TARGET FIRST, BACKGROUND LATER: REGION-PROGRESSIVE LATENT COMPRESSION FOR SATELLITE IMAGE TRANSMISSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent expansion of satellite constellations and onboard sensing brings a rapidly increasing volume of satellite images for earth observation collected onboard. Meanwhile, the satellite downlink can hardly fulfill the huge downloading needs due to the short ground-station contact windows and time-varying link conditions. Although neural image compression approaches achieve strong rate-distortion performance for terrestrial scenarios, they are hard to adopt for satellite image compression for two reasons. First, some approaches require fixed-rate codecs which cannot adapt to dynamic satellite downlink environments. Second, those progressive approaches refine the entire image without considering satellite images' explicit spatial priority. In this paper, our insight is that satellite images usually contain a few small but high-value objects with a large and sparse background. This highly nonuniform information distribution inspires us that the earliest transmitted bits should preserve the region of interest (ROI), whereas background information can be deferred and supplemented using historical observations available at the ground station. Thus, we propose Region-Progressive Latent Compression (RPLC), which partitions a quantized latent into spatially masked channel-slice units and serializes them once as an ROI-first bitstream that is progressively decodable at unit boundaries. At the ground-side decoder, a lightweight latent-completion network predicts unreceived ROI elements without additional coded bits, while an aligned historical observation guides background reconstruction through a spatially gated Reference ControlNet. Image-quality evaluation and downstream ROI classification and segmentation experiments on the fMoW dataset demonstrate that RPLC achieves state-of-the-art ROI performance against 24 baselines at ultra-low bit rates, attaining 66.0% reference-assisted top-1 ROI classification accuracy under a paired-gallery protocol and 81.48% Mask IoU at only 0.0089 bpp.

1 INTRODUCTION

The explosive growth of satellite constellations and the significant enhancement of on-board sensing capabilities have generated huge high-value satellite images. However, downloading these images suffers from two bottlenecks. 1) The data volume far exceeds the capacity limit of existing satellite-terrestrial links (Esch et al., 2018; Devaraj et al., 2017; Denby et al., 2023); 2) Influenced by the physical space environment, these links suffer from severe bandwidth fluctuations and unstable connections (Zhang, 2024; Vasisht et al., 2021). Such instability can cause burst packet loss and temporarily interrupt an ongoing image transfer, leaving only an incomplete bitstream at the receiver (Sha et al., 2025; Vasisht et al., 2021).

Although neural image compression demonstrates superior rate–distortion performance (Ballé et al., 2018; Minnen et al., 2018; Cheng et al., 2020b; He et al., 2022) and is explored for in-orbit deployment for satellite images (Zhang et al., 2024b), it still has two issues. 1) **Limited bitrate adaptability.** Existing methods (Ballé et al., 2018; Minnen et al., 2018; Cheng et al., 2020b; He et al., 2022) follow the one-model-per-rate training paradigm. Under unstable satellite-terrestrial links, incomplete bitstream reception leads to decoding failures, making them unsuitable for dynamic conditions. 2) **Region-agnostic bit allocation.** Although some progressive methods (Toderici et al., 2017; Theis

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.



Figure 1: Qualitative comparison at nearby operating points (bpp ≈ 0.01–0.05; each panel lists its bpp). Center: current image with the ROI in the red box. Left: full-image reconstructions, including RPLC at $k = 3$. Right: ROI crops, including RPLC at $k = 5$, the transmitted ROI mask, and the co-located historical reference used by the ground decoder.

et al., 2022; Zhang et al., 2026b) allow for bitstream truncation, they allocate bits uniformly across the entire image, failing to explicitly prioritize the target regions that are critical for downstream analysis. As Figure 1 shows, at comparable ultra-low bitrates such codecs often distort or replace the target structures, whereas RPLC preserves them.

Considering the above two issues, our key insight is that *not all pixels in a satellite image are equally important: task-critical targets are usually sparse and surrounded by a massive background*. Therefore, rather than adopting a uniform progressive strategy, the transmission should be entirely target-driven. By prioritizing the ROI during transmission, the system preserves the target even when the link terminates early. Figure 2 illustrates this target-first transmission principle under an interrupted satellite downlink.

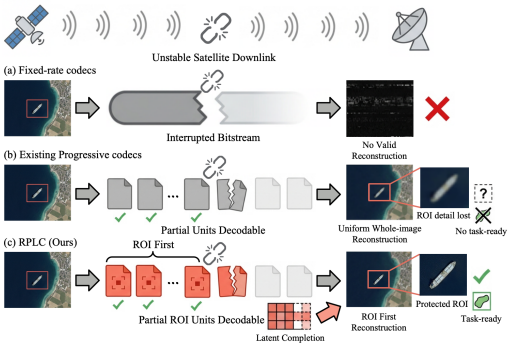

Figure 2: Task setting under an interrupted downlink. Unlike fixed-rate and uniformly progressive codecs, RPLC sends ROI units first and completes missing ROI latents on the ground.

Based on this observation, we propose Region-Progressive Latent Compression (RPLC). RPLC encodes a satellite image into a single ROI-first bitstream that is progressively decodable at unit boundaries, thereby preserving task-critical content when transmission terminates early. For onboard image encoding, a lightweight YOLOE-26n instance-segmentation model fine-tuned on fMoW efficiently generates the ROI mask (Wang et al., 2025; Jocher et al., 2026). The compression encoder then partitions the quantized latent into spatially masked channel-slice units, frames each unit separately, and serializes them once with all ROI units preceding the background units. At the ground-side decoder, a lightweight latent-completion network predicts unreceived ROI elements from the decoded context without additional coded bits. An aligned historical observation, made available by periodic satellite revisits, guides background reconstruction through a spatially gated Reference ControlNet, while background gating and ROI pixel locking prevent the reference path from altering the reconstructed ROI.

We compare RPLC with 24 baselines on fMoW, spanning conventional learned compression, progressive compression, ROI-oriented coding, and codecs commonly used for satellite imagery. RPLC provides leading ROI DISTS and CLIP performance throughout the evaluated range and consistently surpasses the rate-matched satellite baselines in reconstruction, ROI classification, and ROI segmentation. Ablation studies further validate the effectiveness of the proposed components.

2 Related Work

2.1 Remote-Sensing Image Compression.

Satellite codecs must combine coding efficiency, low encoder complexity, and robustness over short, variable contacts (Drusch et al., 2012; Roy et al., 2014; Zhang et al., 2024b). Operational systems rely on JPEG2000 (Skodras et al., 2001) and CCSDS standards (Consultative Committee for Space Data Systems, 2012), which suit constrained hardware but not learned perceptual or task-oriented optimization. Onboard accelerators such as Φ-Sat-1 (Giuffrida et al., 2022) have motivated learned codecs with frequency, non-local, and global-context modeling (Xiang & Liang, 2024; Zhang et al., 2024a), federated onboard training of a codec across a constellation (Gómez & Meoni, 2024), and COSMIC (Zhang et al., 2024b), which pairs a lightweight onboard encoder with ground-side diffusion compensation. These methods improve full-image reconstruction but are fixed-rate or lack a target-prioritized progressive bitstream.

2.2 Natural-Image Learned, Generative, and Progressive Compression.

Learned codecs pair analysis and synthesis transforms with a learned entropy model (Ballé et al., 2017); hyperpriors (Ballé et al., 2018) and autoregressive or spatial–channel contexts (Minnen et al., 2018; Cheng et al., 2020b; He et al., 2021; 2022) further reduce rate, but their representations are not designed for truncation. At very low rates, generative codecs counter over-smoothing (Blau & Michaeli, 2019) with GAN texture (Mentzer et al., 2020) or diffusion denoising (Yang & Mandt, 2023; Careil et al., 2024), but neither adapts the bitrate nor chooses which region to send first. Progressive codecs decode several reconstructions from one stream via recurrent refinement (Toderici et al., 2017), diffusion-based reverse-channel coding (Theis et al., 2022), or coarse-to-fine autoregressive tokens (Zhang et al., 2026b). Trit-plane coding orders latent information by rate–distortion importance (Lee et al., 2022a), refines partially received latents with context models (Jeon et al., 2023), and selective compression masks latent elements by learned importance (Lee et al., 2022b). RPLC extends these progressive-selection and partial-latent ideas from global importance over the complete frame to ordering around a task-defined ROI.

2.3 ROI-Based and Semantically Structured Compression.

ROI and spatially adaptive codecs allocate more bits to salient objects, detections, or semantic segments than to the background (Cai et al., 2020; Song et al., 2021), including mask-guided transformer autoencoders (Li et al., 2023), soft-attention bit allocation (Hu et al., 2025), and text-selected ROIs (Jin et al., 2026), and semantically structured methods such as GIT-SSIC (Feng et al., 2023) use irregular group masks and group-independent transforms for selective transmission. These approaches target object-centric natural images, whereas remote-sensing targets are small and sparse within a dominant background, and none combines ROI-first progression, decoder-side latent completion, and a co-located historical reference; RPLC unifies all three.

3 Method

Figure 3 gives an overview of `RPLC`. Building on a pretrained one-step SANA/DC-AE generative codec, RPLC contributes ROI-first region–channel coding, a latent-completion network, and a reference branch, trained in three successive stages that each freeze the modules trained before.

3.1 Preliminaries and Problem Formulation

Learned image compression. For a current image x, the DC-AE and ELIC encoders extract generative-space and compression-oriented multiscale features, respectively. The main analysis transform g_a fuses them, and the hyper-analysis transform h_a produces the side-information latent:

$$\begin{aligned} f_{\mathrm{dc}} &= \mathcal{E}_{\mathrm{dc}}(x), & f_{\mathrm{elic}} &= \mathcal{E}_{\mathrm{elic}}(x), \\ y &= g_a(\mathrm{Concat}[f_{\mathrm{dc}}, f_{\mathrm{elic}}]), & z &= h_a(y), \\ \hat{y} &= \mathcal{Q}(y), & \hat{z} &= \mathcal{Q}(z). \end{aligned} \tag{1}$$

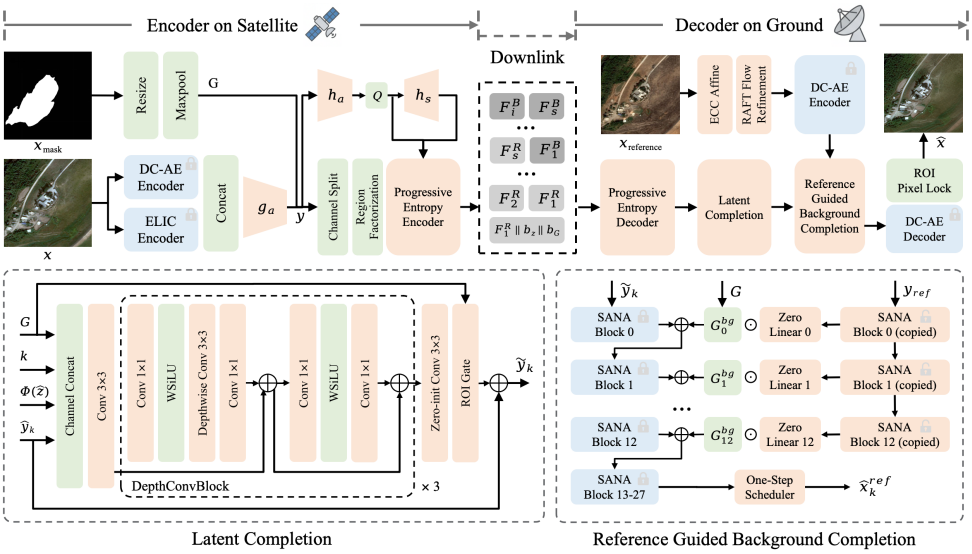

Figure 3: Overview of RPLC with reference-guided decoding. On board, an fMoW-fine-tuned YOLOE-26n model generates $x_{\text{mask}}^{\text{raw}}$, which is mapped to the latent gate G. The compression encoder entropy-codes the ROI payloads b_i^R and background payloads b_i^B independently and serializes their frames once in ROI-first order; the first unit carries $H\|b_z\|b_G\|F_1^R$, and receiving the first k complete frames yields $\hat{y}_k$. On the ground, latent completion takes $(k, G, \hat{y}_k, \Phi(\hat{z}))$, and the aligned x_{ref} provides background-gated ControlNet guidance.

Here $\mathcal{Q}$ denotes quantization. For a 512×512 input, $f_{\text{dc}} \in \mathbb{R}^{32\times16\times16}$ is the generative DC-AE latent and $f_{\text{elic}} \in \mathbb{R}^{320\times32\times32}$. Learned convolutional adapters (not interpolation) match both to 16×16, and after concatenation a stride-2 stage of g_a yields the entropy-coded latent $\hat{y} \in \mathbb{Z}^{320\times8\times8}$, and h_a yields $\hat{z} \in \mathbb{Z}^{160\times2\times2}$. Region partitioning, the gate G, and entropy coding all operate on this 8×8 grid. The hyper-synthesis transform maps $\hat{z}$ to a 320-channel context feature $\Phi(\hat{z}) = h_s(\hat{z})$ on the latent grid. $\hat{z}$ is coded into b_z under a learned channel-wise factorized prior. Each symbol of $\hat{y}$ is coded as the integer residual $s = \text{round}(y-\mu)$ under the discretized Gaussian $\mathcal{N}(\mu, \sigma^2) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2})$ and reconstructed as $\hat{y} = \mu + s$; μ and σ come from the context model of Eq. (4), and σ selects one of 64 CDFs on a log-uniform scale table over $[0.11, 256]$. $\mathcal{R}(\hat{y}, \hat{z})$ sums the negative log-likelihoods of both. At the decoder, the synthesis transform g_s (depthwise-convolution blocks and a learned $2\times$ upsampling) maps the decoded 320×8×8 latent to $\ell = g_s(\tilde{y}) \in \mathbb{R}^{32\times16\times16}$, which a one-step SANA transformer (Xie et al., 2025) denoises and a DC-AE decoder (Chen et al., 2025) maps to $\hat{x}$. The backbone minimizes $\mathcal{L}_{\text{RDP}} = \mathcal{R}(\hat{y}, \hat{z}) + \lambda_D \mathcal{D}(x, \hat{x}) + \lambda_P \mathcal{P}(x, \hat{x})$ (Blau & Michaeli, 2019), where the three terms measure rate, distortion, and perceptual discrepancy, and λ_D, λ_P set their trade-off. Because the downlink budget is unknown at encoding time, RPLC orders the units of $\hat{y}$ such that every received unit-boundary prefix is decodable and short prefixes favor the ROI.

Conditional diffusion generation. ControlNet (Zhang et al., 2023) steers a frozen diffusion denoiser through a trainable copy of its early blocks joined by zero-initialized layers, preserving the pretrained prior at initialization. The one-step SANA decoder can thus use untransmitted side information; periodic revisits supply an archived image of the same site, most useful for the background, whose bits arrive last (Section 3.5).

3.2 Onboard ROI Mask Generation

The satellite generates the binary ROI mask $x_{\text{mask}}^{\text{raw}}$ with YOLOE-26n (Wang et al., 2025; Jocher et al., 2026), fine-tuned on SAM (Kirillov et al., 2023) pseudo-masks prompted by the official boxes of the fMoW training split; no test image or manual mask is used (Appendix B). Acquisitions are task-driven: the ground station uplinks the target category, and the onboard model keeps the highest-scoring instance of that category, falling back to the most confident detection; our evaluation uses the annotated fMoW category, with one target per image. Nearest-neighbor resizing of $x_{\text{mask}}^{\text{raw}}$ to the

8×8 grid of $\hat{y}$ and 3×3 max-pooling, which adds a one-cell boundary margin, yield $G \in \{0,1\}^{8\times8}$, broadcast over the 320 channels. G travels bit-exactly in the 60-byte gate packet b_G of the first unit and is shared by unit formation, latent completion, reference gating, and ROI locking.

3.3 Region-Progressive Latent Coding

We divide the quantized latent $\hat{y}$ into S channel slices and use G to split each slice into ROI and background subsets, yielding the ordered units $(R,1),\dots,(R,S),(B,1),\dots,(B,S)$. Let $s(c) \in \{1,\dots,S\}$ denote the slice containing channel c, and let b_i^R and b_i^B be the range-coded payloads of the two regions in slice i. A prefix of $k \in \{1,\dots,2S\}$ complete units contains $k_R = \min(k,S)$ ROI units and $k_B = \max(0,k-S)$ background units. For spatial position p and channel c, the reception mask is

$$M_k(p,c) = \begin{cases} 1, & G(p)=1,\ s(c) \le k_R, \\ 1, & G(p)=0,\ s(c) \le k_B, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Thus, $M_k(p,c) = 1$ exactly when the unit containing (p,c) belongs to the received prefix. Both endpoints then construct the partial latent as

$$\hat{y}_k = M_k \odot \hat{y} + (1 - M_k) \odot \bar{\mu}_k, \tag{3}$$

where $\odot$ denotes elementwise multiplication and $\bar{\mu}_k$ holds, for every unreceived coordinate, the conditional mean last predicted by the context model below from $\hat{z}$ and the received prefix. The mask $G \odot (1 - M_k)$ therefore selects exactly the unreceived ROI elements handled by latent completion.

To keep ROI-first entropy coding causal, let u_t be the t-th ordered unit with position mask Δ_t. Each unit is coded in four channel-aware checkerboard phases $\Pi_0,\dots,\Pi_3$, which assign each channel quarter its own spatial checkerboard pattern and visit every position–channel pair once. Both endpoints keep a latent-shaped context state $c_{t,\phi}$, initialized as $c_{1,0} = \Phi(\hat{z})$ and carried over between units ($c_{t+1,0} = c_{t,4}$). Phase ϕ predicts $(\mu_{t,\phi}, \sigma_{t,\phi}) = f_\phi(c_{t,\phi})$ with phase-specific adapters around a shared context network, codes the symbols $\Delta_t \cap \Pi_\phi$ with these parameters, and updates the state as

$$c_{t,\phi+1}(p,c) = \begin{cases} \hat{y}(p,c), & (p,c) \in \Pi_\phi \cap \mathcal{R}_{\le t}, \\ \mu_{t,\phi}(p,c), & (p,c) \in \Pi_\phi \setminus \mathcal{R}_{\le t}, \\ c_{t,\phi}(p,c), & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathcal{R}_{\le t}$ collects the positions of units $u_1,\dots,u_t$, and decoded values include a deterministic latent-residual-prediction correction. Untransmitted coordinates thus enter the state only through predicted means, never through their true values, including every background symbol while ROI units are coded, so encoder and decoder compute identical means, scales, and CDF indices for any prefix. Within each phase, symbols are ordered by channel quarter and then in raster order, and the four phases are concatenated into the unit's single rANS payload; both endpoints rebuild this index sequence from shared masks. A single encoding pass produces all $2S$ units, and any lower-rate reconstruction retains a shorter prefix whose frames are byte-identical to those of the complete stream, so transmitter needs no knowledge of the eventual downlink budget.

Bitstream framing and interruption handling. RPLC decodes at unit boundaries. Each payload is wrapped in a frame F_i^r, $i \in \{1,\dots,S\}$ and $r \in \{R,B\}$, behind a 20-byte frame header h_i^r that stores its length and group index, and the first unit U_1 prepends a 20-byte stream header H (which stores the length of b_z), the hyperprior payload, and the fixed-size gate packet:

$$\begin{aligned} U_1 &= H \| b_z \| b_G \| F_1^R, \qquad F_i^r = h_i^r \| b_i^r, \\ \mathcal{T} &= U_1 \| (F_2^R \| \cdots \| F_S^R) \| (F_1^B \| \cdots \| F_S^B). \end{aligned} \tag{5}$$

The decoder thus parses b_z and b_G before b_1^R, whose decoding requires both. The entropy coder restarts at each boundary, so an incomplete frame cannot corrupt earlier units. All of these bytes are counted in the reported rates: the first unit U_1 averages 292 bytes, of which 100 are fixed structures (H, the 60-byte b_G, and one frame header) and 192 carry b_z and b_1^R, and each later unit adds one 20-byte frame header to its payload. If a contact ends mid-stream, an interruption at an arbitrary byte rolls back only to the preceding unit boundary; across 41,216 random and boundary-adjacent

byte cuts of the fMoW test streams, the parser recovered the maximal complete prefix every time, discarding 142 bytes on average (336 at the 95th percentile; Appendix C). We evaluate tail truncation of a contiguous stream; erasures inside a received prefix require retransmission and are outside this setting, and a contact ending before U_1 is complete yields no reconstruction.

3.4 Decoder-Side Latent Completion

For a short prefix, the unreceived ROI elements of $\hat{y}_k$ contain only conditional means. A lightweight network $\mathcal{C}$ takes the prefix level k, the gate G, the partial latent $\hat{y}_k$, and the spatial hyper-synthesis feature $\Phi(\hat{z})$, and predicts a residual restricted to those positions:

$$\tilde{y}_k = \hat{y}_k + G \odot (1 - M_k) \odot \mathcal{C}(k, G, \hat{y}_k, \Phi(\hat{z})), \tag{6}$$

where G is broadcast over channels, so $G \odot (1 - M_k)$ prevents $\mathcal{C}$ from changing the background or any received symbol. $\mathcal{C}$ (Figure 3) embeds and concatenates the four inputs, fuses them with a 3×3 convolution, and applies three residual depthwise-convolution blocks; its zero-initialized output layer starts exactly from conditional-mean filling. With the codec frozen, $\mathcal{C}$ is trained on prefixes $k \in \{1, \ldots, 5\}$ (probabilities $(4, 3, 2, 1, 1)/11$); $\tilde{y}_k$ is decoded to $\hat{x}_k^{\mathrm{cur}}$ and compared with x inside the ROI through ROI-normalized image-space terms,

$$\mathcal{L}_{\mathrm{C}} = 3\,\ell_1^{\mathrm{ROI}} + \ell_{\mathrm{LPIPS}}^{\mathrm{ROI}} + 0.5\,\ell_{\mathrm{DISTS}}^{\mathrm{ROI}} + 0.5\,\ell_{\mathrm{DINO}}^{\mathrm{ROI}}, \tag{7}$$

namely pixel L1, LPIPS and DISTS on the ROI composite, and a DINOv2 token-cosine term. Because $G \odot (1 - M_k)$ confines the residual to unreceived ROI symbols, it is inactive at $k = 5$, and completion improves the ROI without changing the bitstream or its rate.

3.5 Reference-Guided Background Completion

The historical image x_{ref} is first aligned to the reference-free reconstruction $\hat{x}_k^{\mathrm{cur}}$: ECC (Evangelidis & Psarakis, 2008) removes global affine offsets and RAFT (Teed & Deng, 2020) refines local correspondence with residual optical flow. A frozen DC-AE encoder maps the aligned $x_{\mathrm{ref}}^{\mathrm{ali}}$ into the current latent space, and a soft background gate G^{bg} derived from G suppresses reference features inside the ROI.

We adapt ControlNet from a convolutional U-Net to the SANA DiT. The frozen main branch processes $\tilde{y}_k$, while trainable copies of Blocks 0–12 process the reference latent. Their hidden states $\bar{h}_l$ enter the corresponding main blocks through background-gated, zero-initialized linear projections:

$$h_{l+1} = \mathcal{S}_l(h_l) + c_s(k)\, G_l^{\mathrm{bg}} \odot \mathcal{Z}_l(\bar{h}_l), \quad c_s(k) = \mathrm{clip}\Big(\tfrac{2S-k}{S}, 0, 1\Big), \quad l = 0, \ldots, 12, \tag{8}$$

where $\mathcal{S}_l$, G_l^{bg}, and $\mathcal{Z}_l$ denote the l-th frozen SANA block, the background gate bilinearly resized to the token grid, and the zero-initialized linear projection. The prefix-adaptive scale $c_s(k)$ keeps full reference guidance while $k \le S$ and fades it linearly to zero as the S background units arrive. The $\mathcal{Z}_l$ are the token-space counterparts of ControlNet zero convolutions. Only the copied branch and $\mathcal{Z}_l$ are trained; frozen Blocks 13–27, the one-step scheduler, and DC-AE decoder then produce $\hat{x}_k^{\mathrm{ref}}$.

To prevent the historical reference from modifying the transmitted target, we apply ROI pixel locking. With $\hat{x}_k^{\mathrm{cur}}$ and $\hat{x}_k^{\mathrm{ref}}$ denoting the reference-free and reference-guided reconstructions, the final output is

$$\hat{x}_k = m_{\mathrm{rx}} \odot \hat{x}_k^{\mathrm{cur}} + (1 - m_{\mathrm{rx}}) \odot \hat{x}_k^{\mathrm{ref}}. \tag{9}$$

Pixels inside the mask are thus decoded from the current-image bitstream, through latent completion and the generative decoder, and the historical reference contributes only outside it. At the decoder, b_G yields G bit-exactly and the received pixel mask m_{rx} by nearest-cell expansion of its 16×16 local mask within the transmitted box, so locking uses only transmitted information.

4 Experiments

4.1 Experimental Setup

Datasets. We use the multi-temporal fMoW (Christie et al., 2018), whose repeated observations provide historical references and whose annotated boxes and categories define the tasked targets,

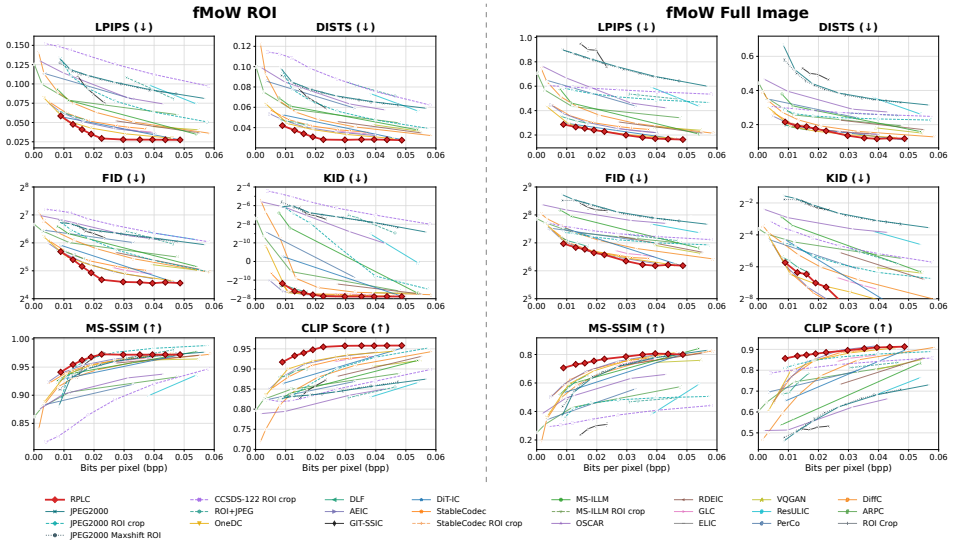

Figure 4: Rate–quality comparison with 24 baselines on fMoW (the two learned ROI crops are single points); left: ROI, right: full image (RPLC with the reference). RPLC's ROI is reference-free.

with the split and preprocessing of COSMIC (Zhang et al., 2024b) and a fixed test set of 256 images randomly sampled from the fMoW test split. Appendix F also evaluates generalization to DIOR (Li et al., 2020).

Metrics. We report MS-SSIM (Wang et al., 2003), KID (Bińkowski et al., 2018), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), FID (Heusel et al., 2017), and CLIP (Radford et al., 2021) cosine similarity on the evaluated region (ROI, background, or full image). Near-zero unbiased KID differences are treated as ties. ROI metrics are computed within the ROI mask transmitted in b_G and restored at the decoder, i.e., the locked region m_{rx} of Eq. (9): each image is cropped to the mask's bounding box and pixels outside the mask are set to zero, identically for every method and operating point. Target parts missed by the onboard detector thus lie outside the evaluated ROI, and their effect appears in the segmentation task, which is scored against manual masks.

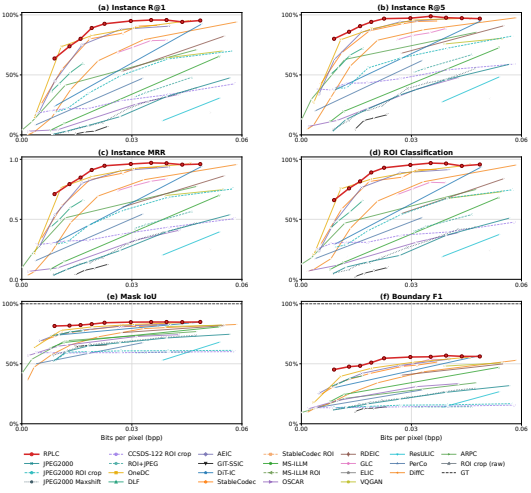

Figure 5: Downstream evaluation on fMoW against 24 baselines (RPLC and ROI crops use the reference): (a–c) instance retrieval, (d) Category Recall@1, (e,f) segmentation vs. manual masks.

ROI classification task. ROI classification (Cheng et al., 2020a) predicts the fMoW category of each target box. Without task-specific training, each target-box crop (minimum side 48 pixels) from the complete reconstruction $\hat{x}_k$ is matched by one-nearest-neighbor search against a gallery holding the uncompressed crop and fMoW category of every test instance. Crops are resized to 224×224 and ranked by cosine similarity of normalized mean-pooled patch tokens from a frozen DINOv2-S/14 (Oquab et al., 2024); the nearest sample's category is the prediction (Category Recall@1). Because the gallery holds each query's uncompressed counterpart, this is a paired-gallery test of how much category and instance evidence a reconstruction retains, not held-out classification; an uncompressed query scores 100% by construction. Exact-instance

Recall@1, Recall@5, and mean reciprocal rank (MRR) measure whether instance-specific evidence survives. DINOv2 also supplies a term of the completion loss (Eq. 7), so this probe shares features with RPLC training. All methods share boxes, crop rule, gallery, and encoder.

ROI segmentation task. To measure target-shape preservation, we manually delineate the target in every test image as pixel-level ground truth. A frozen SAM ViT-H (Kirillov et al., 2023) segments each complete reconstruction $\hat{x}_k$ from the fMoW target box and one positive center point, and we keep the returned mask with the highest model-predicted IoU, without consulting the ground truth. We report Mask IoU and Boundary F1 (two-pixel tolerance). The box-and-point prompt is an oracle input identical for every method, so this task measures shape preservation under a fixed prompt rather than autonomous segmentation.

Baselines and implementation. The 24 baselines form four groups. (i) **Conventional learned compression**: non-progressive learned and generative codecs, namely DiT-IC (Shi et al., 2026), StableCodec (Zhang et al., 2025), MS-ILLM (Muckley et al., 2023), OSCAR (Guo et al., 2025), RDEIC (Li et al., 2025), GLC (Jia et al., 2024), ELIC (He et al., 2022), VQGAN (Esser et al., 2021), ResULIC (Ke et al., 2025), PerCo (Careil et al., 2024), OneDC (Xue et al., 2025b), DLF (Xue et al., 2025a), and AEIC (Zhang et al., 2026a). (ii) **Progressive compression**: DiffC (Theis et al., 2022) and ARPC (Zhang et al., 2026b). (iii) **ROI-oriented compression**: the selective group-transmission path of GIT-SSIC (Feng et al., 2023) and learned ROI crops coded by StableCodec and MS-ILLM (one operating point each). (iv) **Satellite-oriented conventional baselines**: full-frame JPEG2000 (Skodras et al., 2001), JPEG2000-coded ROI crop (Skodras et al., 2001), CCSDS-122-coded ROI crop (Consultative Committee for Space Data Systems, 2017), full-frame JPEG2000 Maxshift ROI coding driven by the transmitted mask (the 60-byte mask packet included in its rate), ROI+JPEG (Wallace, 1991), and an uncompressed RGB ROI crop. ROI-crop variants transmit the target box and composite its reconstruction over the paired historical image; their rates include all header and payload bytes, normalized by the 512×512 image area; the raw crop (6.08 bpp) appears only in the legend. Learned baselines start from official or released checkpoints and are fine-tuned on fMoW (DiffC is training-free; GIT-SSIC is trained on fMoW); Appendix E lists each checkpoint and measured bitrate range. We summarize comparisons by the *rate-matched envelope*, the best baseline value at each RPLC operating rate, where each baseline enters only within its measured bitrate range: bracketed rates are linearly interpolated, other rates are excluded (no extrapolation), and image-quality envelopes also keep a measured point within 0.003 bpp; fewer baselines thus enter at the lowest rates. The *expanded* envelope includes the satellite-oriented baselines. COSMIC and the loss-resilient codec of Sha et al. (2025) lack bitrate support in our 0.0089–0.0486 bpp range and are not compared quantitatively.

Scope of decoder-side information. RPLC and the four ROI-crop configurations use the historical image, so full-image curves compare deployment configurations. ROI image-quality metrics involve no side information for RPLC, because the evaluated ROI coincides with the locked region, which is reference-free by construction. The fMoW classification and segmentation probes run on the complete output $\hat{x}_k$, which can include reference-guided pixels outside m_{rx}, and are thus end-to-end reference-assisted; DIOR uses the reference-free path.

Training details. Three stages are trained on 8 A800 GPUs with AdamW, each freezing earlier modules: (1) the five-slice ROI-first codec, initialized from the merged one-step SANA/DC-AE codec; (2) the 5.33M-parameter completion network; and (3) the Reference ControlNet, on aligned fMoW temporal pairs (schedules and loss weights in Appendix A).

4.2 Main Results

Rate–distortion comparison. Figure 4 reports ROI and full-image results on the fMoW test set. **ROI.** RPLC exceeds the expanded rate-matched envelope on DISTS and CLIP at all ten operating points. Even at the first prefix ($k = 1$, 0.0089 bpp, one of five ROI units), RPLC outperforms every rate-matched satellite-oriented baseline on all six ROI metrics (LPIPS 0.0585, DISTS 0.0420). Against JPEG2000 Maxshift ROI coding with the same transmitted mask and mask packet (0.0086 bpp), RPLC at $k = 1$ lowers ROI LPIPS by 0.068 and DISTS by 0.049 and raises PSNR by 2.98 dB; paired-bootstrap 95% intervals exclude zero on all seven ROI metrics. Further prefixes

and the learned ROI-crop comparison are in Appendix D. **Full image.** For the complete reference-assisted configuration, RPLC leads primarily on LPIPS, FID, MS-SSIM, and CLIP, an advantage that includes the historical reference most baselines lack; background-region and DIOR results are in Appendices D and F.

ROI classification. Figure 5(d) reports top-1 ROI classification accuracy under the paired-gallery protocol (Category Recall@1). At the lowest-rate prefix (0.0089 bpp), RPLC reaches 66.0% and lies above the expanded rate-matched envelope (instance retrieval in Figure 5(a–c)). Because box crops extend beyond m_{rx} (51% box coverage on average), the reference aids this probe: without the reference branch, $k = 1$ Category Recall@1 falls from 65.6% to 48.8% in a controlled rerun. A DINOv2 linear probe trained on a disjoint fMoW split gives 21.1% top-1 accuracy at $k = 1$, versus 0.0% for Maxshift at 0.0086 bpp.

Table 1: Regional leave-one-out ablation on fMoW (anchor: full RPLC). Each module is evaluated where it operates; positive BD-rate, integrated over the common quality support, is the extra rate needed after removing it.

Model variant	**BD-rate ↓**			
	DISTS	**FID**	**CLIP**	**KID**
ROI quality (k = 1–5)				
w/o ROI-first RPLC	+100.9%	+92.7%	+93.1%	+84.6%
w/o Latent Compl.	+13.2%	+6.3%	+12.1%	+14.7%
Background quality (k = 6–10)				
w/o Ref. ControlNet	+33.3%	+33.7%	+19.3%	+36.6%

ROI segmentation. Figure 5(e,f) measures target-shape preservation against manual masks. RPLC reaches 81.48% Mask IoU at 0.0089 bpp and lies above the expanded envelope at all ten Mask IoU points and nine of ten Boundary F1 points.

4.3 Ablation Study

Table 1 shows that each module saves rate where it operates; replacing ROI-first transmission with uniform full-image slice coding (w/o ROI-first RPLC) has the largest effect.

Wrong-reference robustness. With every historical reference replaced by a seeded derangement, background gating and ROI pixel locking keep ROI values identical to the reference-free output in all 60 metric–prefix comparisons, confirming the locking guarantee. For background and full-image metrics over k = 1–5, the protected mismatch still retains 68.9% of the correct-reference gain on average, reflecting the learned prior. An all-zero guide raises $k = 1$ background FID from 97.7 to 223.7 (Appendix G.3), so the gain is not added capacity.

Onboard cost and mask robustness. Only YOLOE-26n (2.7M parameters) and the 272M-parameter encoder run on board, which targets satellites with AI accelerators. Mask perturbations change first-prefix ROI DISTS by at most 0.0014; missed coverage is the larger risk (Appendix B).

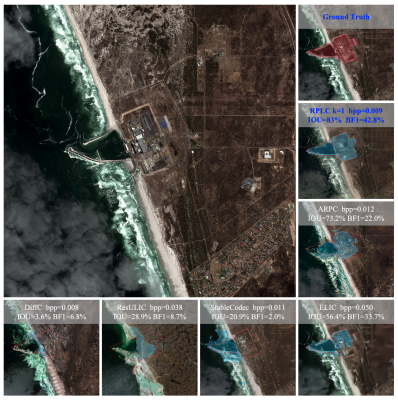

Figure 6: ROI segmentation comparison. Red regions denote the manually annotated ground-truth masks; cyan regions denote the SAM predictions.

5 Conclusion

RPLC encodes each satellite image only once into an ROI-first stream that can be truncated and decoded at every unit boundary, eliminating the need for re-encoding at different rates. When the stream is truncated, RPLC reconstructs the ROI latents beyond the truncation point without transmitting additional bits. It then restores the background from historical observations without affecting the reconstructed ROI.

REPRODUCIBILITY STATEMENT

The appendix documents the three-stage training schedule (optimizer, learning-rate schedules, update counts, and loss weights), the backbone architectures, the physical bitstream format with field-level byte widths, and the source, checkpoint status, and measured bitrate range of every baseline. All operating points are obtained by encoding once and retaining prefixes, and a decode audit of a fixed stream under five random seeds produced identical tensor and image hashes, confirming deterministic ground-side inference.

ETHICS STATEMENT

RPLC is designed to deliver operator-tasked Earth-observation targets over constrained downlinks; it adds no sensing capability beyond the acquiring satellite and uses historical images only where the operator already archives them under its data-governance policy. Background outside the locked ROI is synthesized with help from archived observations and can be outdated or hallucinated, so it should not support operational decisions about current conditions; deployments should mark reference-assisted regions for analysts.

AI USE STATEMENT

We used large language models (LLMs) to polish the prose of the main text. We also used several LLMs to produce simulated reviews of the manuscript and revised the presentation in response. The experimental code was designed and written by the authors; AI tools were then used to check it for bugs and runtime issues. AI tools assisted in plotting the figures from our actual experimental results and in drawing the architecture diagram from the actual code structure. The authors reviewed all AI-assisted output and take full responsibility for the content of this work, including its text, claims, code, and figures.

REFERENCES

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Chunlei Cai, Li Chen, Xiaoyun Zhang, and Zhiyong Gao. End-to-end optimized ROI image compression. *IEEE Transactions on Image Processing*, 29:3442–3457, 2020.

Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *International Conference on Learning Representations*, 2024.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *International Conference on Learning Representations*, 2025.

Gong Cheng, Xingxing Xie, Junwei Han, Lei Guo, and Gui-Song Xia. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:3735–3756, 2020a.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized Gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7948, 2020b.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Consultative Committee for Space Data Systems. Lossless multispectral & hyperspectral image compression. CCSDS Blue Book CCSDS 123.0-B-1, Consultative Committee for Space Data Systems, May 2012.

Consultative Committee for Space Data Systems. Image data compression. CCSDS Blue Book CCSDS 122.0-B-2, Consultative Committee for Space Data Systems, September 2017.

Bradley Denby, Krishna Chintalapudi, Ranveer Chandra, Brandon Lucia, and Shadi Noghabi. Kodan: Addressing the computational bottleneck in space. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pp. 392–403, 2023.

Kiruthika Devaraj, Ryan Kingsbury, Matt Ligon, Joseph Breu, Vivek Vittaldev, Bryan Klofas, Patrick Yeon, and Kyle Colton. Dove high speed downlink system. In *Proceedings of the AIAA/USU Conference on Small Satellites*, 2017.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012.

Thomas Esch, Soner Üreyen, Julian Zeidler, Annekatrin Metz-Marconcini, Andreas Hirner, Hubert Asamer, Markus Tum, Martin Böttcher, S Kuchar, Vaclav Svaton, et al. Exploiting big Earth data from space – first experiences with the TimeScan processing chain. *Big Earth Data*, 2(1):36–55, 2018.

Patrick Esser, Robin Rombach, and Björn 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.

Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008.

Ruoyu Feng, Yixin Gao, Xin Jin, Runsen Feng, and Zhibo Chen. Semantically structured image compression via irregular group-based decoupling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17237–17247, 2023.

Gianluca Giuffrida, Luca Fanucci, Gabriele Meoni, Matej Batič, Léonie Buckley, Aubrey Dunne, Chris Van Dijk, Marco Esposito, John Hefele, Nathan Vercruyssen, et al. The Φ-Sat-1 mission: The first on-board deep neural network demonstrator for satellite Earth observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Pablo Gómez and Gabriele Meoni. Tackling the satellite downlink bottleneck with federated onboard learning of image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6809–6818, 2024.

Jinpei Guo, Yifei Ji, Zheng Chen, Kai Liu, Min Liu, Wang Rao, Wenbo Li, Yong Guo, and Yulun Zhang. OSCAR: One-step diffusion codec across multiple bit-rates. In *Advances in Neural Information Processing Systems*, volume 38, pp. 85267–85286, 2025.

Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14771–14780, 2021.

Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5718–5727, 2022.

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.

Kai Hu, Han Wang, Renhe Liu, Zhilin Li, Shenghui Song, and Yu Liu. ROI-based deep image compression with implicit bit allocation. *arXiv preprint arXiv:2511.08918*, 2025.

Seungmin Jeon, Kwang Pyo Choi, Youngo Park, and Chang-Su Kim. Context-based trit-plane coding for progressive image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14348–14357, 2023.

Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26088–26098, 2024.

Jian Jin, Fanxin Xia, Feng Ding, Xinfeng Zhang, Meiqin Liu, Yao Zhao, Weisi Lin, and Lili Meng. Customizable ROI-based deep image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(1):566–578, 2026.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics YOLO26: Unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

Anle Ke, Xu Zhang, Tong Chen, Ming Lu, Chao Zhou, Jiawen Gu, and Zhan Ma. Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29626–29650, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Jae-Han Lee, Seungmin Jeon, Kwang Pyo Choi, Youngo Park, and Chang-Su Kim. DPICT: Deep progressive image compression using trit-planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16113–16122, 2022a.

Jooyoung Lee, Seyoon Jeong, and Munchurl Kim. Selective compression learning of latent representations for variable-rate image compression. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13146–13157, 2022b.

Binglin Li, Jie Liang, Haisheng Fu, and Jingning Han. ROI-based deep image compression with Swin transformers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Ajmal Mian. RDEIC: Accelerating diffusion-based extreme image compression with relay residual diffusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11):11540–11552, 2025.

Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11913–11924, 2020.

David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Matthew J. Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25426–25443, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

David P Roy, Michael A Wulder, Thomas R Loveland, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Dennis Helder, James R Irons, David M Johnson, Robert Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014.

Hongwei Sha, Muchen Dong, Quanyou Luo, Ming Lu, Hao Chen, and Zhan Ma. Towards loss-resilient image coding for unstable satellite networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12506–12514, 2025.

Junqi Shi, Ming Lu, Xingchen Li, Anle Ke, Ruiqi Zhang, and Zhan Ma. DiT-IC: Aligned diffusion transformer for efficient image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25719–25729, 2026.

Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. The JPEG 2000 still image compression standard. *IEEE Signal Processing Magazine*, 18(5):36–58, 2001.

Myungseo Song, Jinyoung Choi, and Bohyung Han. Variable-rate deep image compression through spatially-adaptive feature transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2380–2389, 2021.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419. Springer, 2020.

Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy compression with Gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022.

George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5306–5314, 2017.

Deepak Vasisht, Jayanth Shenoy, and Ranveer Chandra. L2D2: Low latency distributed downlink for LEO satellites. In *Proceedings of the 2021 ACM SIGCOMM Conference*, pp. 151–164, 2021.

Gregory K Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34 (4):30–44, 1991.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24591–24602, 2025.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, pp. 1398–1402. IEEE, 2003.

Shao Xiang and Qiaokang Liang. Remote sensing image compression based on high-frequency and low-frequency components. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025.

Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. DLF: Extreme image compression with dual-generative latent fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19227–19236, 2025a.

Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. One-step diffusion-based image compression with semantic distillation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 37108–37144, 2025b.

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64971–64995, 2023.

Lei Zhang, Xugang Hu, Tianpeng Pan, and Lili Zhang. Global priors with anchored-stripe attention and multiscale convolution for remote sensing image compression. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:138–149, 2024a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

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.

Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. StableCodec: Taming one-step diffusion for extreme image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17379–17389, 2025.

Tianyu Zhang, Dong Liu, and Chang Wen Chen. Ultra-low bitrate perceptual image compression with shallow encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12118–12128, 2026a.

Weimin Zhang. Two-ray channel models with Doppler effects for LEO satellite communications. *ITU Journal on Future and Evolving Technologies*, 5(2):243–259, 2024.

Ziyuan Zhang, Han Qiu, Maosen Zhang, Jun Liu, Bin Chen, Tianwei Zhang, and Hewu Li. COSMIC: Compress satellite image efficiently via diffusion compensation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 91951–91982, 2024b.

Ziyuan Zhang, Yichong Xia, Bin Chen, Tianwei Zhang, Hao Wang, and Han Qiu. Autoregressive-based progressive coding for ultra-low bitrate image compression. In *International Conference on Learning Representations*, 2026b.

APPENDIX

This appendix documents the implementation of RPLC (Appendix A), the onboard target-mask generator (Appendix B), the physical bitstream and its interruption behavior (Appendix C), additional fMoW results (Appendix D), the provenance of all 24 baselines (Appendix E), cross-dataset results on DIOR (Appendix F), further ablations (Appendix G), and qualitative progressive reconstructions (Appendix H). Unless stated otherwise, fMoW experiments use the 256-image test set randomly sampled from the fMoW test split, ROI definitions, historical-reference pairs, and ten progressive levels of the main text, and every rate counts all transmitted bytes of the corresponding prefix.

A IMPLEMENTATION DETAILS

A.1 DATASETS AND REPORTED CONFIGURATION

The fMoW evaluation follows the preprocessing and temporal-pair construction described in the main text (Christie et al., 2018). The reported RPLC configuration uses ROI-first coding, ROI-only latent completion, and aligned temporal background control. RPLC targets periodic-revisit Earth-observation operations: the ground station archives previously downlinked observations, so a co-located historical image is available as decoder-side information during a later revisit. We additionally use a random 128-image DIOR test set (Li et al., 2020) solely to examine cross-dataset generalization. Because DIOR provides no temporal pairs, this evaluation isolates the reference-free ROI path (ROI-first coding and latent completion). Every method and operating point is evaluated on the same image set. ROI metrics are computed within the ROI mask transmitted in b_G and restored at the decoder, which is also the ROI locking region; this per-image region is shared by all methods.

A.2 TRAINING SCHEDULE AND OBJECTIVES

The five-slice codec is initialized from the merged one-step SANA/DC-AE codec and trained in the three stages summarized in Table A.1; each later stage keeps the previously trained modules frozen. All operating points are obtained by encoding the ten region–channel units once and retaining a prefix; no rate-specific re-encoding is performed.

Table A.1: Audited training schedule of the reported RPLC configuration. Update counts are recovered from training logs rather than nominal checkpoint counters. All stages use AdamW without warm-up on 8 GPUs.

Trainable stage	**Batch/GPU**	**Updates**	**Epochs**	**Learning-rate schedule**
ROI-progressive codec	2	18,000	0.79	5×10^{-5}; $\times0.5$ at 11k, $\times0.2$ at 15.5k
Low-prefix latent completion	4	45,380	3.99	1×10^{-4}; $\times0.5$ at 28k, $\times0.2$ at 40k
Temporal background control	5	7,807	4.76	2×10^{-5}; constant

Objectives. The codec objective combines ROI-composite LPIPS, DISTS, squared-pixel, entropy-rate, adaptive adversarial, and self-distance terms with weights 1, 0.2, 1, 0.5, 0.5, and 0.05; ROI pixels receive twice the pixel-loss weight. Its prefix distribution assigns probability $5/18$ to each of $k = 5$ and $k = 10$ and $1/18$ to every other prefix. Temporal background control uses background L1/LPIPS/DISTS weights $1/1/0.5$ and an ROI L1 weight of 0.5. Latent completion uses ROI-normalized L1/LPIPS/DISTS/DINOv2-cosine weights $3/1/0.5/0.5$ and samples $k = 1, \ldots, 5$ with probabilities $(4, 3, 2, 1, 1)/11$; at $k = 5$ every ROI unit is received, so the completion residual is inactive and these samples provide no completion gradient, although the image-space terms are still evaluated. The DINOv2 encoder of this loss is also the frozen feature extractor of the ROI classification probe.

Data and determinism. Codec and completion training use 363,572 fMoW RGB/mask pairs with official-box rasterized masks, paired random 512×512 crops, and horizontal flipping. The two background stages use 65,613 multi-temporal instances with shared geometric augmentation; the reference is selected by cloud/GSD ranking and aligned with ECC followed by frozen RAFT-small. A

decode audit of one fixed entropy stream, mask packet, reference, and all ground-side modules under seeds 0–4 produced identical tensor and PNG hashes, confirming deterministic inference.

A.3 Encoder and Decoder Backbones

Figure A.1 summarizes the three image transforms. We adopt SANA's `dc-ae-f32c32-sana-1.0` DC-AE (Xie et al., 2025; Chen et al., 2025): its encoder maps a 512×512 RGB image to a $32 \times 16 \times 16$ latent, combining high-resolution ResNet blocks with low-resolution EfficientViT blocks, and its decoder reverses this hierarchy. The ELIC branch (He et al., 2022) serves only as an auxiliary analysis encoder, not as a stand-alone codec; its 5×5 convolutional stages, residual bottlenecks, and attention blocks extract compression-oriented multiscale features ($320 \times 32 \times 32$ for a 512×512 input), which the RPLC analysis transform g_a fuses with the DC-AE features through learned adapters rather than interpolation:

$$
\begin{aligned}
f'_{\mathrm{dc}} &= \mathrm{Conv}_{3\times3}^{32\to128}(f_{\mathrm{dc}}) \in \mathbb{R}^{128\times16\times16}, \qquad f'_{\mathrm{elic}} = \mathrm{Conv}_{3\times3,\, s=2}^{320\to64}(f_{\mathrm{elic}}) \in \mathbb{R}^{64\times16\times16}, \\
y &= \mathrm{DCB}_{320}^{\times2}\Big(\mathrm{Conv}_{3\times3,\, s=2}^{192\to320}\left(\mathrm{DCB}_{192}^{\times2}(\mathrm{Concat}[f'_{\mathrm{dc}}, f'_{\mathrm{elic}}])\right)\Big) \in \mathbb{R}^{320\times8\times8},
\end{aligned}
\tag{A.1}
$$

where DCB_C is a depthwise-convolution block with C channels. For a 512×512 input, the coded latent is thus $\hat{y} \in \mathbb{Z}^{320\times8\times8}$ and the hyper-latent $\hat{z} \in \mathbb{Z}^{160\times2\times2}$; with $S = 5$, each channel slice holds 64 channels. On the decoder side, the synthesis transform g_s maps the completed latent $\tilde{y}_k \in \mathbb{R}^{320\times8\times8}$ to the $32 \times 16 \times 16$ SANA/DC-AE latent $\ell_k = g_s(\tilde{y}_k)$ through three DCB_{320} blocks, a learned $2\times$ upsampling (3×3 convolution, GELU, 1×1 convolution to 1,280 channels, and pixel shuffle), and a final 3×3 convolution to 32 channels; the one-step SANA DiT denoises ℓ_k, and the DC-AE decoder produces the image.

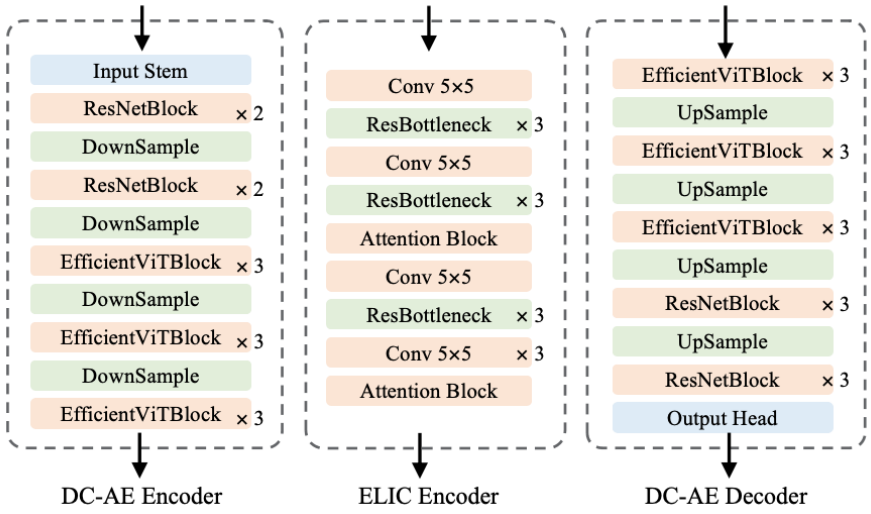

Figure A.1: Stage-level image-transform backbones of RPLC. The DC-AE encoder and decoder provide the 32-channel SANA-compatible latent space at $32\times$ spatial compression, and the ELIC encoder supplies auxiliary multiscale compression features. "$\times n$" denotes n repeated blocks; arrows indicate tensor flow.

A.4 Onboard Resource Profile

Table A.2 lists the satellite-side modules. The satellite runs the fMoW-fine-tuned YOLOE-26n mask generator and the compression path (DC-AE encoder, ELIC auxiliary analysis transform, RPLC analysis and hyperprior transforms, context model, and entropy coder). The Reference ControlNet, one-step SANA denoiser, and DC-AE decoder run only at the ground station.

Table A.2: Satellite-side model and memory profile. YOLOE-26n requires 9.0 GFLOPs at 640×640; peak encoder memory is measured for one 512×512 image.

Onboard component	Parameters	Footprint
YOLOE-26n mask generator	2.7M	10.6 MB (ONNX)
Compression encoder	272M	1.04 GiB (FP32 weights)
Peak encoder memory	–	5.51 GiB

B Onboard Target-Mask Generation and Robustness

B.1 Mask-Entry Protocol

The deployed encoder uses YOLOE-26n, the nano YOLOE open-vocabulary segmentation model (Wang et al., 2025) built on the YOLO26 architecture (Jocher et al., 2026). Because fMoW provides only boxes, training labels are SAM pseudo-masks prompted by the official boxes of the fMoW training split; the masks are cropped and rescaled to the target region, converted to polygon labels, and used to fine-tune the model on 30,055 training and 3,015 validation samples. No test image or manual mask enters training. Acquisitions are task-driven: the ground station uplinks the target category of each tasked image, and the model selects the highest-scoring instance of that category, falling back to the highest-confidence instance from the fixed 62-class vocabulary when the category is not detected. In our evaluation, the annotated fMoW category serves as the tasked category. On the fMoW test set, the target class is present in 252 of 256 images, and the fallback keeps every transmitted mask nonempty. The single deployed mask box obtains 46.79% AP_{50} and 59.77% localization recall at IoU 0.5 against the official box; these are single-mask localization measures, not unrestricted detector AP. On an NVIDIA A800 at batch size one and 640-pixel resolution, warm end-to-end mask generation, including loading, preprocessing, non-maximum suppression, and mask post-processing, takes 22.57 ± 2.38 ms per image (44.3 images/s). Relative to the official fMoW box, the transmitted mask covers 51.4% of the box area on average (median 50.5%; 5th/25th/75th/95th percentiles 17.9%/41.8%/60.2%/94.0%), 97.5% of its pixels lie inside the box, and the mean mask–box IoU is 0.504. The mask therefore stays within the annotated target extent, while the box, a loose outer bound, also contains surrounding context; target-box crops used by the downstream probes consequently include pixels outside the locked region.

B.2 Mask-Perturbation Stress Test

We perturb the predicted mask before both ROI coding and gate-packet (RLM-16) packing in four ways: an 8-pixel erosion and an 8-pixel dilation, each matching the scale of one 8×8-pixel mask cell; random removal of 25% of the positive 8×8-pixel ROI cells (fixed seed 3407); and false-positive background cells whose total area equals 10% of the original ROI. Every condition is re-encoded, packed, decoded, and evaluated at $k \in \{1, 3, 5\}$, yielding 3,072 additional reconstructions. All ROI metrics are measured within the unperturbed transmitted mask, the evaluation region of the main text, so a perturbed mask cannot redefine the region being scored.

Table B.1: First-prefix ($k = 1$) robustness to mask perturbations on the fMoW test set. All conditions, including the unperturbed reference, are re-encoded within this study under the byte accounting of Table C.2; the unperturbed stream operates at the main-text rate of 0.00891 bpp.

Mask condition	bpp	LPIPS↓	DISTS↓	FID↓	MS-SSIM↑	CLIP↑
Unperturbed	.00891	.0585	.0420	51.50	.9410	.9171
Erosion (8 px)	.00834	.0608	.0434	50.41	.9443	.9134
Dilation (8 px)	.00941	.0596	.0429	50.62	.9440	.9153
25% ROI-cell loss	.00855	.0607	.0430	50.78	.9437	.9130
10% false positive	.00952	.0596	.0427	50.20	.9444	.9150

Table B.1 and Figure B.1 show that moderate perturbations shift the protected area and the stream rate and cause a bounded loss in ROI quality. Erosion and 25% ROI-cell loss lower the first-prefix rate by 6.4% and 4.0% and raise ROI LPIPS by 3.9% and 3.8%; dilation and false positives raise

the rate by 5.6% and 6.8% and ROI LPIPS by 1.9%. Every condition changes $k = 1$ ROI DISTS by at most 0.0014. Semantic coverage is more sensitive: erosion reduces the received-mask recall from 76.95% to 67.27%, whereas dilation raises recall to 79.94% but lowers precision from 62.93% to 53.88%. Compression quality thus degrades gracefully under moderate mask-shape errors, while missed target coverage remains the main entry-point risk.

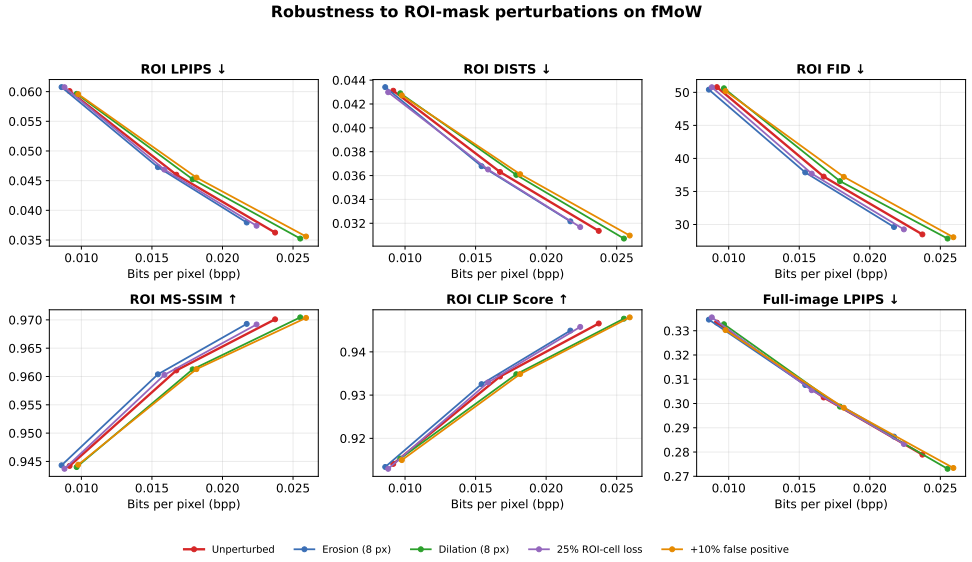

Figure B.1: Rate–quality robustness to target-mask perturbations. Each condition codes and locks with its own perturbed mask, while all ROI metrics use the unperturbed transmitted mask as the evaluation region. All streams are re-encoded within this study; the unperturbed stream operates at 0.00891 bpp for $k = 1$.

C Physical Bitstream and Interruption Handling

C.1 Stream Format

At the transport layer, the implementation serializes the stream as

$$\mathcal{T}_{\text{tx}} = \underbrace{H \| b_z \| b_G \| F_1^R}_{U_1^R} \| F_2^R \| \cdots \| F_S^R \| F_1^B \| \cdots \| F_S^B, \tag{C.1}$$

where H is the global header, $S = 5$, and each frame F_i^r carries an independently framed rANS payload behind a 20-byte frame header. The first unit U_1^R packages H, the hyperprior b_z, the 60-byte RLM-16 gate packet b_G, and F_1^R; it must arrive intact before the first advertised reconstruction is decoded. A frame carries no region flag: the region follows from the ROI-first order and the within-image group index. Within a unit, symbols are serialized phase-major over four channel-aware checkerboard phases ($0 \to 1 \to 2 \to 3$), whose masks split the 320 channels into four quarters with different spatial checkerboard patterns so that each position–channel pair is coded exactly once; within a phase, symbols follow the channel quarter and then a row-major (C-order) spatial scan, and the four phases are concatenated into the unit's single rANS payload. Each phase is coded with the conditional means $\mu_{t,\phi}$ and scales $\sigma_{t,\phi}$ predicted from the context state of Eq. (4), initialized with $\Phi(\hat{z}) = h_s(\hat{z})$ and updated after every phase with decoded values or predicted means, so true values of untransmitted symbols never enter the context; a region–slice delta mask keeps only the symbols of the current unit, and both endpoints rebuild the identical index sequence from the phase, region, slice, and delta masks before rANS coding. At the receiver, the gate G is restored bit by bit from the 8-byte exact gate, and the received pixel locking mask m_{rx} from the 16×16 local mask by deterministic nearest-cell expansion within the transmitted bounding box. Table C.1 lists the field widths. Every packet-restored 8×8 coding gate exactly matches the encoder gate on the fMoW test

set. In the implementation, the codec container stores $H\|b_z\|F_1^R\|\cdots\|F_S^B$, and the transport bundle inserts the RLM-16 packet b_G between b_z and F_1^R, which yields the order above. A truncated frame header or payload invalidates only the current group; an incomplete b_z prevents decoding altogether, and an incomplete b_G withholds the $k = 1$ reconstruction. The codec handles only tail truncation of a contiguous stream; losses inside a received prefix are left to the link layer.

Table C.1: Field widths of the fixed-size stream structures. The group index is a within-image sequence index, and the flags byte is currently zero. The format has no cross-image sequence number.

Structure	**Fields (bytes)**	**Total**
Global header H	Magic 8; version 1; order 1; slice count 1; group count 1; z_h 2; z_w 2; b_z length 4	20
Frame header	Marker 4; group index 1; flags 1; reserved 2; symbol count 4; payload length 4; checksum 4	20/frame
RLM-16 packet b_G	Magic 4; version 1; grid height 1; grid width 1; flags 1; bounding box 4×2; checksum 4; exact gate 8; local mask 32	60

C.2 Byte Accounting

All RPLC rates count every transmitted byte of a prefix, normalized by the 512×512 image area (1 bpp = 32,768 bytes). Table C.2 converts the measured mean rates of the main text into bytes and separates the fixed-size structures from the entropy-coded content. The first advertised prefix ($k = 1$) averages 292 bytes; the header, gate packet, and first frame header account for 100 of them, and the hyperprior b_z with the first ROI payload b_1^R for the remaining 192. Each additional unit adds one 20-byte frame header and its payload.

Table C.2: Byte accounting of the transmitted RPLC stream on the fMoW test set (means over 256 images). Cumulative sizes are the main-text rates expressed in bytes.

	Prefix			
Stream component	$k = 1$	$k = 2$	$k = 3$	$k = 10$
Global header H	20.0	20.0	20.0	20.0
Gate packet b_G	60.0	60.0	60.0	60.0
Frame headers	20.0	40.0	60.0	200.0
Hyperprior b_z and latent payloads	192.0	305.7	384.6	1,312.5
Cumulative prefix (bytes)	292.0	425.7	524.6	1,592.5
Rate (bpp)	.00891	.01299	.01601	.0486

C.3 Random Byte Interruption

Because b_G is a fixed subrecord of the bootstrap unit U_1^R, we first verify, in a separate test over all 60 partial lengths of the 60-byte packet for each of the 256 images (15,360 cases), that every partial gate packet is rejected. The audit uses the stream format of Table C.1, whose first unit $U_1 = H\|b_z\|b_G\|F_1^R$ averages 292 bytes. All cuts below are conditioned on a complete b_G: we uniformly sample 128 tail-cut offsets from each of the 256 image streams (32,768 random cuts; seed 20260731) and add 8,448 deterministic cuts immediately before, at, and after the end of U_1 and each later group boundary. For a cut at byte c, the parser selects the largest $k^*(c)$ whose framed group ends no later than c. Table C.3 shows that all 41,216 cuts recover exactly this maximal complete prefix and that the discarded suffix is confined to the single incomplete group. To test the entropy decoder rather than only the parser, we also decode one interior cut per image: all 256 interrupted buffers are bit-exact with explicit truncation at $k^*(c)$.

The receiver accepts U_1^R only after H, b_z, b_G, and F_1^R have all arrived and passed their checks; before that boundary, no $k = 1$ reconstruction is released. An implementation may form a diagnostic hyperprior-only mean-fill output from a complete $H\|b_z$, but this output is not an advertised operating point. The audit covers tail interruption of a contiguous stream; it does not claim resynchronization after deletion or corruption of bytes inside an already received prefix.

Table C.3: Random tail-interruption audit on the fMoW test streams. Discarded-byte statistics are computed over the uniformly sampled cuts.

Check	Result
Random byte cuts / boundary-adjacent cuts	32,768 / 8,448
Cuts exposing an advertised prefix ($k \geq 1$)	27,446
Partial gate packets rejected	15,360 / 15,360
Maximal complete prefix recovered	41,216 / 41,216
Cuts losing more than one group	0
Discarded bytes (mean / P95 / max)	142.2 / 336 / 606
Interior cuts decoded bit-exactly	256 / 256 (max. latent difference 0)

D Additional fMoW Results

D.1 Low-Prefix ROI Results

Table D.1 reports the first three fMoW prefixes, including the pixel-fidelity metrics PSNR and MS-SSIM. The complete reference-assisted configuration is used, and ROI pixel locking prevents the historical observation from altering the reconstructed target.

Table D.1: Low-prefix ROI performance on the 256-image fMoW test set. Rates count every transmitted byte of the prefix (Table C.2).

k	bpp	PSNR↑	MS-SSIM↑	LPIPS↓	DISTS↓	FID↓	CLIP↑
1	.00891	20.96	.9410	.0585	.0420	51.50	.9171
2	.01299	22.08	.9546	.0476	.0373	42.19	.9328
3	.01601	22.98	.9621	.0409	.0340	35.66	.9398

D.2 Additional ROI Controls and Paired Bootstrap

JPEG2000 Maxshift ROI coding. The Maxshift baseline of the main text encodes the full frame with JPEG2000 Maxshift ROI coding driven by the same transmitted mask as RPLC and add the 60-byte mask packet to each codestream before converting to bpp. Table D.2 lists its ten operating points; its lowest point (0.00864 bpp) lies just below RPLC's first prefix (0.00891 bpp). ROI metrics use the same evaluation region as the main text.

Table D.2: JPEG2000 Maxshift ROI coding on the fMoW test set with RPLC's transmitted mask; rates include the 60-byte mask packet. All metrics are computed on the ROI; Instance R@1 follows the paired-gallery protocol of the main text.

bpp	LPIPS↓	DISTS↓	FID↓	KID↓	MS-SSIM↑	CLIP↑	Inst. R@1↑
.00864	.1270	.0912	111.20	.0203	.9099	.8278	0.39%
.01276	.1165	.0834	101.22	.0141	.9341	.8324	3.13%
.01569	.1130	.0798	93.05	.0104	.9430	.8338	5.86%
.01863	.1097	.0769	88.32	.0087	.9493	.8358	7.81%
.02206	.1060	.0741	85.47	.0079	.9549	.8379	10.55%
.02894	.0997	.0695	79.47	.0063	.9625	.8454	21.88%
.03440	.0955	.0671	75.37	.0052	.9666	.8534	28.13%
.03843	.0920	.0653	72.27	.0045	.9691	.8578	30.86%
.04258	.0897	.0639	69.23	.0039	.9708	.8617	34.38%
.04730	.0868	.0623	66.07	.0030	.9731	.8677	39.06%

Learned ROI-crop baselines. We also code the target crop with two learned codecs, StableCodec and MS-ILLM, including the 23-byte crop header in the rate. Each point is encoded independently and cannot be truncated into intermediate prefixes, and no compatible DiT-IC crop checkpoint was available. Table D.3 compares them with the first two RPLC prefixes: at $k = 2$, RPLC is better on

all five metrics at a 4–6% higher rate, and at $k = 1$, with 27–29% fewer bytes, it is better on all metrics except FID relative to the StableCodec crop.

Table D.3: Learned ROI-crop baselines versus RPLC prefixes on the fMoW ROI.

Method	bpp	LPIPS↓	DISTS↓	FID↓	MS-SSIM↑	CLIP↑
StableCodec ROI crop	.01247	.0743	.0528	46.97	.9305	.9039
MS-ILLM ROI crop	.01220	.0930	.0696	80.80	.9258	.8364
RPLC, $k = 1$	.00891	.0585	.0420	51.50	.9410	.9171
RPLC, $k = 2$	.01299	.0476	.0373	42.19	.9546	.9328

Paired bootstrap. Table D.4 reports 95% intervals for the per-image paired difference between RPLC at $k = 1$ and the lowest Maxshift point, from 1,000 paired image resamples. For FID and KID, the intervals are percentile intervals of the plug-in estimators recomputed on each resample. All seven intervals exclude zero. BD-rate intervals are not reported because the stored baseline curves do not retain per-image values at every operating point.

Table D.4: Paired-bootstrap 95% intervals of the ROI difference, RPLC ($k = 1$, 0.00891 bpp) minus JPEG2000 Maxshift (0.00864 bpp), on the 256 fMoW test images.

Metric	Difference	95% interval
LPIPS	-0.0684	$[-0.0759, -0.0611]$
DISTS	-0.0492	$[-0.0529, -0.0460]$
MS-SSIM	$+0.0311$	$[+0.0282, +0.0341]$
CLIP	$+0.0893$	$[+0.0841, +0.0949]$
PSNR (dB)	$+2.98$	$[+2.70, +3.31]$
FID	-69.42	$[-75.62, -63.70]$
KID	-0.0224	$[-0.0261, -0.0192]$

Held-out linear probe. To complement the paired-gallery protocol, we train a linear classifier on frozen DINOv2-S/14 features of a disjoint fMoW training split and apply it to square crops of the official ROI box enlarged by 15% context; no ground-truth pixels are used. Table D.5 evaluates the ten Maxshift reconstructions of Table D.2 (256 images each) under this protocol: top-1 accuracy stays below 10% over the whole range and is 0.00% at 0.00864 bpp. In the controlled run of Appendix G.3, RPLC reaches 21.09% top-1 (50.39% top-5) at $k = 1$ and 31.25% (61.72%) at $k = 10$ under the same protocol.

Table D.5: Held-out DINOv2-S/14 linear-probe classification of JPEG2000 Maxshift reconstructions on the 256-image fMoW test set. Macro top-1 averages per-class accuracy.

bpp	Top-1↑	Top-5↑	Macro top-1↑
.00864	0.00%	2.73%	0.00%
.01276	1.17%	6.64%	6.52%
.01569	0.78%	6.25%	4.35%
.01863	1.95%	12.50%	5.58%
.02206	3.13%	12.50%	8.62%
.02894	4.69%	16.80%	7.23%
.03440	5.08%	24.22%	7.41%
.03843	7.42%	24.61%	11.88%
.04258	7.81%	26.95%	12.29%
.04730	9.38%	30.08%	10.97%

D.3 Complete Background-Region Results

Figure D.1 reports all ten background-region operating points on fMoW. Prefixes $k = 1$–5 contain ROI units but no background unit, so their background quality is supplied by the aligned historical-reference branch; prefixes $k = 6$–10 then incorporate the five transmitted background units. This

explains why the background curves are meaningful before background payloads arrive and continue to improve after $k = 5$. The figure includes methods with consistent full-frame background reconstructions; ROI-crop baselines, whose uncoded background is composited directly from a historical image, are not treated as codec-only background references.

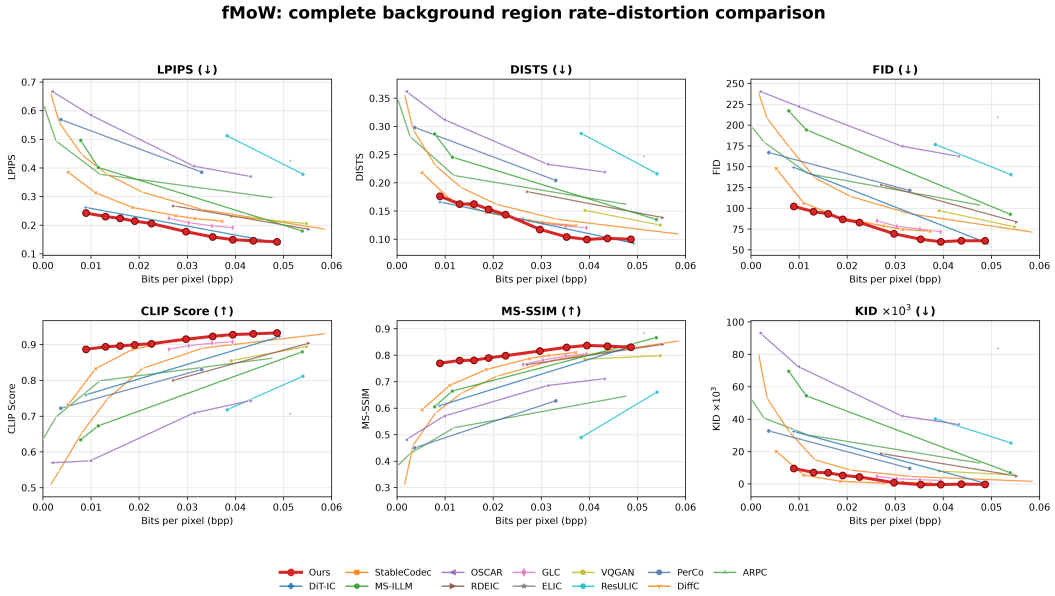

Figure D.1: Background-region rate–quality comparison on the fMoW test set with the aligned reference-guided configuration. Curves connect measured operating points.

E Baseline Sources and Evaluated Rate Ranges

Table E.1 records the implementation and measured fMoW bitrate range of every comparison method, grouped as in the main text. Standards-based and direct-transmission baselines have no learned checkpoint. Rates are measured from the produced streams or persisted files rather than copied from nominal quality labels; the archived machine-readable audit additionally records code revisions, checkpoint paths, training data, inference settings, and every evaluated operating point.

Table E.1: Implementation, fMoW adaptation, and measured fMoW bitrate range of the 24 baselines. ROI-crop rates include crop coordinates, headers, and payload bytes, normalized by the full 512×512 image.

Method	Evaluated implementation	fMoW	Measured bpp
Conventional learned and generative compression			
DiT-IC (Shi et al., 2026)	Official q1/q2 checkpoints	fine-tuned	.008902–.049138
StableCodec (Zhang et al., 2025)	Official target-rate checkpoints	fine-tuned	.005229–.037191
MS-ILLM (Muckley et al., 2023)	Official target-rate checkpoints	fine-tuned	.007825–.053901
OSCAR (Guo et al., 2025)	Official multi-rate checkpoint	fine-tuned	.001900–.043000
RDEIC (Li et al., 2025)	Official released checkpoints	fine-tuned	.027196–.055265
GLC (Jia et al., 2024)	Official released checkpoint	fine-tuned	.026200–.039400
ELIC (He et al., 2022)	Released re-implementation checkpoint	fine-tuned	.051357
VQGAN (Esser et al., 2021)	Released pretrained checkpoints	fine-tuned	.039093–.054718
ResULIC (Ke et al., 2025)	Official released checkpoints	fine-tuned	.038301–.054039
PerCo (Careil et al., 2024)	Official released checkpoints	fine-tuned	.003690–.032990
OneDC (Xue et al., 2025b)	Official released checkpoints	fine-tuned	.003418–.140495
DLF (Xue et al., 2025a)	Official released checkpoints	fine-tuned	.008058–.016718
AEIC (Zhang et al., 2026a)	Official released checkpoints	fine-tuned	.004817–.040249
Progressive compression			
DiffC (Theis et al., 2022)	Official training-free code with Stable Diffusion 1.5	–	.001594–.058602
ARPC (Zhang et al., 2026b)	Official released checkpoints	fine-tuned	.000162–.144767
ROI-oriented compression			
GIT-SSIC (Feng et al., 2023)	Trained from scratch; no test leakage	trained	.014635–.023376
StableCodec ROI crop (Zhang et al., 2025)	StableCodec applied to the target crop; 23-byte crop header counted	–	.01247
MS-ILLM ROI crop (Muckley et al., 2023)	MS-ILLM applied to the target crop; 23-byte crop header counted	–	.01220
Satellite-oriented conventional baselines			
JPEG2000 (Skodras et al., 2001)	OpenJPEG 2.4.0 full-frame streams	–	.008293–.056877
JPEG2000 ROI crop (Skodras et al., 2001)	OpenJPEG 2.4.0 ROI-crop streams	–	.008904–.057531
CCSDS-122 ROI crop (Consultative Committee for Space Data Systems, 2017)	Standards-based local implementation	–	.003998–.058014
JPEG2000 Maxshift ROI (Skodras et al., 2001)	Full-frame Maxshift streams with the transmitted mask; 60-byte mask packet counted	–	.008642–.047299
ROI+JPEG (Wallace, 1991)	Local JPEG ROI-crop implementation	–	.030830–.128972
ROI crop (raw)	Uncompressed RGB crop transmission	–	6.075669

F Cross-Dataset ROI Generalization on DIOR-128

F.1 Scope and Protocol

RPLC addresses periodic revisits, in which earlier observations of the same location are retained at the ground station. DIOR is a single-temporal benchmark and therefore cannot exercise the reference-guided background path; we use it solely to test whether the ROI-first representation and latent-completion network generalize beyond fMoW. Accordingly, we report ROI reconstruction, recognition, and segmentation, and omit background and full-image metrics, which would evaluate an incomplete reference-free configuration instead of the deployed system.

We evaluate this reference-free ROI path and all 24 fMoW baselines on a random 128-image DIOR test set. Unlike the fMoW downstream probes, which read the complete reference-assisted output, DIOR recognition and segmentation use the reference-free reconstruction $\hat{x}_k^{\mathrm{cur}}$. DIOR ROI masks

are produced by the same YOLOE-26n model, and ROI metrics follow the main-text protocol (crop to the mask's bounding box, pixels outside the mask set to zero). Following the main-text protocol, ROI category recognition uses frozen DINOv2-S/14 features and a paired gallery of the uncompressed ROI crops with their DIOR labels. For segmentation, we manually delineate the target in every test image; a frozen SAM ViT-H model segments each reconstruction from the same box and center-point prompt, and its prediction is compared with the annotation by Mask IoU and Boundary F1. The annotations are used only for evaluation. ROI FID and KID use 2,048-dimensional Inception features over the same 128 images, and rate-matched comparisons interpolate only within each baseline's measured bitrate range, never extrapolating.

F.2 RESULTS

Table F.1 lists all ten RPLC operating points. From $k = 1$ to $k = 10$, RPLC reduces ROI LPIPS from .0786 to .0309 and FID from 69.91 to 26.66, while CLIP similarity increases from .9086 to .9583 (Figure F.1). Target representations and boundaries also recover progressively (Figure F.2): at $k = 1$, RPLC obtains 76.56% Instance Recall@1, 84.04% Mask IoU, and 54.03% Boundary F1, and at $k = 3$, Instance Recall@1 and Mask IoU reach 99.22% and 89.13%. These results support cross-dataset generalization of the target-first current-image path.

Table F.1: Reference-free RPLC results on the random DIOR-128 test set. All metrics are evaluated on the ROI.

k	**bpp**	**LPIPS↓**	**FID↓**	**CLIP↑**	**Inst. R@1↑**	**Mask IoU↑**	**Bound. F1↑**
1	.00944	.0786	69.91	.9086	76.56%	84.04%	54.03%
2	.01378	.0587	52.22	.9297	89.84%	87.29%	63.79%
3	.01697	.0485	42.36	.9450	99.22%	89.13%	69.42%
4	.02020	.0399	33.67	.9519	98.44%	89.61%	71.16%
5	.02395	.0326	28.44	.9567	98.44%	90.23%	74.71%
6	.02994	.0314	27.93	.9585	98.44%	91.93%	76.93%
7	.03460	.0312	27.58	.9595	99.22%	93.00%	77.18%
8	.03808	.0310	27.12	.9587	100.00%	92.98%	78.32%
9	.04163	.0310	26.97	.9587	99.22%	92.04%	77.66%
10	.04572	.0309	26.66	.9583	99.22%	92.06%	77.53%

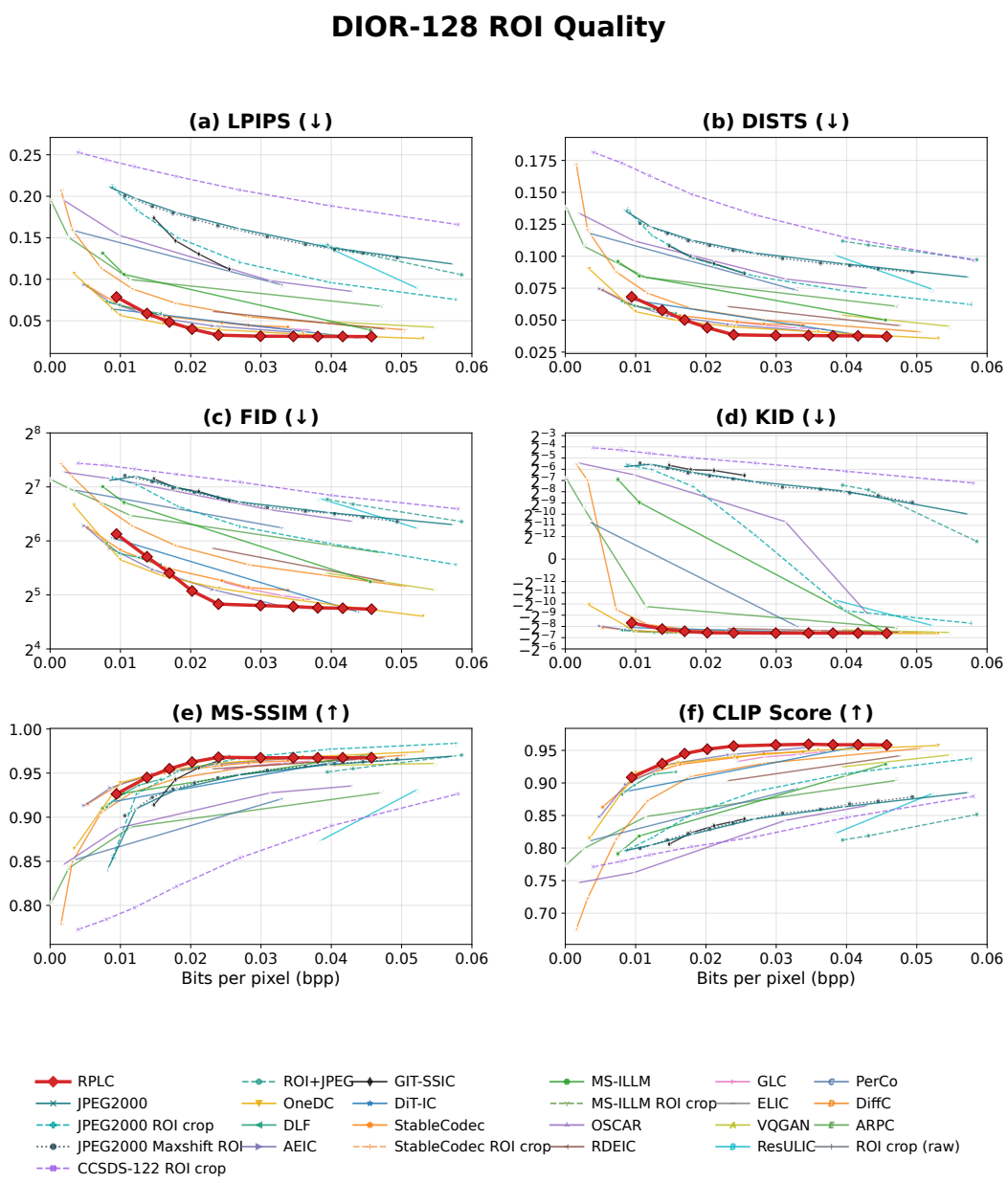

Figure F.1: ROI rate–quality comparison on the random DIOR-128 test set against 24 baselines; the two learned ROI crops are single points, and the uncompressed ROI crop (raw) lies off-axis. FID uses a base-2 logarithmic scale; unbiased KID uses a signed symmetric base-2 scale. Curves connect measured operating points.

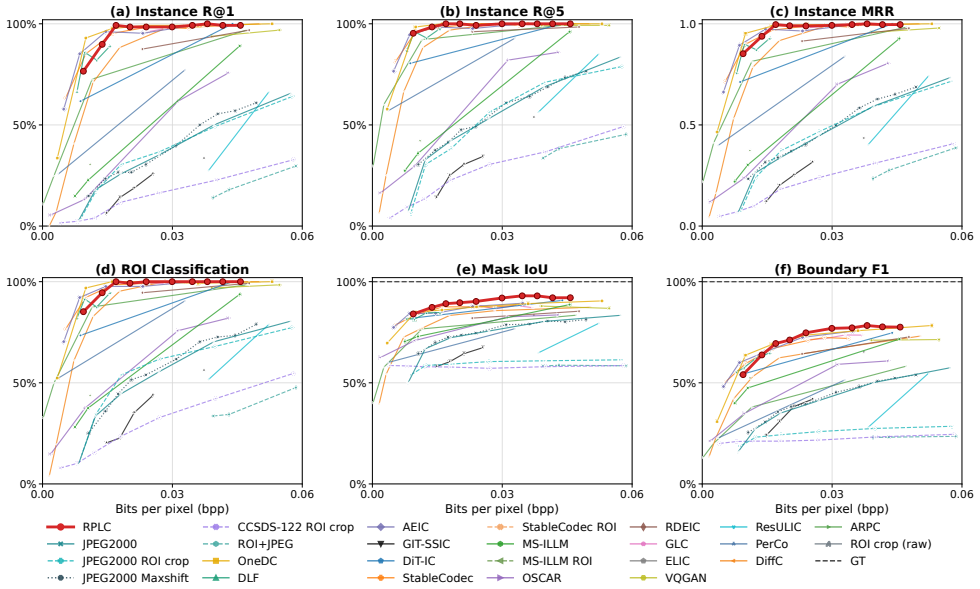

Figure F.2: ROI downstream evaluation on the random DIOR-128 test set against 24 baselines. Recognition uses frozen DINOv2-S/14 features and a paired 128-image gallery; Mask IoU and Boundary F1 compare the frozen SAM prediction on each reconstruction with the manually annotated ROI mask.

G Additional Ablations

G.1 Transmission Order

To isolate ordering from representation learning, we keep the same trained networks, quantized latent, and region–channel units, and re-run entropy coding under each serialization; because the entropy contexts follow the transmission order, the payloads are re-encoded, but no network is retrained. ROI-first sends $(R, 1), \ldots, (R, S)$ before all background units; background-first reverses the region priority; interleaved alternates ROI and background units within each channel slice; and channel-progressive transmits each complete spatial slice before the next. Figure G.1 shows that ROI-first gives the strongest early-prefix ROI quality on all six metrics, and all schedules reach the same endpoint once every unit has arrived, confirming that the early gain comes from transmission priority.

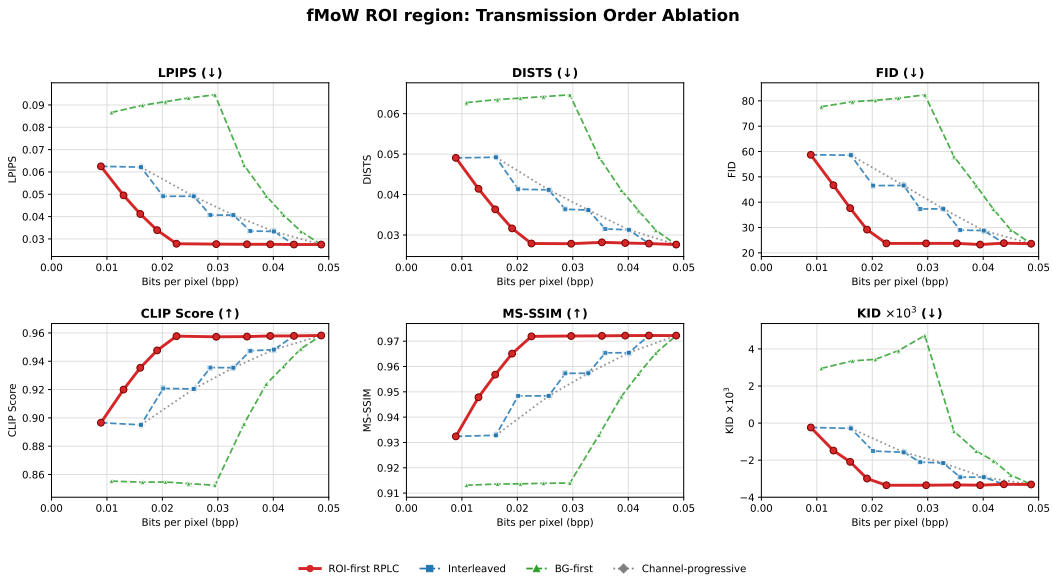

Figure G.1: ROI transmission-order ablation on fMoW. The codec and region–channel units are shared across schedules; only their serialized order changes.

G.2 Reference-Branch Design

Figure G.2 evaluates three design choices of the ground-side reference path. Replacing ECC-plus-RAFT alignment with ECC alone lowers background PSNR by about 0.3 dB at the first prefix, showing that local flow refinement matters most when little current background has arrived. Replacing the prefix-adaptive conditioning scale $c_s(k) = \mathrm{clip}((2S - k)/S, 0, 1)$ of the main text with a fixed value raises background DISTS from 0.096 to 0.120 at $k = 10$. Applying latent completion to the fully mean-filled background also degrades several intermediate prefixes, which motivates the ROI-only completion mask.

G.3 Same-Capacity Zero-Reference Control

To separate reference information from added decoder capacity, we compare three variants over all ten prefixes on the 256 fMoW test images: the trained reference branch with the aligned historical image (*reference*); the same branch fed an all-zero guide (*zero reference*); and the reference branch bypassed (*no reference*). This controlled run uses ECC alignment only, because RAFT weights were unavailable on its host, and its staged stream gives 0.00952 bpp at $k = 1$; it is therefore a controlled comparison rather than a reproduction of the deployed ECC+RAFT numbers. All variants keep ROI pixel locking, so their ROI metrics coincide (e.g., ROI LPIPS .0585 and DISTS .0420 at $k = 1$).

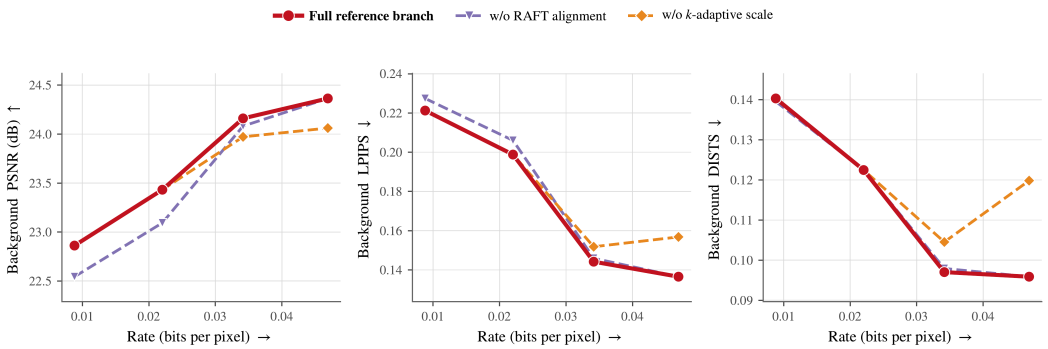

Figure G.2: Leave-one-out ablation of the reference branch. Optical-flow refinement mainly affects low-rate reconstruction, whereas fixed conditioning degrades the high-rate DISTS points.

Table G.1 shows that the zero guide severely degrades the background at $k = 1$, so the reference gain stems from the reference content rather than from the extra branch. At $k = 10$, where $c_s(k) = 0$, the variants coincide. Table G.2 shows that the reference also affects the downstream probes, because target-box crops extend beyond the locked region: at $k = 1$, paired-gallery Category Recall@1 is 65.63% with the reference, 48.83% without the branch, and 36.33% with the zero guide.

Table G.1: Background and full-image quality of the zero-reference control on fMoW.

			Background					Full
Variant	k	**bpp**	**LPIPS↓**	**DISTS↓**	**FID↓**	**MS-SSIM↑**	**CLIP↑**	**FID↓**
Reference	1	.00952	.2904	.1627	97.70	.7500	.8860	119.31
Zero reference	1	.00952	.4559	.3209	223.68	.5649	.6955	260.58
Reference	5	.02316	.2657	.1453	85.75	.7770	.8959	97.47
Ref. / zero ref.	10	.04926	.1420	.1003	60.67	.8303	.9331	73.07

Table G.2: Downstream probes of the zero-reference control on fMoW. Linear-probe accuracy uses a DINOv2-S/14 classifier trained on a disjoint fMoW split; recall values use the paired gallery of the main text.

		Linear probe		Category	Instance	
Variant	k	**Top-1**	**Top-5**	**R@1**	**R@1**	**R@5**
Reference	1	21.09%	50.39%	65.63%	62.89%	76.56%
No reference	1	15.23%	42.97%	48.83%	33.98%	53.52%
Zero reference	1	9.77%	29.69%	36.33%	16.02%	33.98%
Reference	5	25.39%	62.89%	89.45%	87.11%	90.63%
Ref. / zero ref.	10	31.25%	61.72%	94.53%	94.14%	96.48%

G.4 Mismatched-Reference Robustness

We replace every aligned historical image with a seeded random derangement of the same 256-image reference pool, so that no image receives its correct pair. With background gating and ROI pixel locking enabled, the reference-free, correct-reference, and protected mismatched-reference variants have identical ROI values in all 60 metric–prefix checks (ten prefixes, six metrics; Figure G.3); without these protections, a mismatched reference changes the target region. Across background and full-image metrics for $k = 1$–5, the protected mismatch retains 68.9% of the correct-reference gain on average. Because the evaluated ROI is the locked region, this identity holds by construction; the check verifies that the implementation realizes the locking guarantee, and it does not imply that an unrelated reference suffices for high-quality background reconstruction.

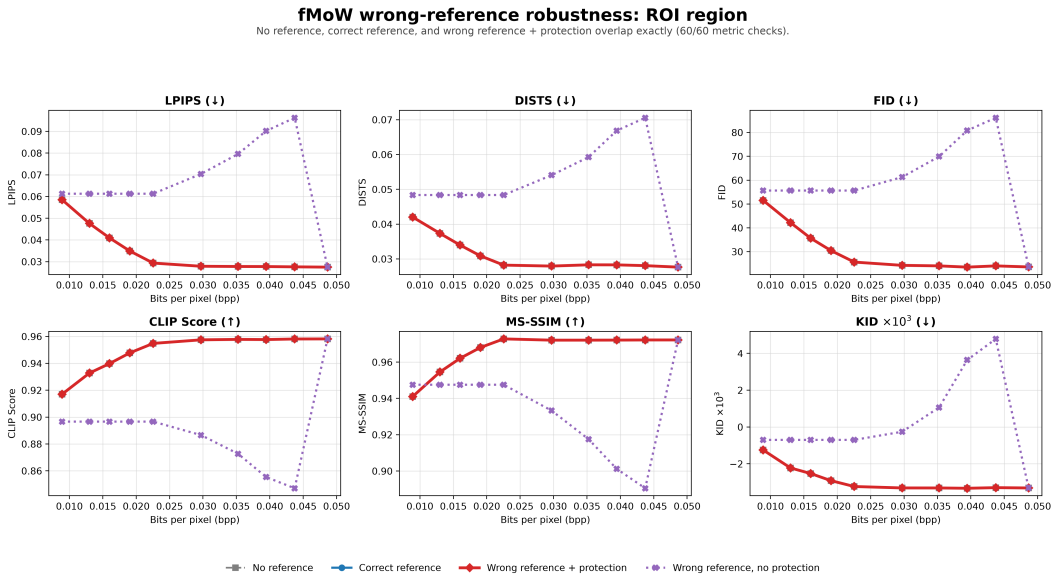

Figure G.3: Mismatched-reference stress test on the fMoW ROI. Reference-free, correct-reference, and protected mismatched-reference curves coincide in all 10×6 metric–prefix checks; without background gating and ROI pixel locking, a mismatched reference changes the ROI reconstruction.

H QUALITATIVE EXAMPLES

fMoW progressive reconstructions (Figures H.1–H.3). These figures show 24 fMoW examples whose decoder-restored masks cover 1.5%–51.8% of the image. The cases span four diagnostic strata, derived reproducibly from image, reference, and mask statistics rather than assigned by hand: small ROIs with sparse backgrounds, temporal appearance changes, registration-sensitive pairs, and fragmented received masks. Each row shows the current image, the co-located historical reference, the decoder-restored mask, and RPLC reconstructions at $k \in \{1, 3, 5, 6, 10\}$. Every displayed rate is measured per image from the complete transmitted prefix.

ROI and background details (Figures H.4–H.9). These figures enlarge predefined crops for 16 representative cases. In each full view, the red box selects the *ROI zoom* crop and the cyan box a separate *background zoom* crop for comparing texture and structure away from the target; the boxes are visualization aids and are never transmitted. The three cases with mask coverage below 5% (03497: 3.45%, 01533: 4.35%, 03307: 4.47%) are grouped at the end of the last two detail figures for comparison; this ordering does not alter the quantitative test set.

DIOR reference-free reconstructions (Figure H.10). Eight DIOR cases at $k = 3, 4, 5, 6, 10$ show raw decoder outputs without ground-truth background compositing or historical-reference guidance. Background quality begins to recover when the first background unit arrives at $k = 6$ and continues to improve toward $k = 10$.

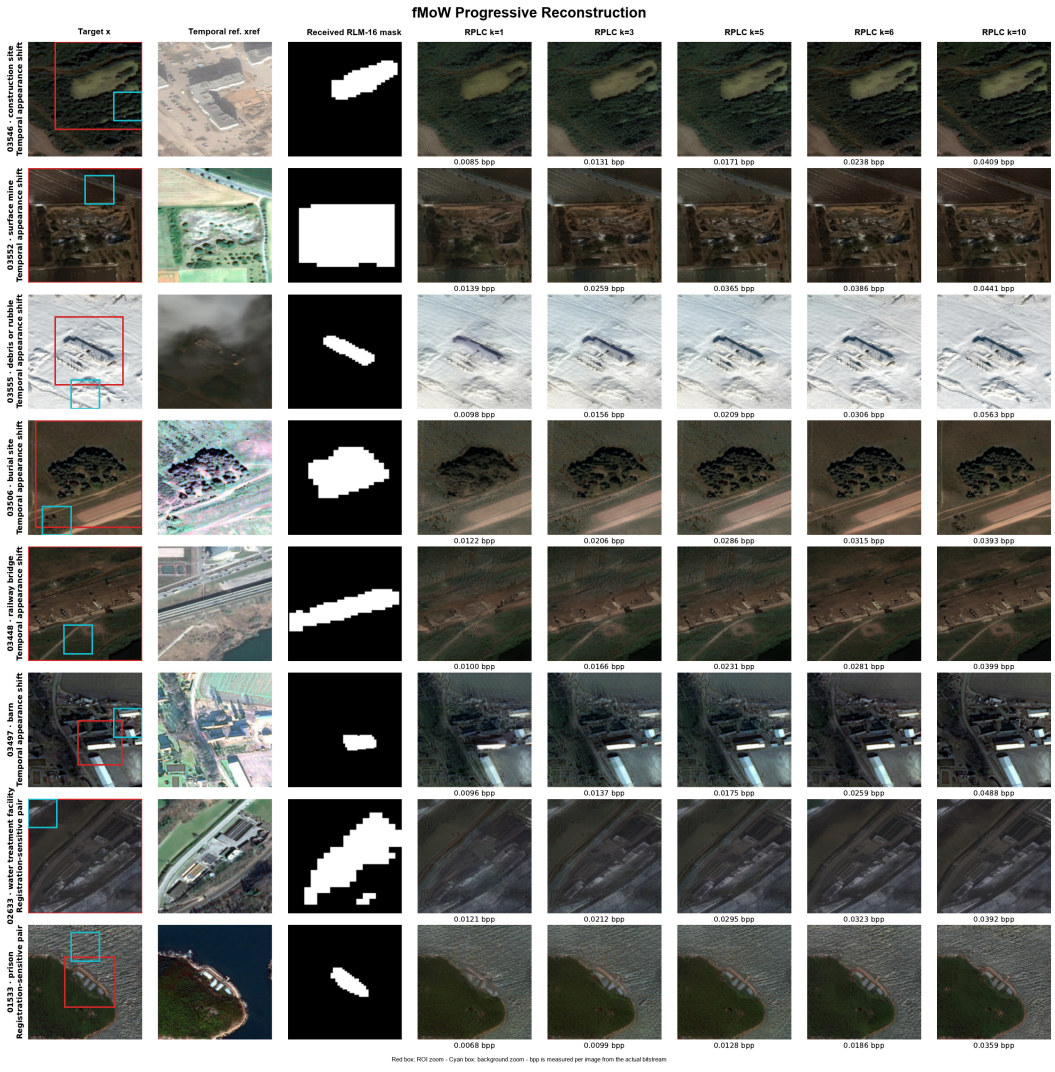

Figure H.1: fMoW progressive reconstructions (1/3): current image, historical reference, decoder-restored mask, and RPLC at $k \in \{1, 3, 5, 6, 10\}$ with per-image rates.

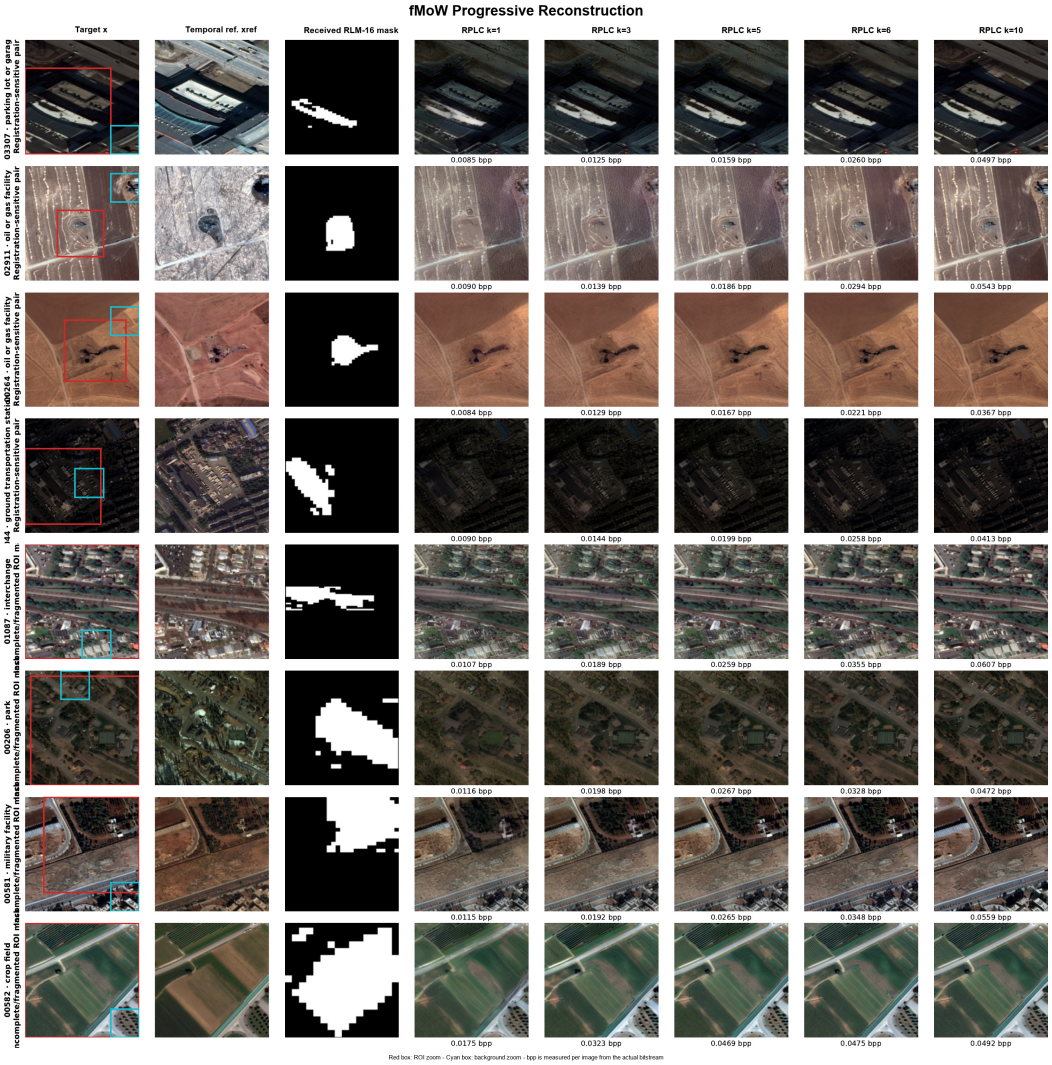

Figure H.2: fMoW progressive reconstructions (2/3).



Figure H.3: fMoW progressive reconstructions (3/3).

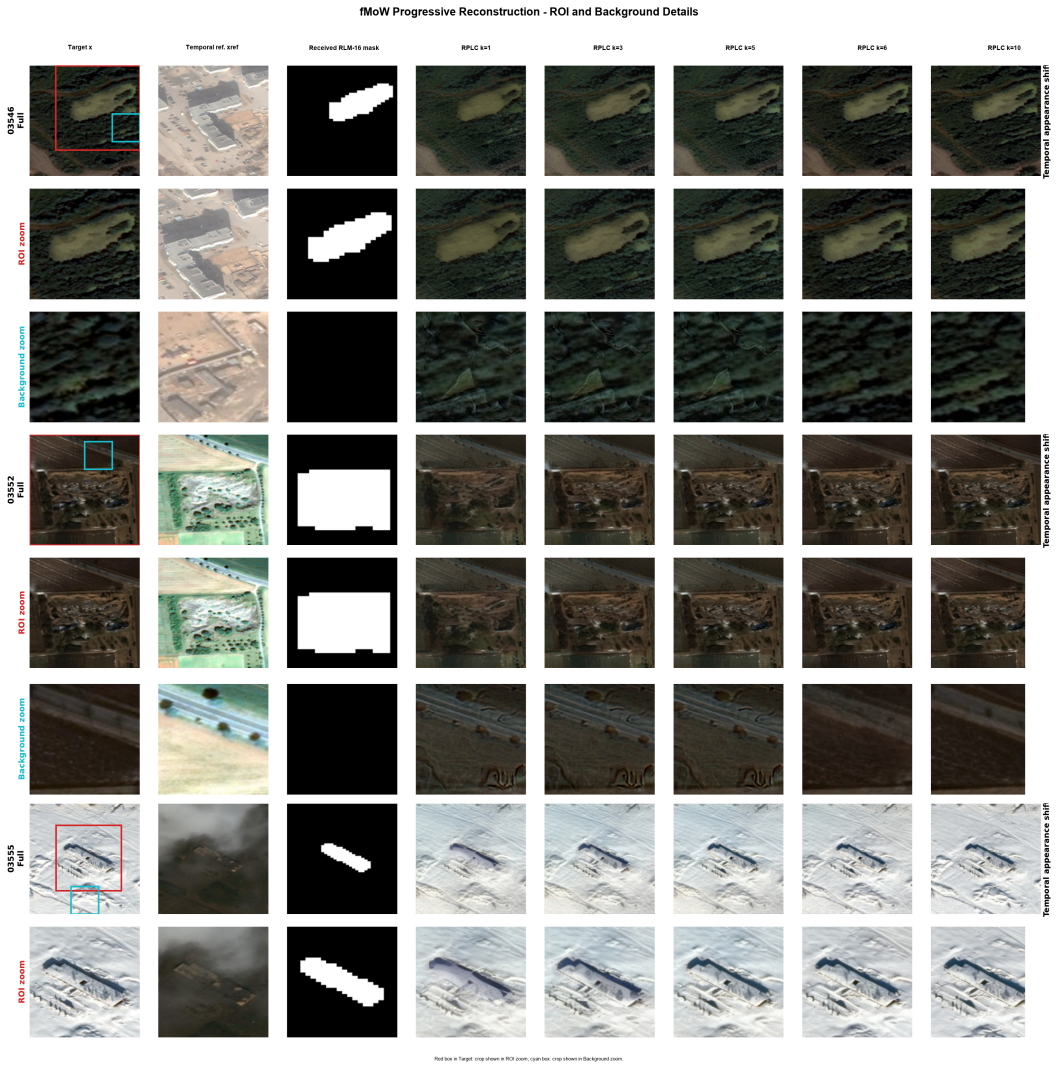

Figure H.4: ROI and background details (1/6). Red box: ROI zoom crop; cyan box: background zoom crop.



Figure H.5: ROI and background details (2/6).



Figure H.6: ROI and background details (3/6).

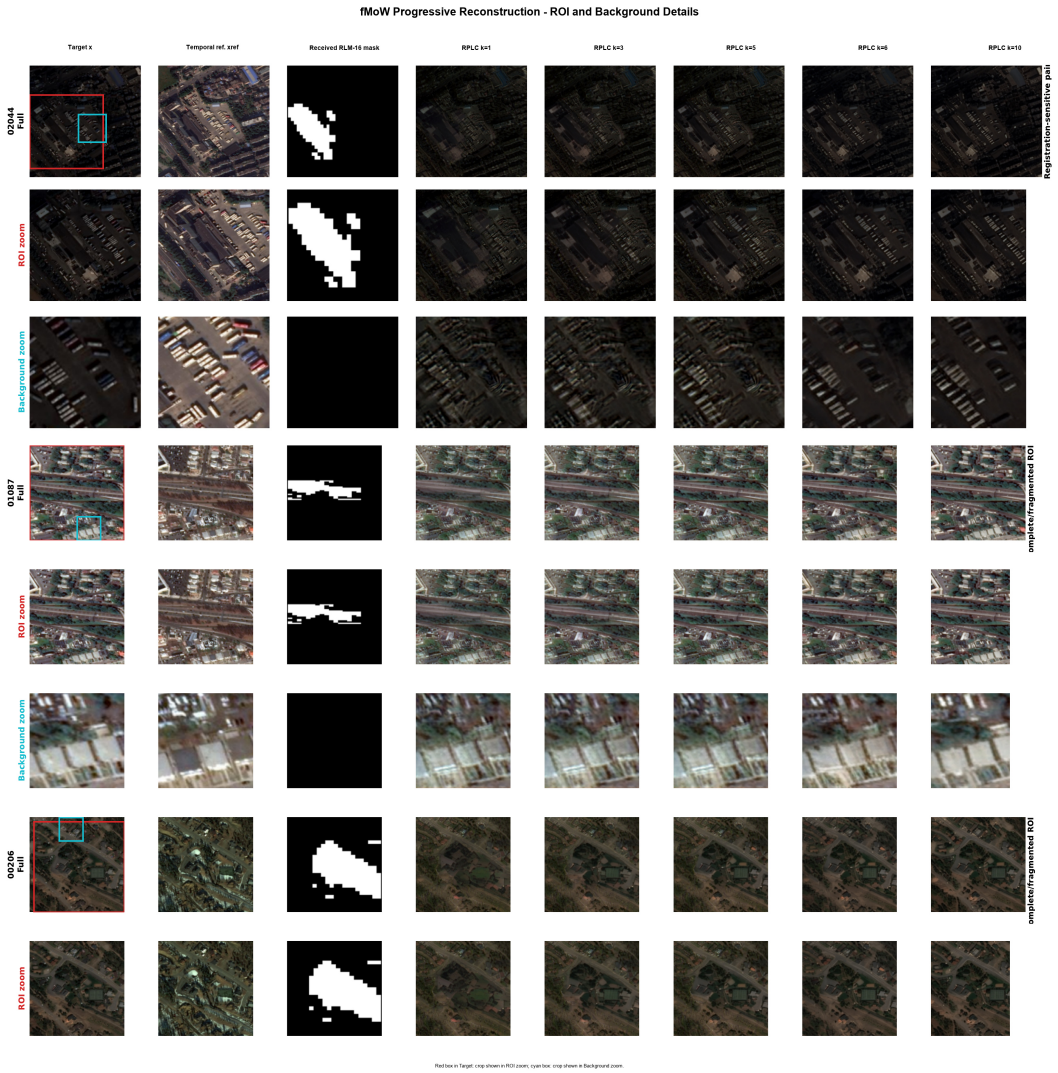

Figure H.7: ROI and background details (4/6).

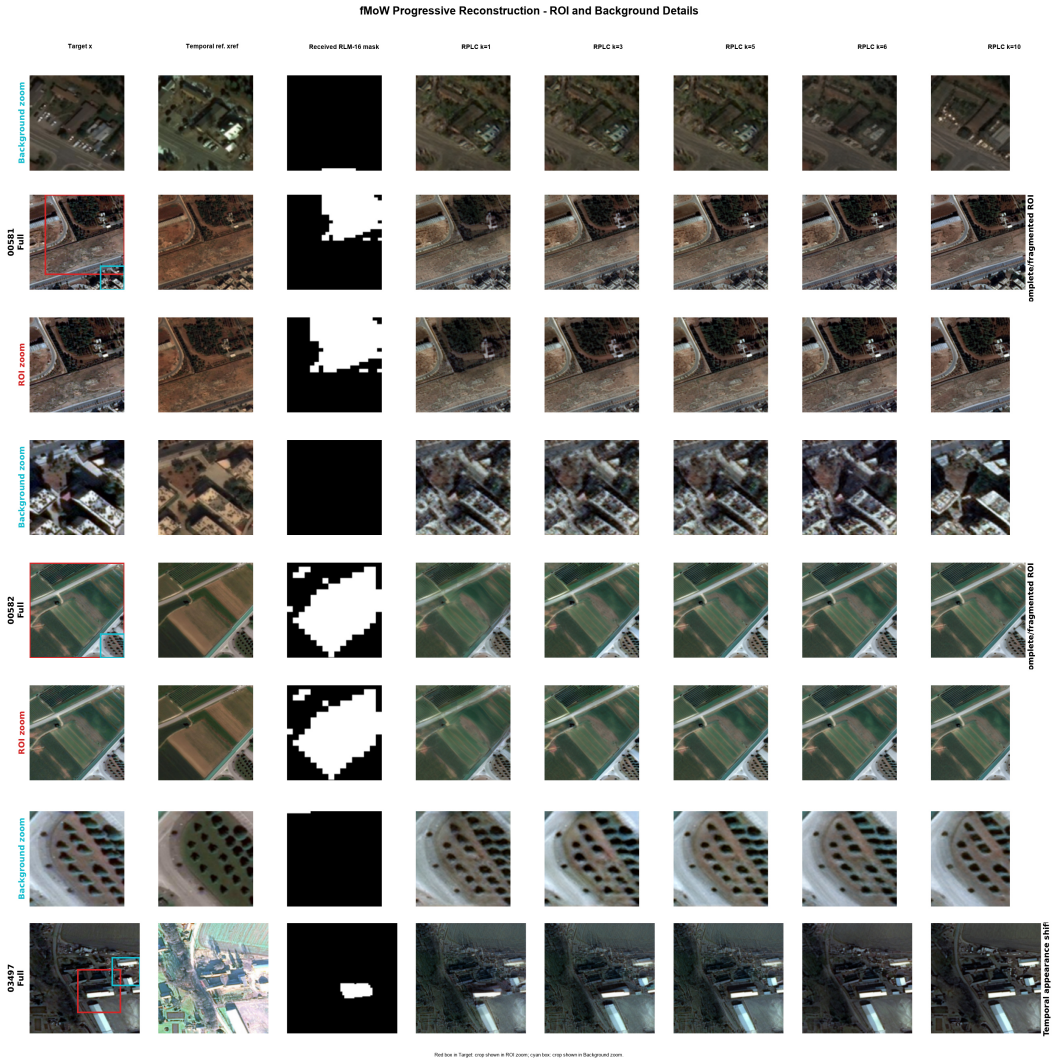

Figure H.8: ROI and background details (5/6), including low-coverage masks.

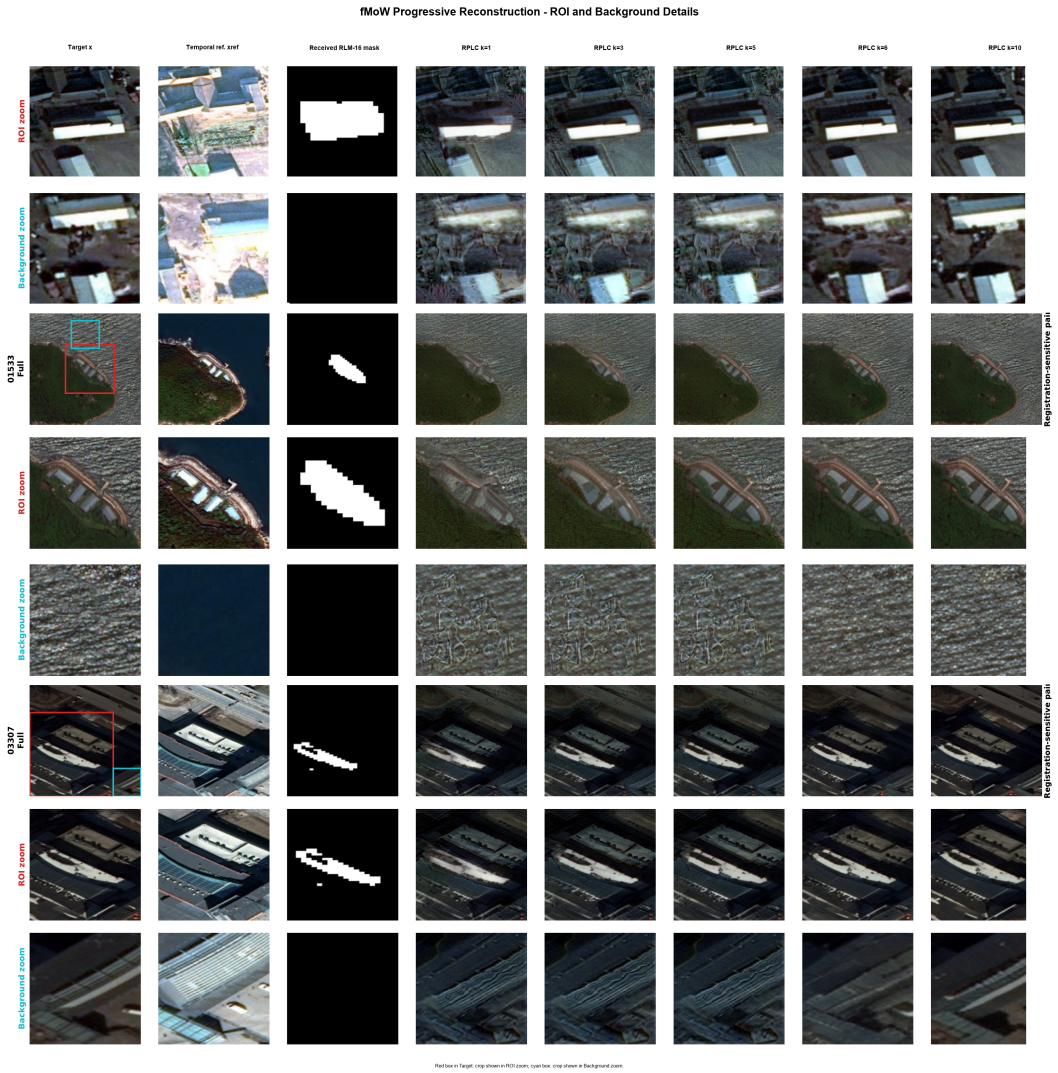

Figure H.9: ROI and background details (6/6), including low-coverage masks.

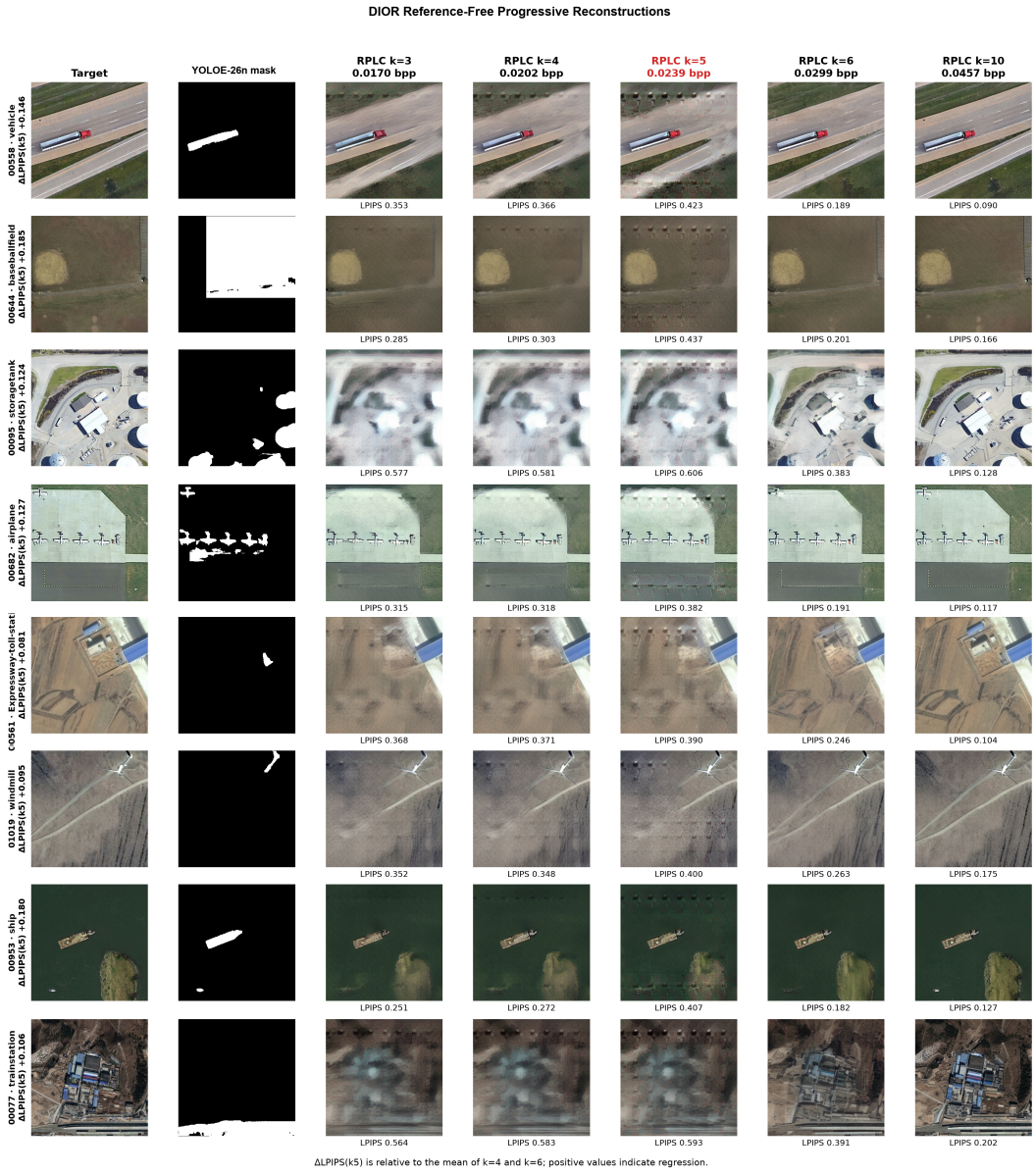

Figure H.10: DIOR reference-free progressive reconstructions at $k = 3, 4, 5, 6, 10$.