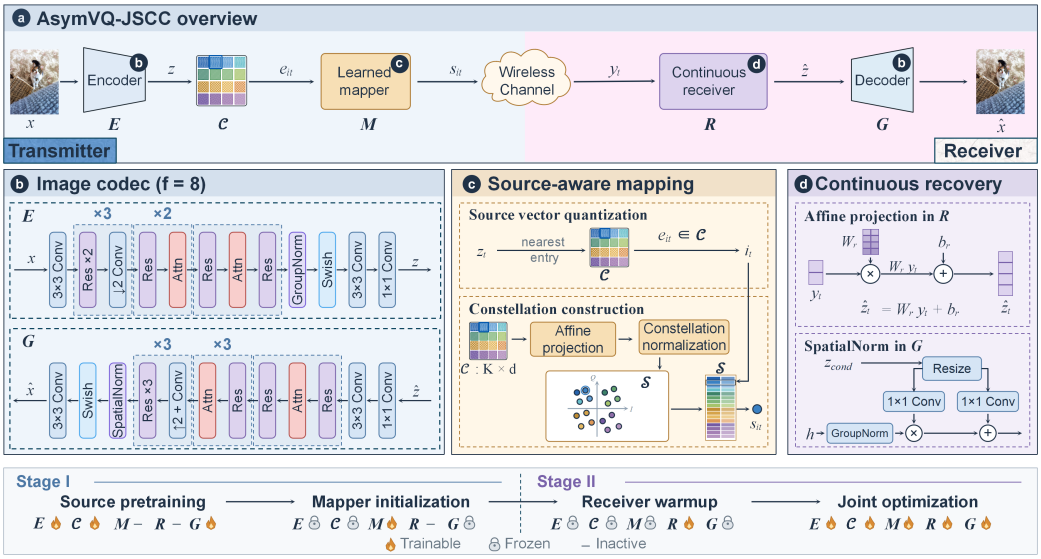

Figure 2: **Architecture of AsymVQ-JSCC.** (a) Communication pipeline; (b) image codec; (c) source quantization and irregular constellation construction; (d) continuous receiver and decoder conditioning. Res and Attn denote residual blocks (ResBlock) and attention blocks (AttentionBlock), respectively. $\mathbf{z}_{\text{cond}}$ and $\mathbf{h}$ denote the conditioning latent and intermediate decoder features. Constellations are schematic.

where $P = \sum_i \pi_i \|\mathbf{s}_i\|_2^2$ and π_i is the source-codeword usage probability. The receiver R transforms observations directly into decoder latents:

$$\hat{\mathbf{z}}_t = R(\mathbf{y}_t), \qquad \hat{\mathbf{x}} = G(\hat{\mathbf{Z}}). \tag{3}$$

Token-grid reshaping is implicit; Figure 2(a) denotes encoder and receiver grids by $\mathbf{z}$ and $\hat{\mathbf{z}}$. E and G include latent projections. One complex use per token gives $\rho = N/(3HW) = 1/(3f^2)$ per input scalar, or $\rho_{\text{pixel}} = N/(HW) = 1/f^2$ per pixel. At fixed f, K changes constellation cardinality but neither budget.

3.2 Learning Geometry for Continuous Recovery

Geometry determines confusable source representations; recovery determines their reconstruction cost. Fix $\mathcal{C}$ and $\mathcal{S}$, draw $I \sim \pi$, and observe $\mathbf{y} = \mathbf{s}_I + \mathbf{n}$ under AWGN. For affine reconstruction $g(\mathbf{v}) = \mathbf{B}\mathbf{v} + \mathbf{b}$ targeting $g(\mathbf{e}_I)$, the conditional probabilities are

$$p_i(\mathbf{y}) = \frac{\pi_i \exp\!\big(-\|\mathbf{y} - \mathbf{s}_i\|_2^2/(2\sigma^2)\big)}{\sum_j \pi_j \exp(-\|\mathbf{y} - \mathbf{s}_j\|_2^2/(2\sigma^2))}. \tag{4}$$

Write $p_i = p_i(\mathbf{y})$, $\boldsymbol{\mu} = \sum_i p_i \mathbf{e}_i$, and $\Delta_{ij} = \|\mathbf{B}(\mathbf{e}_i - \mathbf{e}_j)\|_2^2$ for the reconstruction cost of confusing codewords i and j.

Proposition 1 (Reconstruction cost of codeword constraints). For the affine model, the conditional risk decomposes as

$$\begin{aligned} \mathcal{R}(\mathbf{v} \mid \mathbf{y}) &= \mathbb{E}\big[\|g(\mathbf{v}) - g(\mathbf{e}_I)\|_2^2 \mid \mathbf{y}\big] \\ &= \underbrace{\frac{1}{2}\sum_{i,j} p_i p_j \Delta_{ij}}_{\mathcal{V}(\mathbf{y})\text{: residual ambiguity}} + \|\mathbf{B}(\mathbf{v} - \boldsymbol{\mu})\|_2^2. \end{aligned} \tag{5}$$

An unconstrained estimate attains $\mathcal{V}(\mathbf{y})$ at $\mathbf{v} = \boldsymbol{\mu}$; restricting the estimate to the source codebook adds

$$\min_{\mathbf{v} \in \mathcal{C}} \mathcal{R}(\mathbf{v} \mid \mathbf{y}) - \min_{\mathbf{v} \in \mathbb{R}^d} \mathcal{R}(\mathbf{v} \mid \mathbf{y}) = \min_j \|\mathbf{B}(\mathbf{e}_j - \boldsymbol{\mu})\|_2^2 \geq 0. \tag{6}$$

This penalty is positive when the conditional mean reconstruction differs from every decoded codeword. Appendix A gives the proof and a two-codeword example.

The decomposition motivates joint geometry and recovery: continuous estimates can reduce the discrete-choice penalty, while residual ambiguity depends on shared noisy neighborhoods and reconstruction costs. Source-aware irregular QAM organizes these neighborhoods; the receiver learns to decode their observations. The affine surrogate supplies the rationale, with its posterior used only for analysis. In the nonlinear system, source-space confusion costs initialize geometry, which image reconstruction then refines.

3.3 Stage I: Source-Aware Constellation Initialization

Source vocabulary. We first train $E, G, \mathcal{C}$ with the image objective

$$\ell_{\text{img}}(\mathbf{x}, \tilde{\mathbf{x}}) = d_{\text{MSE}}(\mathbf{x}, \tilde{\mathbf{x}}) + \lambda_p d_{\text{LPIPS}}(\mathbf{x}, \tilde{\mathbf{x}}) + \lambda_g \ell_{\text{adv}}(\tilde{\mathbf{x}}), \tag{7}$$

and source-VQ penalty

$$\mathcal{L}_{\text{VQ}} = \frac{\|\operatorname{sg}(\mathbf{Z}^q) - \mathbf{Z}\|_F^2 + \beta\|\mathbf{Z}^q - \operatorname{sg}(\mathbf{Z})\|_F^2}{Nd}, \tag{8}$$

Here sg is stop-gradient and ℓ_{adv} the generator's PatchGAN loss. Minimizing $\mathbb{E}_{\mathbf{x}}[\ell_{\text{img}}(\mathbf{x}, G(\mathbf{Z}^q)) + \mathcal{L}_{\text{VQ}}]$ establishes the vocabulary before physical mapping.

Source-aware physical mapping. The mapper centers and normalizes projections $\mathbf{u}_i = \mathbf{W}_m \mathbf{e}_i + \mathbf{b}_m$ using codeword usage probabilities:

$$\mathbf{s}_i = \alpha(\mathbf{u}_i - \bar{\mathbf{u}}), \qquad \bar{\mathbf{u}} = \sum_i \pi_i \mathbf{u}_i, \qquad \alpha = \left(\frac{P}{\sum_j \pi_j \|\mathbf{u}_j - \bar{\mathbf{u}}\|_2^2}\right)^{1/2}. \tag{9}$$

Thus $\sum_i \pi_i \mathbf{s}_i = \mathbf{0}$ and $\sum_i \pi_i \|\mathbf{s}_i\|_2^2 = P$. Projecting the full codebook constructs $\mathcal{S}$; source index i_t selects $\mathbf{s}_{i_t}$. M learns the geometry; estimated usage probabilities are fixed at inference.

We initialize M with $E, G, \mathcal{C}$ frozen, using the source vocabulary to organize physical neighbors. Normalized weights ω_{tj} decay with squared observation-to-symbol distance (Appendix B.2), giving

$$\begin{aligned} \mathcal{L}_{\text{map}} &= \mathbb{E}\left[\frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{K}\omega_{tj}\|\operatorname{sg}(\mathbf{z}_t) - \operatorname{sg}(\mathbf{e}_j)\|_2^2\right], \\ \mathcal{L}_{\text{init}} &= \mathcal{L}_{\text{map}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}}, \end{aligned} \tag{10}$$

where $\mathcal{L}_{\text{sep}}$ penalizes all distinct point pairs within a separation margin. Expectations average images and channel realizations, including SNR sampling where used. Continuous-token–codeword distances proxy reconstruction-dependent confusion costs: nearby candidates contribute according to their source mismatch, encouraging compatible physical neighborhoods. With source representations fixed, gradients pass through physical coordinates and soft weights. Stage II refines the geometry through image reconstruction; the deployed receiver omits the proxy weights.

3.4 Stage II: Continuous Semantic Recovery and Joint Optimization

Continuous receiver. To decode observations of the initialized constellation, we introduce the shared affine map

$$R(\mathbf{y}_t) = \mathbf{W}_r \mathbf{y}_t + \mathbf{b}_r, \qquad \mathbf{W}_r \in \mathbb{R}^{d\times 2}. \tag{11}$$

Outputs need not be codewords: $G \circ R$ combines affine reception with nonlinear semantic conversion, denoising, and spatial reconstruction, without detection or lookup. We fit R by minimizing $d_{\text{MSE}}(\mathbf{x}, G(R(\mathbf{S})))$ on noiseless symbols with the pretrained codec and constellation frozen, establishing a decodable interface before joint noise adaptation.

Discrete transmission with joint gradients. We unfreeze the codec and mapper for joint training. Differentiable source-space assignments $\mathbf{a}_t$ (Appendix B.2) propagate reconstruction gradients through discrete selection using

$$\mathbf{a}_t^{\text{ST}} = \text{onehot}(i_t) + \mathbf{a}_t - \operatorname{sg}(\mathbf{a}_t), \qquad \mathbf{s}_{i_t} = \sum_i a_{ti}^{\text{ST}} \mathbf{s}_i. \tag{12}$$

Similarly, $\mathbf{z}_t^q = \sum_i a_{ti}^{\mathrm{ST}} \mathbf{e}_i$ in Eq. (8). Forward selection chooses one codeword and symbol; gradients pass through the soft assignment. Initialization uses this discrete selection; inference uses Eq. (1)'s hard rule.

Joint reconstruction objective. The noiseless reconstruction $\hat{\mathbf{x}}^{(0)} = G(R(\mathbf{S}))$ shares symbols and parameters with its noisy counterpart. The generator minimizes

$$\begin{aligned} \mathcal{L}_{\mathrm{joint}} &= \mathbb{E}\big[w_n\{\ell_{\mathrm{img}}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{L}_{\mathrm{VQ}}\} + w_c \ell_{\mathrm{img}}(\mathbf{x}, \hat{\mathbf{x}}^{(0)})\big] + \mathcal{L}_{\mathrm{const}}, \\ \mathcal{L}_{\mathrm{const}} &= \lambda_{\mathrm{nn}} \mathcal{L}_{\mathrm{nn}} + \lambda_{\mathrm{papr}} \mathcal{L}_{\mathrm{papr}}. \end{aligned} \tag{13}$$

The channel term trains reconstruction over noisy neighborhoods; the noiseless auxiliary term preserves constellation-point fidelity. The VQ penalty is counted once. With $[a]_+ = \max(0, a)$, the constellation regularizers are

$$\begin{aligned} \mathcal{L}_{\mathrm{nn}} &= \frac{1}{K} \sum_i \left[m_{\mathrm{nn}} - \min_{j \neq i} \|\mathbf{s}_i - \mathbf{s}_j\|_2 \right]_+^2, \\ \mathcal{L}_{\mathrm{papr}} &= \left[\frac{\max_i \|\mathbf{s}_i\|_2^2}{K^{-1} \sum_i \|\mathbf{s}_i\|_2^2} - \kappa \right]_+. \end{aligned} \tag{14}$$

The first penalizes nearest neighbors below margin m_{nn}, versus all pairs during initialization. The second controls peak-to-average power ratio (PAPR) under uniform constellation weights. We jointly update $E, \mathcal{C}, M, R, G$, recomputing the constellation from the current codebook and mapper. Separate discriminator updates use hinge losses on real images and both reconstructions. Appendix B specifies schedules, weights, and power conventions.

4 Experiments

We evaluate perceptual reconstruction across channel noise and transmission budgets using LPIPS and DISTS, with fidelity and reconstruction statistics in the appendix.

4.1 Experimental Setup

Operating points and evaluation. We evaluate Kodak, Tecnick, DIV2K, and CLIC2020 under AWGN and Rayleigh fading. The channel bandwidth ratio $\rho = N/(3HW)$ counts complex uses per RGB scalar. One symbol per token gives rates $\{1/48, 1/192, 1/768\}$ for downsampling factors $\{4, 8, 16\}$. SNR sweeps target $\rho = 1/192$; bandwidth sweeps use 10 dB. Coded transmission counts parity and padding, using recorded actual rates where available.

Systems and baselines. We compare with continuous DeepJSCC (Bourtsoulatze et al., 2019), a local ProGIC reproduction (Cao et al., 2026) with LDPC and 64-QAM, an adapted released CDM-JSCC model (Yang et al., 2025b), SwinJSCC (Yang et al., 2024), and MambaJSCC (Wu et al., 2025). All methods are evaluated on Kodak, Tecnick, DIV2K, and CLIC2020. For the Rayleigh experiments, all methods are evaluated under the same fading model and receiver-CSI assumption at each operating point. Appendix C specifies the baseline implementations, training and evaluation protocols, and channel conventions.

Metrics. LPIPS (Zhang et al., 2018) measures perceptual feature discrepancy, and DISTS (Ding et al., 2020) measures structural and textural similarity; lower values are better for both. Appendix C.2 reports PSNR and MS-SSIM (Wang et al., 2003), together with dataset-level FID (Heusel et al., 2017).

4.2 Perceptual Reconstruction across Channel Conditions

Figures 3 and 4 vary SNR at fixed token budget (top) and token budget at 10 dB (bottom).

Response to channel noise. Across the four datasets, AsymVQ-JSCC improves gradually with SNR, with lower LPIPS than DeepJSCC, ProGIC+LDPC, and CDM-JSCC at shared operating

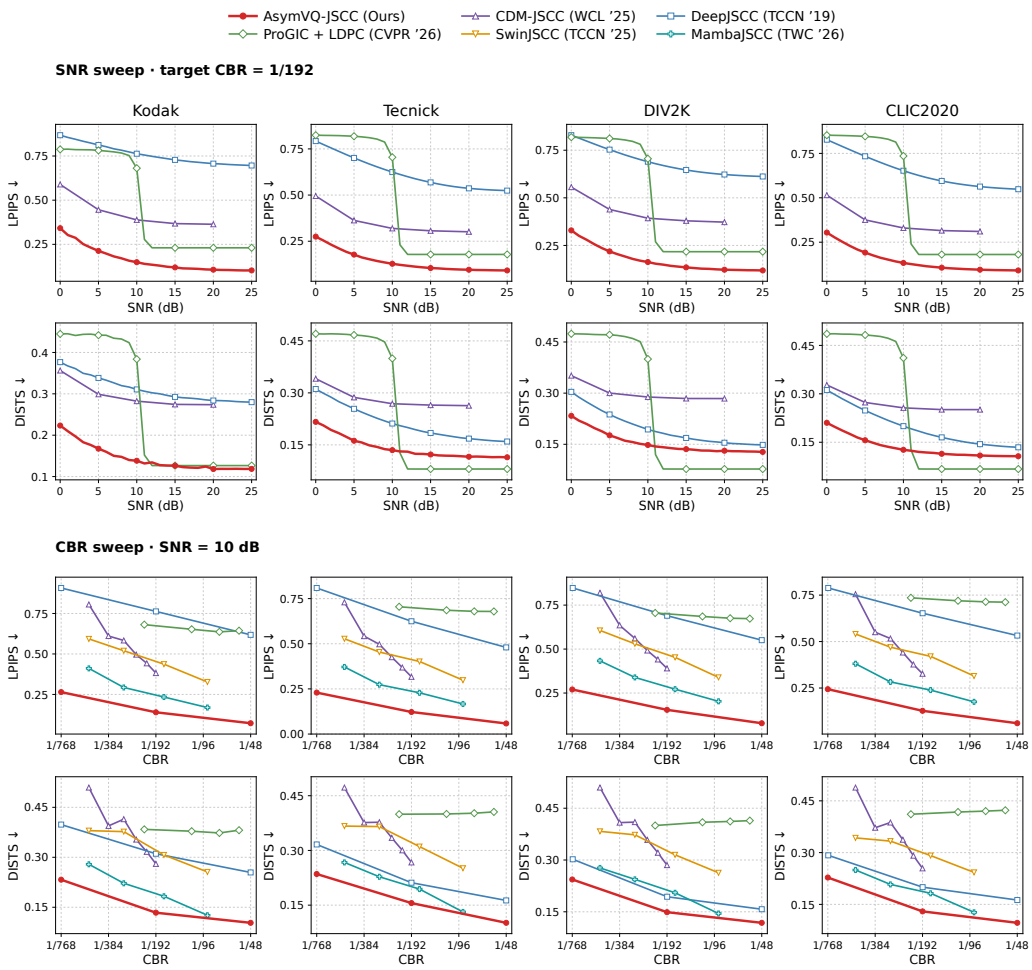

Figure 3: **Perceptual reconstruction under AWGN.** LPIPS/DISTS versus SNR at target CBR $1/192$ (top) and versus CBR at 10 dB (bottom). Lower is better; axes are independent and CBR is logarithmic.

points. ProGIC+LDPC instead exhibits a sharp transition as channel decoding becomes reliable. At CBR $1/192$, Kodak LPIPS/DISTS fall from 0.342/0.223 at 0 dB to 0.102/0.119 at 25 dB under AWGN; the Rayleigh values fall from 0.345/0.236 to 0.122/0.137. This response is consistent with reconstruction without a hard receiver decision.

Perceptual quality and token density. At 10 dB, increasing token density improves both perceptual measures. For Kodak, CBRs $1/768$, $1/192$, and $1/48$ give AWGN LPIPS values 0.265, 0.140, and 0.072 and DISTS values 0.233, 0.134, and 0.103; Rayleigh LPIPS is 0.302, 0.188, and 0.110. Other datasets follow this trend. Appendices C.3 and C.4 examine design sensitivity and qualitative robustness.

Receiver constraint. At a fixed checkpoint, projecting receiver latents onto the source codebook costs 1.46–2.57 dB PSNR and raises LPIPS by 0.063–0.077 on Kodak. Appendix C.5 reports the paired inference intervention and noise-free references.

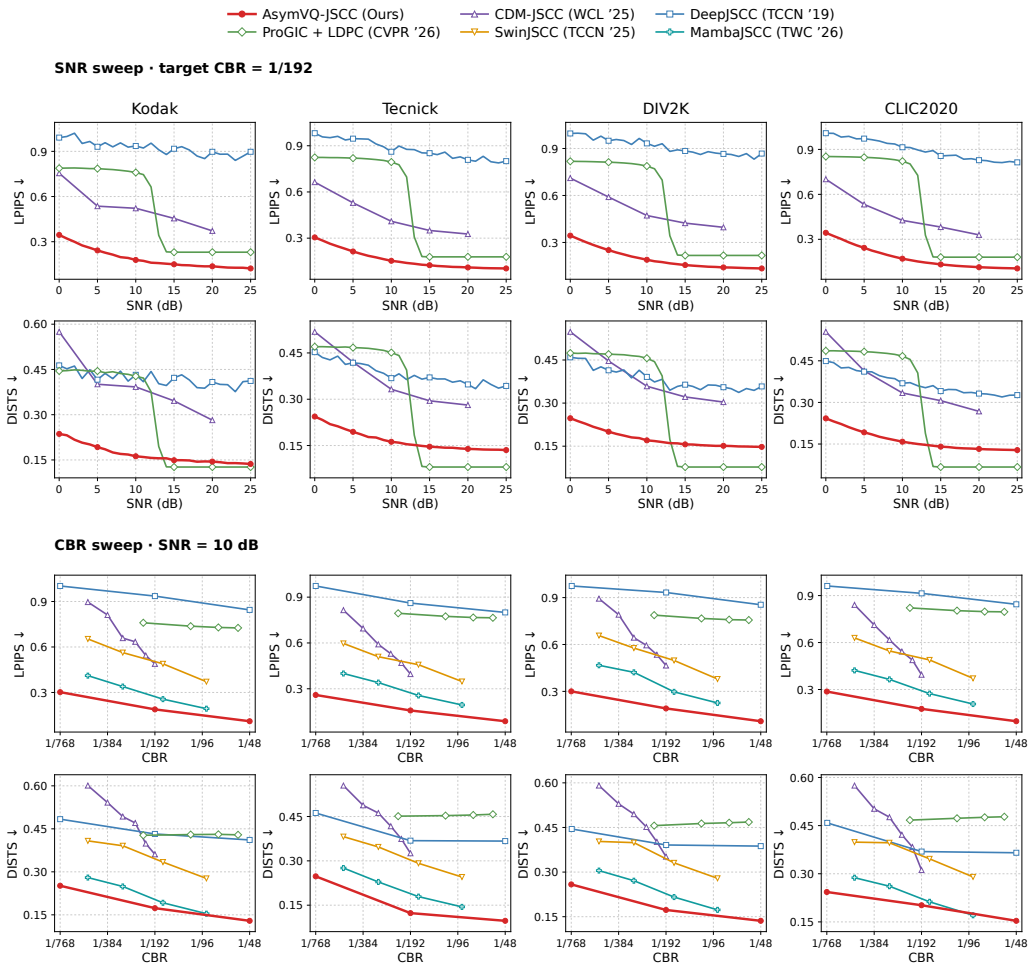

Figure 4: **Perceptual reconstruction under Rayleigh fading.** Arrangement and metric directions follow Figure 3.

5 Conclusions

AsymVQ-JSCC separates the discrete transmission vocabulary from continuous reconstruction latents. Jointly learned irregular QAM and continuous reception connect physical neighborhood geometry to semantic recovery. The affine analysis characterizes the cost of codeword-constrained decoding; experiments show a gradual perceptual response to noise and improved quality with token density. This design makes reconstruction, rather than exact token identity, the objective of the communication interface.

6 Limitations

The analysis uses an affine reconstruction surrogate; the full decoder is nonlinear. The affine, token-wise receiver leaves richer observation models and receivers to future work.

Appendix roadmap. Appendices A–C provide the analysis, training details, and evaluation protocols with supplementary results, respectively.

AI USE STATEMENT

ChatGPT was used only to aid and polish the writing of the manuscript, and for literature retrieval and discovery to identify related work that might otherwise be overlooked. It was not used for research ideation or execution, implementation or debugging, experiments, mathematical proofs, or dataset generation. The authors take full responsibility for all research decisions, interpretations, and the final content, including any AI-assisted material.

ETHICS STATEMENT

This work studies image communication using existing image datasets. Perceptual reconstruction can alter fine details or produce plausible content that differs from the source; applications requiring faithful visual evidence should account for this behavior. Dataset use and any redistribution remain subject to the original terms.

REPRODUCIBILITY STATEMENT

Section 3 specifies the architecture, transmission interface, and training objectives; Section 4 describes evaluation settings and metrics. Appendix A gives the analysis assumptions and derivations, and Appendix B details notation, optimization, and power conventions. Appendix C documents baseline implementations, training and evaluation protocols, channel conventions, and additional quantitative and qualitative results.

REFERENCES

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Yufei Bo, Yiheng Duan, Shuo Shao, and Meixia Tao. Learning based joint coding-modulation for digital semantic communication systems. In *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6. IEEE, 2022.

Yufei Bo, Yiheng Duan, Shuo Shao, and Meixia Tao. Joint coding-modulation for digital semantic communications via variational autoencoder. *IEEE Transactions on Communications*, 72(9): 5626–5640, 2024.

Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz. Deep joint source-channel coding for wireless image transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3):567–579, 2019.

Hao Cao, Chengbin Liang, Wenqi Guo, Zhijin Qin, and Jungong Han. ProGIC: Progressive and lightweight generative image compression with residual vector quantization. In *Findings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.02897.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7936–7945. IEEE, 2020.

Kristy Choi, Kedar Tatwawadi, Aditya Grover, Tsachy Weissman, and Stefano Ermon. Neural joint source-channel coding. In *International Conference on Machine Learning*, pp. 1182–1192. PMLR, 2019.

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

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. URL https://arxiv.org/abs/2004.07728.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Nikos G Evgenidis, Nikos A Mitsiou, Sotiris A Tegos, Panagiotis D Diamantoulakis, Panagiotis Sarigiannidis, Ioannis Krikidis, and George K Karagiannidis. Hybrid semantic-shannon communications. *IEEE Transactions on Wireless Communications*, 23(9):10926–10940, 2024.

Nariman Farvardin. A study of vector quantization for noisy channels. *IEEE Transactions on Information Theory*, 36(4):799–809, 1990.

Maria Fresia, Fernando Peréz-Cruz, H Vincent Poor, and Sergio Verdú. Joint source and channel coding. *IEEE Signal Processing Magazine*, 27(6):104–113, 2010.

Qifan Fu, Huiqiang Xie, Zhijin Qin, Gregory Slabaugh, and Xiaoming Tao. Vector quantized semantic communication system. *IEEE Wireless Communications Letters*, 12(6):982–986, 2023.

Lei Guo, Wei Chen, Yuxuan Sun, and Bo Ai. Digital-sc: Digital semantic communication with adaptive network split and learned non-linear quantization. *IEEE Transactions on Cognitive Communications and Networking*, 11(4):2499–2511, 2024.

Raouf Hamzaoui, Vladimir Stankovic, and Zixiang Xiong. Optimized error protection of scalable image bit streams [advances in joint source-channel coding for images]. *IEEE Signal Processing Magazine*, 22(6):91–107, 2005.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Qiyu Hu, Guangyi Zhang, Zhijin Qin, Yunlong Cai, Guanding Yu, and Geoffrey Ye Li. Robust semantic communications with masked vq-vae enabled codebook. *IEEE Transactions on Wireless Communications*, 22(12):8707–8722, 2023.

Jianhao Huang, Kai Yuan, Chuan Huang, and Kaibin Huang. D^2-jscc: Digital deep joint source-channel coding for semantic communications. *IEEE Journal on Selected Areas in Communications*, 43(4):1246–1261, 2025.

Yoon Huh, Hyowoon Seo, and Wan Choi. Universal joint source-channel coding for modulation-agnostic semantic communication. *IEEE Journal on Selected Areas in Communications*, 43(7): 2560–2574, 2025.

David Burth Kurka and Deniz Gündüz. Deepjscc-f: Deep joint source-channel coding of images with feedback. *IEEE journal on selected areas in information theory*, 1(1):178–193, 2020.

David Burth Kurka and Deniz Gündüz. Bandwidth-agile image transmission with deep joint source-channel coding. *IEEE Transactions on Wireless Communications*, 20(12):8081–8095, 2021.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11513–11522. IEEE, 2022.

Chuanhong Liu, Caili Guo, Yang Yang, Wanli Ni, and Tony QS Quek. Ofdm-based digital semantic communication with importance awareness. *IEEE Transactions on Communications*, 72(10): 6301–6315, 2024.

Zian Meng, Qiang Li, Wenqian Tang, Mingdie Yan, and Xiaohu Ge. Channel-aware vector quantization for robust semantic communication on discrete channels. *arXiv preprint arXiv:2510.18604*, 2025.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33:11913–11924, 2020.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

Mahyar Nemati, Jihong Park, and Jinho Choi. Vq-vae empowered wireless communication for joint source-channel coding and beyond. In *GLOBECOM 2023-2023 IEEE Global Communications Conference*, pp. 3155–3160. IEEE, 2023.

Aria Nosratinia, Jin Lu, and Behnaam Aazhang. Source-channel rate allocation for progressive transmission of images. *IEEE Transactions on Communications*, 51(2):186–196, 2003.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Yashas Malur Saidutta, Afshin Abdi, and Faramarz Fekri. Joint source-channel coding over additive noise analog channels using mixture of variational autoencoders. *IEEE Journal on Selected Areas in Communications*, 39(7):2000–2013, 2021.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Junyong Shin, Yongjeong Oh, Jinsung Park, Joohyuk Park, and Yo-Seb Jeon. Esc-mvq: End-to-end semantic communication with multi-codebook vector quantization. *IEEE Transactions on Wireless Communications*, 2025.

Tze-Yang Tung, David Burth Kurka, Mikolaj Jankowski, and Deniz Gündüz. Deepjscc-q: Constellation constrained deep joint source-channel coding. *IEEE Journal on Selected Areas in Information Theory*, 3(4):720–731, 2022.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Mengyang Wang, Jiahui Li, Mengyao Ma, and Xiaopeng Fan. Constellation design for deep joint source-channel coding. *IEEE Signal Processing Letters*, 29:1442–1446, 2022.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multi-scale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems and Computers*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

Zhenzi Weng and Zhijin Qin. Semantic communication systems for speech transmission. *IEEE Journal on Selected Areas in Communications*, 39(8):2434–2444, 2021.

Tong Wu, Zhiyong Chen, Meixia Tao, Yaping Sun, Xiaodong Xu, Wenjun Zhang, and Ping Zhang. Mambajscc: adaptive deep joint source–channel coding with generalized state space model. *IEEE Transactions on Wireless Communications*, 25:9264–9279, 2025.

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

Songjie Xie, Shuai Ma, Ming Ding, Yuanming Shi, Mingjian Tang, and Youlong Wu. Robust information bottleneck for task-oriented communication with digital modulation. *IEEE Journal on Selected Areas in Communications*, 41(8):2577–2591, 2023.

Ke Yang, Sixian Wang, Jincheng Dai, Kailin Tan, Kai Niu, and Ping Zhang. Witt: A wireless image transmission transformer for semantic communications. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Ke Yang, Sixian Wang, Jincheng Dai, Xiaoqi Qin, Kai Niu, and Ping Zhang. Swinjscc: Taming swin transformer for deep joint source-channel coding. *IEEE Transactions on Cognitive Communications and Networking*, 11(1):90–104, 2024.

Mingyu Yang, Bowen Liu, Boyang Wang, and Hun-Seok Kim. Diffusion-aided joint source channel coding for high realism wireless image transmission. *IEEE Transactions on Machine Learning in Communications and Networking*, 3:1227–1243, 2025a.

Pujing Yang, Guangyi Zhang, and Yunlong Cai. Rate-adaptive generative semantic communication using conditional diffusion models. *IEEE Wireless Communications Letters*, 14(2):539–543, 2025b. doi: 10.1109/LWC.2024.3515656.

Niu Yi, Xu Tianyi, Ma Mingming, and Wang Xinkun. Hypervq: Enabling hyperprior entropy modeling for vq-based generative image compression. *arXiv preprint arXiv:2512.07192*, 2025.

Guangyi Zhang, Hanlei Li, Yunlong Cai, Qiyu Hu, Guanding Yu, and Runmin Zhang. Learned image transmission with hierarchical variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13215–13223, 2025a.

Liang Zhang, Danlan Huang, Xinyi Zhou, Feng Ding, Sheng Wu, and Zhiqing Wei. Channel-adaptive wireless image semantic transmission with learnable prompts. In *2024 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6. IEEE, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023a.

Maojun Zhang, Haotian Wu, Guangxu Zhu, Richeng Jin, Xiaoming Chen, and Deniz Gündüz. Semantics-guided diffusion for deep joint source-channel coding in wireless image transmission. *IEEE Transactions on Wireless Communications*, 2025b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Wenyu Zhang, Haijun Zhang, Hui Ma, Hua Shao, Ning Wang, and Victor CM Leung. Predictive and adaptive deep coding for wireless image transmission in semantic communication. *IEEE Transactions on Wireless Communications*, 22(8):5486–5501, 2023b.

Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22741–22750. IEEE, 2023.

A Analysis of Geometry and Continuous Recovery

A.1 Proof of Proposition 1

Consider the single-token surrogate of Section 3.2, with fixed source codebook, constellation, and affine reconstruction map $g(\mathbf{v}) = \mathbf{B}\mathbf{v} + \mathbf{b}$. Assume $\sigma^2 > 0$ and $\pi_i \geq 0$ with $\sum_i \pi_i = 1$. The conditional probabilities in Eq. (4) follow directly from Bayes' rule and the isotropic Gaussian likelihood.

For a fixed observation $\mathbf{y}$, abbreviate $p_i = p_i(\mathbf{y})$ and $\boldsymbol{\mu} = \sum_i p_i \mathbf{e}_i$. Expanding the conditional risk around $\boldsymbol{\mu}$ gives

$$\begin{aligned} \mathcal{R}(\mathbf{v} \mid \mathbf{y}) &= \sum_i p_i \|\mathbf{B}(\mathbf{v} - \mathbf{e}_i)\|_2^2 \\ &= \|\mathbf{B}(\mathbf{v} - \boldsymbol{\mu})\|_2^2 + \sum_i p_i \|\mathbf{B}(\mathbf{e}_i - \boldsymbol{\mu})\|_2^2, \end{aligned} \tag{15}$$

because the cross term contains $\sum_i p_i(\boldsymbol{\mu} - \mathbf{e}_i) = \mathbf{0}$. The remaining variance term satisfies

$$\begin{aligned} \sum_i p_i \|\mathbf{B}(\mathbf{e}_i - \boldsymbol{\mu})\|_2^2 &= \sum_i p_i \|\mathbf{B}\mathbf{e}_i\|_2^2 - \|\mathbf{B}\boldsymbol{\mu}\|_2^2 \\ &= \frac{1}{2} \sum_{i,j} p_i p_j \|\mathbf{B}(\mathbf{e}_i - \mathbf{e}_j)\|_2^2. \end{aligned} \tag{16}$$

This proves Eq. (5). The unconstrained minimum is attained at $\mathbf{v} = \boldsymbol{\mu}$, or at any $\mathbf{v}$ with $\mathbf{B}\mathbf{v} = \mathbf{B}\boldsymbol{\mu}$. Restricting $\mathbf{v}$ to the finite set $\mathcal{C}$ yields Eq. (6). The penalty is positive exactly when $\mathbf{B}\boldsymbol{\mu} \notin \{\mathbf{B}\mathbf{e}_j\}_{j=1}^K$. A nearest-symbol receiver need not attain even this best codeword-constrained risk, since minimum symbol error and minimum reconstruction error need not select the same codeword.

A.2 A Two-Codeword Example

Suppose $K = 2$, $p_1(\mathbf{y}) = p_2(\mathbf{y}) = 1/2$, and $\Delta_{12} > 0$. Selecting either codeword gives conditional reconstruction risk $\Delta_{12}/2$. The intermediate estimate $\boldsymbol{\mu} = (\mathbf{e}_1 + \mathbf{e}_2)/2$ gives risk $\Delta_{12}/4$. Thus, in this surrogate, a representation outside the codebook halves the conditional squared error at an equally ambiguous observation. This is an illustrative conditional-risk calculation, not a prediction of a fixed percentage gain in image experiments.

More generally, $p_1(\mathbf{y})p_2(\mathbf{y})\Delta_{12}$ separates the frequency and strength of ambiguity from its reconstruction cost. Physical distances and source priors determine the conditional weights, while the source representations and reconstruction map determine Δ_{12}. The relevant constellation objective is therefore sensitive to source relationships, rather than only to symbol separability.

A.3 Scope of the Analysis

The target in Proposition 1 is an affine reconstruction of a discrete source token. It does not include source quantization error, the spatial dependencies of the full image decoder, or the perceptual and adversarial objectives. The result characterizes the cost of a discrete reconstruction constraint in that surrogate; it does not establish Bayes optimality or universal superiority of a trained AsymVQ-JSCC model.

In particular, the actual receiver $R(\mathbf{y}) = \mathbf{W}_r\mathbf{y} + \mathbf{b}_r$ is affine. Its outputs lie in an affine subspace of dimension at most two, even when $d > 2$. Continuous reception removes membership in the finite source codebook as a requirement; it does not imply that the receiver spans all of $\mathbb{R}^d$ or realizes every conditional mean. The nonlinear, spatial decoder and the physical mapping are jointly trained with this receiver. Likewise, continuous reception can be applied to a regular constellation. The analysis motivates learning source-aware geometry for recovery, rather than claiming that irregularity alone guarantees improvement.

B Training Details and Notation

B.1 Notation and Module Boundaries

Lowercase bold symbols denote individual vectors or images, and uppercase bold symbols denote collections of tokens or learned matrices. The source vocabulary and physical constellation are distinguished as $\mathcal{C}$ and $\mathcal{S}$. The same index i identifies the source codeword $\mathbf{e}_i$ and its associated physical point $\mathbf{s}_i$; a token position is indexed by t. We use K for their common cardinality because every source codeword is associated with one constellation point. The codeword dimension is d, and N is the number of transmitted symbols.

Table 1: Notation for source representation, transmission, and reconstruction.

Symbol	Meaning
$\mathbf{x}, \hat{\mathbf{x}}$	Input image and noisy-channel reconstruction
E, G	Source encoder and image decoder, including latent projections
M, R	Normalized physical mapper and continuous receiver
$\mathbf{z}_t, \mathbf{z}_t^q$	Encoder token and selected source-codeword vector
$\mathcal{C}, \mathbf{e}_i$	Source codebook and its ith entry in $\mathbb{R}^d$
$\mathcal{S}, \mathbf{s}_i$	Physical constellation and its ith point in $\mathbb{R}^2$
i_t	Source index selected at token position t
$\mathbf{y}_t, \hat{\mathbf{z}}_t$	Channel observation and receiver latent
$\pi_i, p_i(\mathbf{y})$	Codeword usage prior and analysis-only conditional probability
$\hat{\mathbf{x}}^{(0)}$	Noiseless auxiliary reconstruction
f, N, d, K	Downsampling factor, token count, latent dimension, vocabulary size
P, γ, σ^2	Average symbol power, SNR in dB, noise variance per real axis

The encoder-side 1×1 latent projection is included in E, and the decoder-side 1×1 projection is included in G. Both projections are part of the source codec and are shared between noiseless and noisy reconstruction views. This notation suppresses internal network blocks while preserving the distinction between source-VQ latents, physical observations, and receiver latents.

B.2 Soft Assignments for Initialization and Joint Gradients

For AWGN-based mapper initialization, $\mathbf{y}_t = \mathbf{s}_{i_t} + \mathbf{n}_t$ and the weights in Eq. (10) are

$$\omega_{tj} = \frac{\exp(-\|\mathbf{y}_t - \mathbf{s}_j\|_2^2/\tau_m)}{\sum_k \exp(-\|\mathbf{y}_t - \mathbf{s}_k\|_2^2/\tau_m)}. \tag{17}$$

The source assignment used for the straight-through gradients in Eq. (12) is

$$a_{ti} = \frac{\exp(-\|\mathbf{z}_t - \mathbf{e}_i\|_2^2/\tau_s)}{\sum_j \exp(-\|\mathbf{z}_t - \mathbf{e}_j\|_2^2/\tau_s)}. \tag{18}$$

The positive temperatures τ_m and τ_s control assignment sharpness in physical and source space, respectively: smaller values concentrate weights on closer candidates. Transmission power is set independently by the normalization in Eq. (9). The physical weights ω_{tj} contain no usage-prior factor and therefore differ from the posterior $p_i(\mathbf{y})$ in Proposition 1. They weight source-space distortion during initialization and are absent from Stage II and inference. A hard proxy decision is used only to log the Stage I error rate; it does not enter the training loss.

B.3 Optimization Schedule

The two method stages organize four successive optimization steps:

1. **Source-codec pretraining.** Optimize $E, G, \mathcal{C}$ with image reconstruction, source-VQ, perceptual, and adversarial losses.
2. **Constellation initialization.** Freeze $E, G, \mathcal{C}$ and optimize M with the soft-confusion surrogate and separation regularization. Save the resulting numerical constellation.

3. **Receiver warmup.** Freeze the source codec and numerical constellation, and optimize only R using noiseless image MSE. VQ, LPIPS, and adversarial objectives are inactive in this step.
4. **Joint JSCC training.** Optimize $E, G, \mathcal{C}, M, R$ using Eq. (13). Recompute the constellation as the source codebook and mapper change.

The AWGN reference configuration uses $K = 256$, $d = 4$, and $P = 1$. Supported downsampling factors are $f \in \{4, 8, 16\}$. We use $w_n = w_c = 1$, $\lambda_p = 1$, $\lambda_g = 0.1$, $\beta = 0.25$, $m_{\mathrm{nn}} = 0.05$, $\kappa = 2.5$, $\lambda_{\mathrm{nn}} = 1$, and $\lambda_{\mathrm{papr}} = 0.05$. Constellation initialization uses $\lambda_{\mathrm{sep}} = 0.01$ and no additional peak-power penalty. In joint training, the all-pairs separation and additional peak-power penalties have zero weight; only the nearest-neighbor and uniform-PAPR terms are active. The coefficient β multiplies the second term in Eq. (8), following the source-VQ convention in the method specification. Pixel and latent squared-error losses are expressed as means over their elements, while the source-space distances in Eq. (10) are squared Euclidean distances summed over the d features.

During joint training, source assignments update the empirical prior π_i through an exponential moving average. The prior is not updated during validation or inference. For this reference configuration, training SNR is sampled uniformly on $[0, 25]$ dB, and validation uses the grid $\{0, 5, 10, 15, 20, 25\}$ dB. Checkpoint selection uses the mean noisy-channel PSNR on this grid.

B.4 Adversarial Supervision and Power Conventions

Let D denote the PatchGAN discriminator, with its output averaged over patch scores when used in the generator objective. For either reconstructed view $\tilde{\mathbf{x}}$, the generator term is $\ell_{\mathrm{adv}}(\tilde{\mathbf{x}}) = -D(\tilde{\mathbf{x}})$. The discriminator is optimized separately with the real-image hinge term $[1 - D(\mathbf{x})]_+$ and fake-image hinge term $[1 + D(\mathrm{sg}(\tilde{\mathbf{x}}))]_+$, summing the discriminator objectives for noiseless and noisy reconstructions. In the reference joint-training configuration, adversarial terms are inactive during the first 1,000 steps. The discriminator is absent at inference.

The AWGN variance uses the source-usage-weighted power $P = \sum_i \pi_i \|\mathbf{s}_i\|_2^2$. The uniform-weighted PAPR diagnostic instead uses

$$\mathrm{PAPR}_u(\mathcal{S}) = \frac{\max_i \|\mathbf{s}_i\|_2^2}{K^{-1} \sum_i \|\mathbf{s}_i\|_2^2}. \tag{19}$$

These are different averaging conventions. For a real two-dimensional symbol, $\mathbb{E}\|\mathbf{n}_t\|_2^2 = 2\sigma^2$, explaining the factor of two in Eq. (2). The nominal index length is $\log_2 K$ bits per token when K is a power of two, giving $\log_2 K/f^2$ raw bits per spatial pixel. This nominal index count is distinct from the physical channel-use budget and from reliable payload throughput.

Numerical details. The implementation clamps the proxy temperature below at 10^{-6} and stabilizes the uniform-PAPR denominator with a positive numerical floor. The all-pairs separation penalty differs from the nearest-neighbor term in Eq. (14); only the latter is active in joint training.

C Evaluation Protocols and Supplementary Results

C.1 Baseline Implementations and Data Sources

Table 2: Selected evaluated systems and transmission interfaces. ProGIC+LDPC denotes the local reproduction and communication wrapper.

System	Transmitted representation	Receiver
AsymVQ-JSCC	One learned irregular-QAM symbol per source-VQ token	Continuous latent estimation and image decoding
DeepJSCC	Continuous neural channel symbols	Neural image reconstruction
ProGIC+LDPC	Fixed-width RVQ indices, rate-$1/2$ LDPC, Gray 64-QAM	Soft demodulation, LDPC decoding, and source reconstruction
CDM-JSCC	Rate-adapted released latent representation	Conditional diffusion reconstruction, 17 sampling steps

Local neural and coded baselines. DeepJSCC is evaluated convolutionally at the original image resolution, with padding removed before image metrics are computed. Its AWGN-trained checkpoints are selected by validation MSE and evaluated at SNRs from 0 to 25 dB. The ProGIC reproduction uses fixed-width 10-bit RVQ indices and IEEE 802.11n/ac LDPC blocks with 972 information bits and 1,944 coded bits. Gray 64-QAM transmits six coded bits per complex symbol. Thus a payload of B bits occupies $324\lceil B/972 \rceil$ complex uses, including parity and padding. The decoder uses normalized min-sum with a maximum of 30 iterations; decoded source indices are passed to the image codec without an outer-CRC rejection gate. Both local baselines use the recorded evaluation seed 3407 and complete sets of 24 Kodak, 100 Tecnick, 100 DIV2K, and 428 CLIC2020 images.

The local metrics use LPIPS with AlexNet. DeepJSCC evaluates LPIPS at native resolution and DISTS after area resizing to 256×256; ProGIC evaluates both at native resolution. Their FID implementation accumulates Inception features over the dataset rather than averaging a per-image statistic. Consequently, the FID curves describe the reconstruction distribution within each recorded dataset and evaluation protocol.

Diffusion baseline. The CDM-JSCC records use the released 128-pixel model with a rate adapter and a 17-step sampler. Inputs are center-square crops resized to 256×256; its outputs are resized to this evaluation domain. LPIPS uses AlexNet, and DISTS is evaluated on the same domain. Its CLIC results use CLIC2020, consistent with the CLIC evaluation used for the other compared systems.

Rayleigh fading and receiver information. For the Rayleigh experiments, all compared methods are evaluated under the same complex block-fading model and receiver-CSI assumption at each operating point. Receiver CSI is provided consistently to all methods, and the received symbols are equalized using the same channel coefficient before method-specific receiver processing. SNR is defined before fading, and the same channel-use accounting is used throughout the comparison. Method-specific neural decoding architectures are otherwise left unchanged.

SwinJSCC and MambaJSCC baselines. SwinJSCC and MambaJSCC are trained and evaluated under the dataset and channel settings used for the corresponding comparisons in Figures 3–4. Their Kodak, Tecnick, DIV2K, and CLIC results are obtained from direct model evaluations; the CLIC experiments use CLIC2020. Appendix figures report only the metrics available from the corresponding evaluation records.

C.2 Additional Reconstruction Metrics

PSNR and MS-SSIM characterize pixel fidelity and structural similarity; FID describes reconstruction-distribution similarity. Larger PSNR and MS-SSIM and smaller FID are preferable. Each panel retains its own numerical range. The appendix figures report the additional metrics available from the corresponding evaluation records.

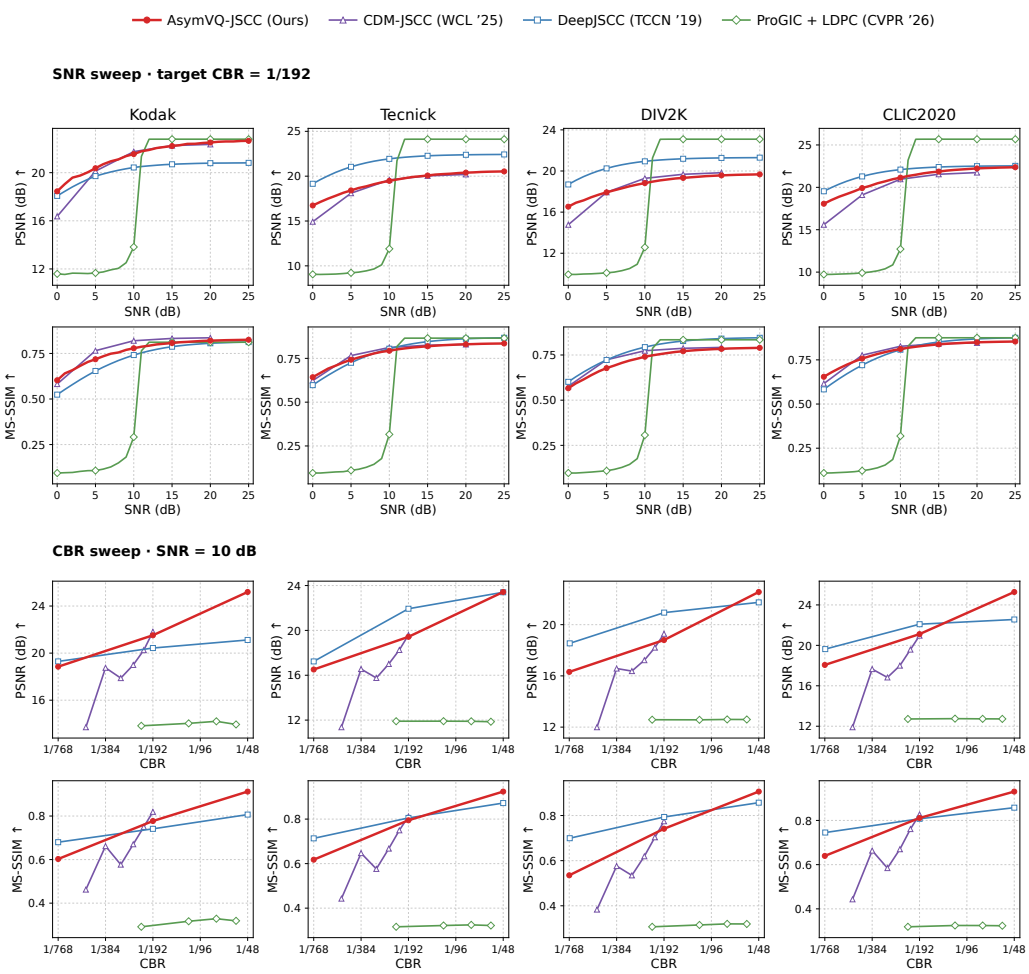

Figure 5: **Pixel and structural fidelity under AWGN.** PSNR and MS-SSIM versus SNR at target CBR $1/192$ (top) and versus CBR at 10 dB (bottom). Larger values are better. Every panel has independent axes; CBR is logarithmic. Protocols follow Appendix C.

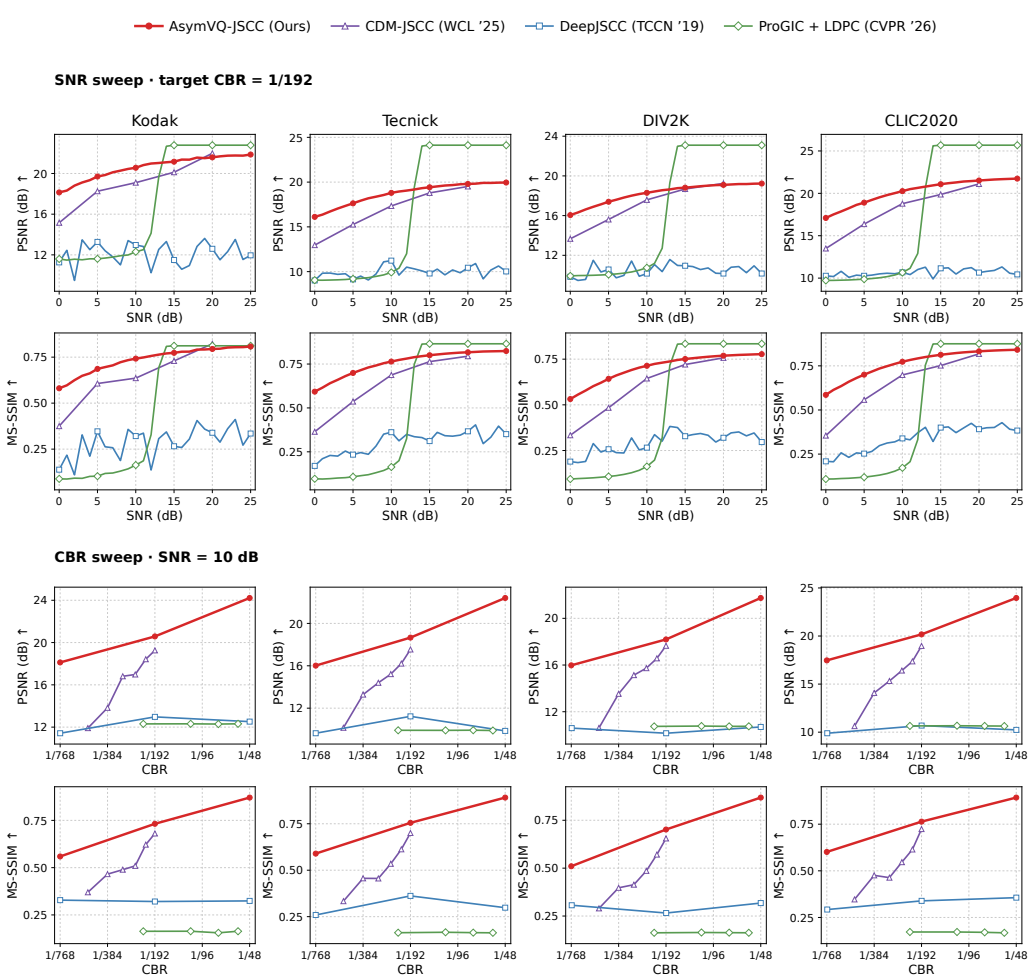

Figure 6: **Pixel and structural fidelity under Rayleigh fading.** PSNR and MS-SSIM versus SNR at target CBR $1/192$ (top) and versus CBR at 10 dB (bottom). Larger values are better. Every panel has independent axes; CBR is logarithmic. Protocols follow Appendix C.

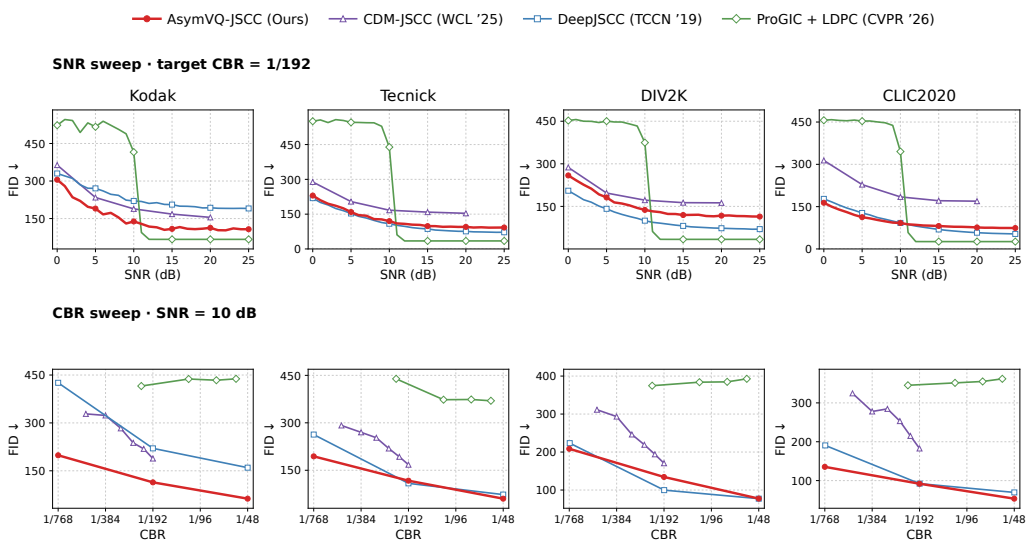

Figure 7: **Reconstruction-distribution similarity under AWGN.** Dataset-level FID versus SNR at target CBR $1/192$ (top) and versus CBR at 10 dB (bottom). Lower values are better. Every panel has independent axes; CBR is logarithmic.

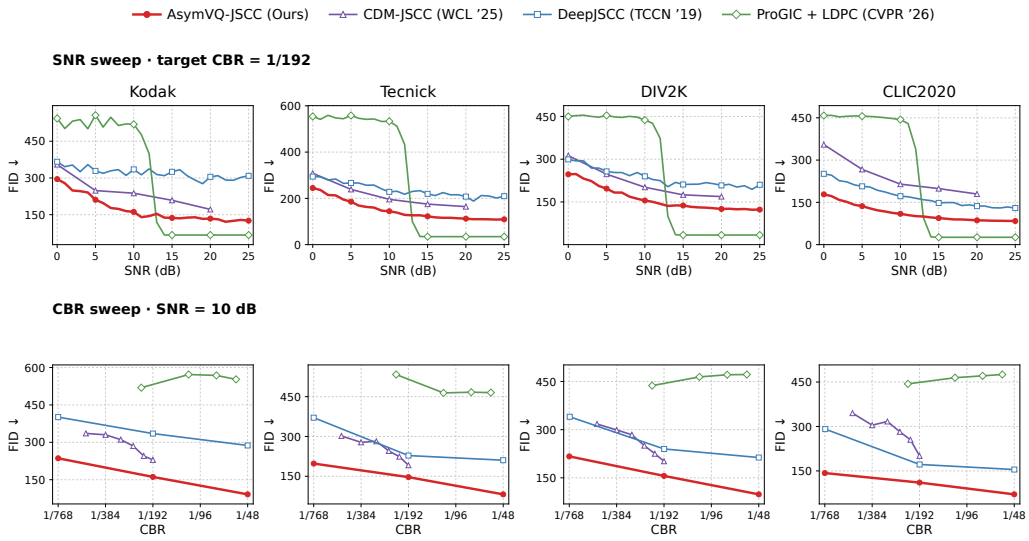

Figure 8: **Reconstruction-distribution similarity under Rayleigh fading.** Dataset-level FID versus SNR at target CBR $1/192$ (top) and versus CBR at 10 dB (bottom). Lower values are better. Every panel has independent axes; CBR is logarithmic.

C.3 Sensitivity to Representation and Mapping Choices

The sensitivity plots examine vocabulary/constellation cardinality, mapper activation, and the PAPR setting. Unless otherwise stated, all studies in this subsection are conducted on Kodak under AWGN, with the original SNR samples and variant labels retained.

Vocabulary and constellation cardinality. At CBR $1/192$, the study ranges from 64 to 2,048 source entries and associated symbols. The 256-point variant combines low LPIPS with a broad SNR response; increasing cardinality beyond 256 does not uniformly improve the curve. More entries change the representation and geometry without increasing the number of transmitted symbols.

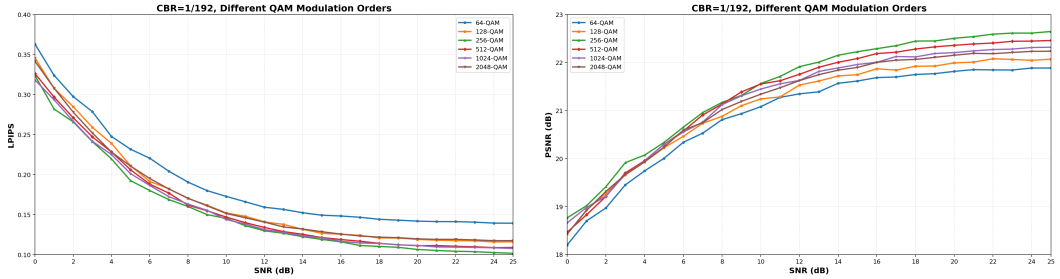

Figure 9: **Vocabulary and constellation cardinality.** LPIPS (left) and PSNR (right) at CBR $1/192$, with 64–2,048 constellation points. Lower LPIPS and higher PSNR are preferable.

Mapper activation. The 256-point comparison contrasts "w/o Act." and "SiLU" at CBRs $1/48$ and $1/192$. The activation-free variant has lower LPIPS across the displayed SNR range at both rates. This concerns the mapper parameterization, not regular versus irregular constellations or continuous versus hard-decision receivers.

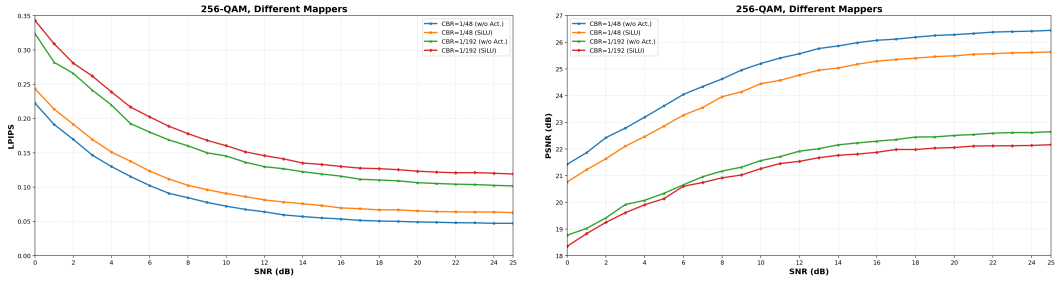

Figure 10: **Mapper activation.** LPIPS (left) and PSNR (right) for the activation-free and SiLU variants, with 256 constellation points at CBRs $1/48$ and $1/192$.

PAPR setting. At CBR $1/48$, the LPIPS curves for settings 2.0 and 2.5 remain close over much of the range; relaxing the setting further does not consistently improve perceptual quality. The setting 2.5 matches the reference regularizer in Appendix B. These labels specify compared settings rather than measured hardware waveform PAPR.

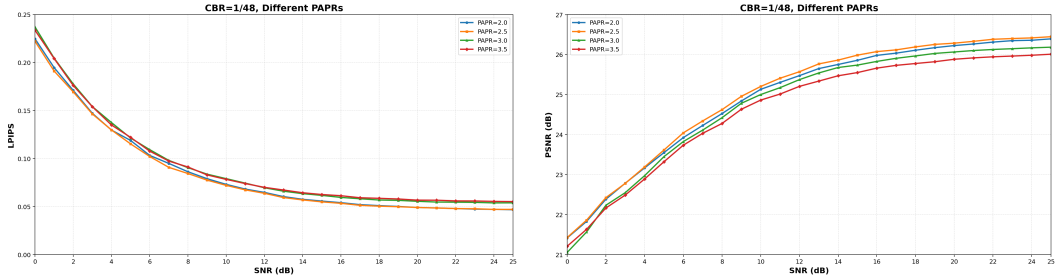

Figure 11: **PAPR sensitivity.** LPIPS (left) and PSNR (right) at CBR $1/48$, retaining the source settings 2.0, 2.5, 3.0, and 3.5.

C.4 Qualitative Robustness

Figures 12–15 show original inputs and reconstructions at 3, 6, 9, 12, 15, 18, and 21 dB. The Kodak AWGN rows use identical 256×256 center crops and CBR $1/192$, with the CDM-JSCC rate adaptation specified in Figure 12. CBR counts transmitted complex payload symbols; CDM-JSCC retains its implementation's known normalization-scale convention, with no additional signaling cost modeled.

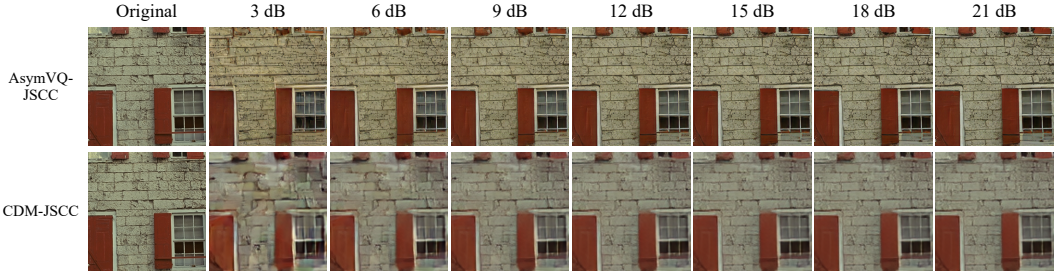

Figure 12: **AWGN reconstructions of Kodak 01 at matched CBR.** Both methods receive the same 256×256 center crop and transmit 1,024 complex payload symbols (CBR $1/192$). CDM-JSCC is rate-adapted through internal 128×128 resizing, its trained uniform 32-real-coordinate branch per patch, and reconstruction resizing to 256×256. Columns show the original and 3, 6, 9, 12, 15, 18, and 21 dB.

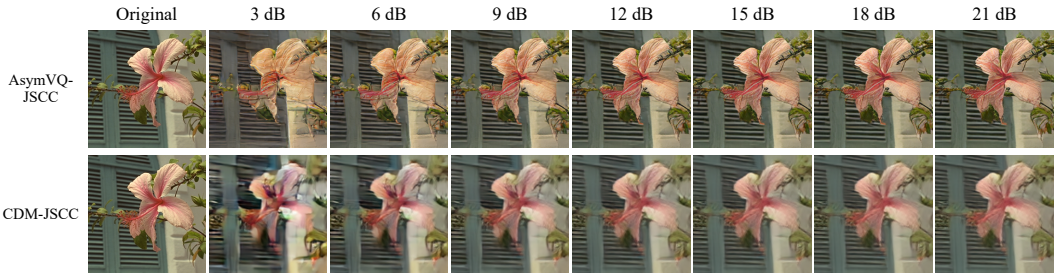

Figure 13: **AWGN reconstructions of Kodak 07 at matched CBR.** The shared input crop, CBR $1/192$, CDM-JSCC rate adaptation, and SNR columns follow Figure 12.

Fading and additional image content. The Rayleigh examples extend the visualization to fading, while the CLIC example follows additional scene content under both channels. The Rayleigh examples follow the common fading and receiver-CSI protocol used in the quantitative evaluation, with channel-specific checkpoints at CBR $1/192$ and bicubic 256×256 inputs. Each image has one complex fading coefficient held fixed across the SNR grid; SNR is defined before fading.

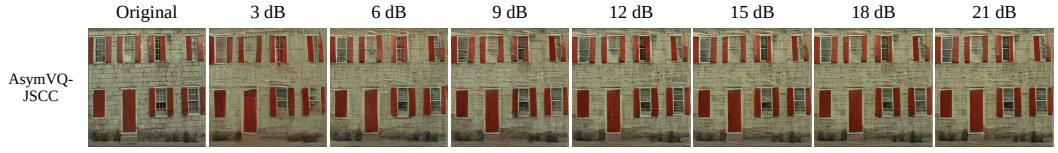

Figure 14: **Rayleigh reconstructions of Kodak 01.** AsymVQ-JSCC at CBR $1/192$ under the common Rayleigh evaluation protocol. Columns show the input and 3, 6, 9, 12, 15, 18, and 21 dB reconstructions.

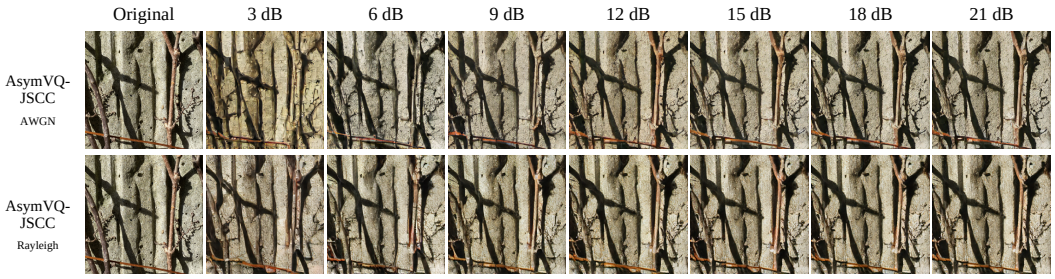

Figure 15: **AWGN and Rayleigh reconstructions on CLIC2020.** AsymVQ-JSCC reconstructions of `img_00001` using the respective channel-specific checkpoints at CBR $1/192$. The original input is bicubically resized to 256×256; SNR columns are 3, 6, 9, 12, 15, 18, and 21 dB.

C.5 Receiver Constraint at a Fixed Checkpoint

We test whether the learned decoder uses receiver latents beyond the source codebook. At inference, we compare continuous reception $G(R(\mathbf{y}))$ with $G(Q_{\mathcal{C}}(R(\mathbf{y})))$, where $Q_{\mathcal{C}}$ projects each latent onto its nearest source codeword in squared Euclidean distance. All weights remain fixed, and both paths use identical source indices, constellation, power normalization, and received observations. The evaluation uses 24 Kodak images bicubically resized to 256×256, a 256-entry codebook with four-dimensional codewords, CBR $1/192$, and AWGN at 3, 9, 15, and 21 dB.

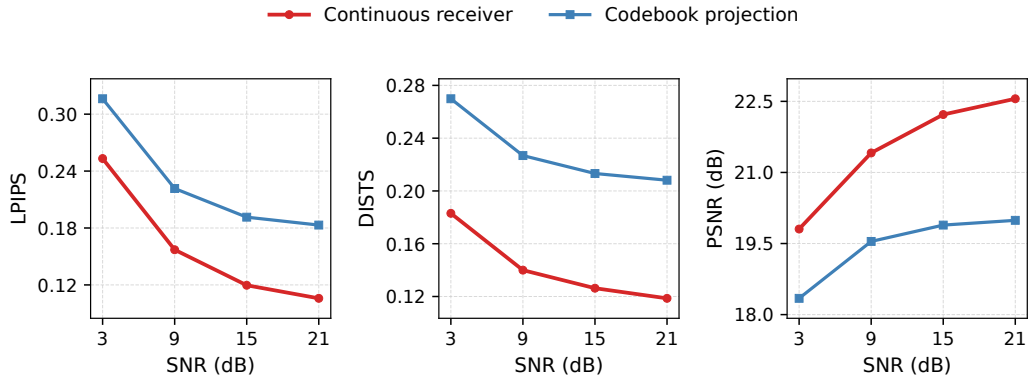

Figure 16: **Effect of enforcing source-codebook membership at inference.** Kodak means for the continuous receiver and nearest-codeword projection. LPIPS and DISTS are lower-is-better; PSNR is higher-is-better. For the intervention evaluation, the transmitter and channel observations are unchanged, and neither receiver path is retrained.

Projection degrades paired PSNR and LPIPS at every tested SNR. At 21 dB, PSNR changes from 22.56 to 19.99 dB and LPIPS from 0.1059 to 0.1831. The final aggregate continuous-receiver DISTS is 0.1186, compared with 0.2081 for the projected path in the intervention evaluation. Table 3 shows a comparable cost without channel noise: the decoder has learned useful continuous representations outside the source vocabulary. The noiseless gap links this result to the learned receiver–decoder interface, rather than a noise-specific projection effect. This intervention evaluates the codebook constraint on the continuously trained model; it is not a comparison with a separately trained RTR receiver.

Table 3: **Noise-free receiver references.** Kodak means using the same checkpoint and transmitted symbols, with channel noise removed.

Receiver latent	PSNR ↑	LPIPS ↓	DISTS ↓
Continuous $R(\mathbf{s})$	22.71	0.1000	0.1174
Projected $Q_{\mathcal{C}}(R(\mathbf{s}))$	20.00	0.1793	0.2079