

# PALR-DNA: A Probability-Aware Latent Recovery Framework for Error-Resilient Image Storage in DNA

**Siqiang Chen**

Southeast University

**Ruijing Zhao**

Southeast University

**Tianyu Wu**

Southeast University

**Ting Huang**

Southeast University

**Xiaozhi Huang**

Southeast University

**Qingjiang Sun**

[sunqj@seu.edu.cn](mailto:sunqj@seu.edu.cn)

Southeast University

---

## Article

**Keywords:** DNA data storage, Image reconstruction, Latent representation, Probabilistic decoding, Transformer, Error-resilient decoding

**Posted Date:** October 27th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11201997/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

In DNA-based image storage, base errors inevitably degrade image reconstruction quality. Existing deep-learning methods are effective in correcting base errors while often overlooking uncertainty encoded in low-probability predictions, thereby limiting reconstruction quality. To address this issue, PALR-DNA, a probability-aware latent recovery framework for error-resilient image storage in DNA, is proposed in this study. It represents images as spatially indexed latent vectors and exploits byte-level probability distributions for recovery from noisy DNA sequences. Specifically, a DNA-to-latent recovery transformer is developed to map variable length noisy sequences directly to byte probability distributions. Top-K candidate beam search is employed to explore high-probability candidates and select sequences with valid spatial indices passing cyclic redundancy checks. Byte-wise maximum-probability estimation approximates unresolved latent vectors from the remaining probability information. Reconstruction results for 543 images demonstrate that PALR-DNA outperforms the baseline methods across 1–5% error rates, with larger gains at higher rates. At 5%, it achieves a latent-vector recovery accuracy of 93.61%, peak signal-to-noise ratio of 27.45 dB, and multiscale structural similarity index measure of 0.918, while remaining stable across error compositions. Overall, PALR-DNA provides a new approach to error-resilient image reconstruction in DNA storage by retaining byte-level probability information.

## 1. Introduction

The rapid expansion of the global datasphere is placing increasing pressure on traditional storage media, and the annual volume of data generated worldwide was projected to reach approximately 175 zettabytes in 2025<sup>[1–3]</sup>. The continued growth of social media and other digital media services has accelerated the production and consumption of digital images, leading to the growing demand for scalable storage systems capable of preserving large image collections efficiently<sup>[4]</sup>. DNA, featuring high information density, molecular stability under suitable preservation conditions, and minimal energy requirements during storage<sup>[5–9]</sup>, can preserve large amounts of image data in a small physical space for extended periods without continuous power, making it a promising archival medium. DNA data storage can be modeled as a noisy communication channel, in which substitutions, insertions, and deletions are introduced across oligonucleotide synthesis, polymerase chain reaction (PCR) amplification, molecular storage, and sequencing<sup>[10–12]</sup>. Substitutions replace intended nucleotides, whereas insertions and deletions alter sequence length and disrupt symbol synchronization during sequence alignment and decoding<sup>[13]</sup>. In image-oriented DNA storage, errors in strand addresses or encoded image data can trigger localized discoloration, severe visual degradation, or reconstruction failure<sup>[14, 15]</sup>. Collectively, these error processes compromise the integrity and retrievability of DNA-encoded information, thereby limiting its reliability in practical storage systems<sup>[16–18]</sup>.

Several methods have been proposed to improve the resilience of DNA-based image storage to nucleotide errors. These methods generally fall into two main categories: DNA sequence error correction<sup>[19–21]</sup> and image restoration<sup>[22–24]</sup>. Press et al.<sup>[19]</sup> introduced HEDGES using hash-based encoding and greedy exhaustive-search decoding to address substitution, insertion, and deletion errors. Welzel et al.<sup>[20]</sup> designed DNA-Aeon using arithmetic coding and a Raptor fountain code to correct substitutions, insertions, deletions, and strand loss while satisfying biochemical constraints. Yan et al.<sup>[21]</sup> presented DNA StairLoop with staircase interleaving and iterative soft-decision decoding for reliable recovery at high error and strand-dropout rates under low sequencing coverage. Using parity encoding and local mean iteration, Cao et al.<sup>[22]</sup> proposed PELMI to reconstruct images from erroneous or missing DNA sequences at the image level. Xu et al.<sup>[23, 24]</sup> established DP-ID and DNA-CTMF through dynamic-programming-based image compression with sequence interleaving and a Pixel-Base codebook, respectively, to transform nucleotide errors into salt-and-pepper noise in reconstructed images and subsequently remove the noise through median filtering. While these methods can effectively reconstruct images under their respective experimental conditions, their reliance on predefined decoding mechanisms or specific image restoration strategies may limit error resilience under complex nucleotide error patterns<sup>[25]</sup>.

Deep learning offers a promising approach to handling the complex errors encountered in DNA storage through its ability to learn error patterns from data and extract useful information from noisy DNA sequences<sup>[26, 27]</sup>. Qin et al.<sup>[28]</sup> introduced RobuSeqNet using attention, a Conformer encoder, and an LSTM decoder to recover reference sequences from noisy read clusters. Bar-Lev et

al. [29] built DNAformer by integrating neural sequence reconstruction with tensor-product error-correcting codes to improve the accuracy and scalability of DNA data retrieval. With the deep-learning-based GatedConvolution and EdgeConnect models, Pan et al. [30] constructed the 2DDNA system to inpaint corrupted regions in images recovered from DNA. To handle missing blocks in biomedical images, Qu et al. [31] proposed HELIX using index-free error correction and deep-learning-based repair. Sequence-reconstruction methods focus on recovering DNA sequences, whereas image-restoration methods adopt recovered image regions to infer missing or corrupted content. Sequence-reconstruction methods generally convert model outputs into hard sequence predictions, whereas image-restoration methods infer missing or corrupted content from recovered image regions. Neither approach fully exploits the uncertainty retained in low-probability byte predictions from noisy DNA sequences, potentially limiting image reconstruction quality.

To address this limitation, we propose PALR-DNA, a probability-aware latent recovery framework for error-resilient image storage in DNA, as illustrated in Fig. 1. First, an autoencoder (AE) [32] is employed to

transform an RGB image into spatially organized, quantized latent vectors, each of which is serialized as a byte payload, combined with its spatial index and a cyclic redundancy check (CRC) field [33], and mapped to an independent DNA strand. This strand-wise organization localizes corruption to individual latent positions, while the index and CRC enable spatial placement and candidate validation. During retrieval, each variable-length noisy DNA sequence is represented using base and positional embeddings and processed by the Transformer encoder [34] in DLRT. Learnable byte queries then attend to the encoded sequence through multi-head cross-attention, producing field-aligned outputs that form a byte probability matrix. This design avoids fixed-length segmentation and reduces alignment errors while retaining prediction uncertainty for subsequent recovery. Top-K candidate beam search (TCBS) therefore retains the K most probable byte values at each output position and preserves the B highest-scoring partial sequences during expansion. Complete candidates are ranked according to their joint probabilities, and the highest-scoring candidate with a valid unoccupied index and a CRC consistent with its index and payload is accepted. However, some severely corrupted sequences may not yield an acceptable candidate at high error rates. For these unresolved sequences, bitwise maximum-probability estimation (BMPE) estimates the payload from the remaining marginal probabilities and selects the most probable valid unoccupied position from the retained index information. The recovered and estimated vectors are subsequently assembled into the latent grid and decoded by the AE to reconstruct the image. The effectiveness of PALR-DNA was evaluated at varying error rates, and the results showed that PALR-DNA consistently outperformed the baseline methods including PELMI and HELIX in image reconstruction quality.

## 2. Methods

### 2.1 DNA Encoding

An AE is utilized to encode an RGB image  $x$  into a spatially organized, quantized latent representation  $Z$ , from which the decoder reconstructs  $x$ . Three stride-2 convolutional stages were used for encoding, and three transposed-convolutional stages were used for decoding. Residual blocks and LeakyReLU activations were used in the intermediate layers, and sigmoid was applied to the output layer. At each spatial index  $r$ ,  $Z$  contains a local latent vector  $z_r$ . For the purpose of preserving pixel-level fidelity and structural similarity, the AE is optimized through a composite loss based on mean squared error (MSE) [35], mean absolute error (MAE) [36], and the multiscale structural similarity index measure (MS-SSIM) [37]. Latent values were quantized by unit-step rounding, clipped to  $[-128, 127]$ , stored as signed 8-bit integers, and converted to floating-point values without rescaling during decoding. Each latent vector  $z_r$  is serialized as a byte payload and combined with its spatial index  $r$  and a CRC field. The index records the position of the latent vector in  $Z$ . The CRC is calculated from the index and payload to verify their consistency during recovery.

The index, payload, and CRC bytes are mapped to DNA codewords using a byte-to-base codebook. The codebook is constructed to maintain sufficient pairwise separation under the Hamming and Levenshtein distances while satisfying the GC-content and homopolymer constraints [38]. For each indexed latent vector, an independent DNA strand  $s_r$  is produced by concatenating the codewords.

### 2.2 DNA-to-Latent Recovery Transformer

Insertions and deletions alter sequence lengths and shift the boundaries between adjacent byte codewords. The resulting alignment errors may propagate throughout a DNA strand when fixed-length nucleotide segments are decoded independently. To avoid explicit codeword segmentation, DLRT is designed to map each variable-length noisy DNA sequence directly to byte probability distributions for the index, latent payload, and CRC fields. For a noisy sequence  $\mathbf{s} = (s_1, \dots, s_T)$ , the input representation at position  $t$  is constructed by combining the corresponding base and positional embeddings:

$$e_t = E_{\text{base}}(s_t) + E_{\text{pos}}(t) \# (1)$$

where  $e_t$  denotes the input embedding at position  $t$ ,  $s_t$  represents the nucleotide token at that position, and  $E_{\text{base}}$  and  $E_{\text{pos}}$  denote the base and positional embedding functions, respectively.

The embedded sequence is processed by a Transformer encoder to obtain contextual representations  $h_1, \dots, h_T$ . Each encoder layer comprises multi-head self-attention and a position-wise feedforward neural network, with residual connections and layer normalization. Padded positions are excluded from self-attention and cross-attention using a validity mask. For each output byte position  $i$ , a learnable byte-query embedding  $q_i$  is introduced. Multi-head cross-attention uses  $q_i$  as the query and the valid contextual representations as keys and values. The corresponding field-aligned output vector for each attention head is computed by:

$$a_i = \sum_{t \in \mathcal{T}_{\text{valid}}} \alpha_{i,t} h_t W^V \# (2)$$

where  $\mathcal{T}_{\text{valid}}$  contains the valid input positions,  $\alpha_{i,t}$  represents the cross-attention weight between query  $q_i$  and input position  $t$ , and  $W^V$  denotes the value-projection matrix. The attention-head index is omitted for simplicity, and the outputs from all heads are combined to obtain the final field-aligned representation  $f_i$ . This representation is obtained without assuming fixed codeword boundaries.

Each field-aligned representation  $f_i$  is layer-normalized and passed through the first linear layer of a shared two-layer MLP, followed by GELU activation and dropout, to obtain the hidden representation  $u_i$ . The second linear layer maps  $u_i$  to logits over the 256 possible byte values, which are normalized using a softmax function:

$$p_i(c | \mathbf{s}) = \frac{\exp[(W_o u_i + b_o)c]}{\sum_{c'=0}^{255} \exp[(W_o u_i + b_o)c']} \# (3)$$

where  $c$  denotes a candidate byte value, and  $W_o$  and  $b_o$  are the weight matrix and bias vector of the MLP output layer, respectively. Collectively, these distributions form a probability matrix covering the index, latent payload, and CRC fields. DLRT is trained on noisy sequences using the bitwise cross-entropy loss:

$$L = -\frac{1}{L} \sum_{i=1}^L \sum_{c \in \mathcal{V}} \mathcal{V} y_{i,c} \log p_i(c | \mathbf{s}) \# (4)$$

where  $L$  is the total number of output byte positions,  $\mathcal{V} = \{0, \dots, 255\}$  is the set of possible byte values, and  $y_{i,c}$  denotes the target distribution for byte value  $c$  at position  $i$ . The resulting probability matrix is subsequently provided to TCBS and BMPE for latent-vector recovery.

## 2.3 DNA Decoding

The latent vectors are recovered through two consecutive stages. TCBS is initially applied to generate a ranked list of candidate byte sequences. At each output position, the  $K$  most probable byte values are retained, and their combinations are explored with a beam size of  $B$ . A candidate sequence  $\mathbf{c} = (c_1, \dots, c_L)$ , comprising the index, latent payload, and CRC fields, is scored by:

$$S(\mathbf{c} | \mathbf{s}) = \sum_{i=1}^L \log p_i(c_i | \mathbf{s}) \# (5)$$

where  $c_i$  denotes the byte value at position  $i$  in candidate sequence  $\mathbf{c}$ , and  $p_i(c_i | \mathbf{s})$  denotes the probability assigned to  $c_i$  given the noisy sequence  $\mathbf{s}$ . After each expansion, only the  $B$  highest-scoring partial candidates are retained. After the complete candidates have been ranked, the highest-scoring candidate with a valid decoded index and a CRC field matching the checksum

recalculated from its index and payload is accepted. Its payload is converted into a recovered latent vector  $\hat{\mathbf{z}}_r$  and assigned to the corresponding unoccupied spatial position  $r$ .

For each unresolved sequence for which TCBS finds no acceptable candidate, BMPE estimates the latent payload and spatial index from the remaining probability information. At payload-byte position  $m$ , the maximum-probability byte value is selected from the corresponding marginal distribution and converted into a quantized latent value, expressed as:

$$\hat{z}_m = g \left( \underset{b \in \mathcal{V}}{\operatorname{argmax}} p_m^{\text{payload}}(b \mid \mathbf{s}) \right) \# (6)$$

where  $m \in \{1, \dots, L_{\text{payload}}\}$ ,  $L_{\text{payload}}$  represents the number of payload bytes, and  $g(\cdot)$  converts the selected byte value into its corresponding quantized latent value. Candidate spatial positions are constructed from the retained index-byte values after invalid and occupied positions are excluded, and the resulting set is denoted by  $\mathcal{A}_K$ . The position with the highest joint index probability is expressed as

$$\hat{r} = \underset{r \in \mathcal{A}_K}{\operatorname{argmax}} \sum_{i=1}^{L_{\text{idx}}} \log p_i^{\text{idx}}(b_i^{(r)} \mid \mathbf{s}) \# (7)$$

where  $L_{\text{idx}}$  represents the number of index bytes, and  $b_i^{(r)}$  denotes the  $i$ -th byte of candidate index  $r$ . The estimated latent vector is assigned to the selected position  $\hat{r}$ . The sequence remains unresolved if  $\mathcal{A}_K$  is empty. After all sequences have been processed, the recovered latent vectors are assembled into  $\hat{\mathbf{Z}}$ , which is decoded to obtain the reconstructed image  $\hat{\mathbf{x}}$ . An example of the BMPE procedure is detailed in Supplementary Fig. S1.

## 2.4 Evaluation Metrics

Image reconstruction quality is evaluated using MSE, peak signal-to-noise ratio (PSNR) [39], structural similarity index measure (SSIM) [40], and MS-SSIM. MSE and PSNR quantify pixel-level reconstruction fidelity. SSIM and MS-SSIM assess structural similarity at single and multiple scales, respectively. Lower MSE and higher PSNR, SSIM, and MS-SSIM indicate better reconstruction quality.

Furthermore, latent recovery is assessed using latent-vector recovery accuracy. It is defined as the proportion of latent vectors for which both the recovered spatial index and the complete payload match the ground truth. Missing, incorrectly placed, or payload-mismatched vectors are considered recovery failures. The mean and standard error of each metric are reported across the evaluation images.

## 3. Results

### 3.1 Setup for PALR-DNA

In this study, 5,431 images were selected from the WikiART Face dataset [41] and divided into 4,346 training, 543 validation, and 542 test images. All WikiART Face images were converted to RGB and resized to  $512 \times 512$  pixels. A portrait of Confucius obtained from Wikimedia Commons [42] was cropped and resized to the same dimensions and added to the test set for visualizing reconstruction quality, yielding 543 evaluation images. Using Adam [43], the AE was trained on random  $256 \times 256$  crops for 100 epochs with a learning rate of  $1 \times 10^{-4}$  and a batch size of 16. The MSE, MAE, and  $1 - \text{MS-SSIM}$  loss terms were weighted at 20:2:1. Every  $512 \times 512$  image was encoded into a  $16 \times 64 \times 64$  latent grid, quantized to signed 8-bit values, and divided into 4,096 16-byte payloads, each with a 2-byte index and a 4-byte CRC-32 field. Mapping each byte to a 7-nt codeword yielded a 154-nt DNA strand. The codebook enforced minimum Hamming and Levenshtein distances of 3, three or four GC bases per codeword, and a maximum homopolymer length of four nucleotides. The changes in data representation and size are summarized in Supplementary Fig. S2.

For DLRT, three Transformer layers, a 128-dimensional embedding, four attention heads, and 22 learned byte queries were utilized for the index, payload, and CRC fields. Input sequences were truncated or padded to 192 nucleotides. The model was trained for 50 epochs using AdamW<sup>[44]</sup> with a learning rate of  $3 \times 10^{-4}$  and a batch size of 1,024. During training, substitution rates were sampled from 0.4% to 3.5%, while insertion and deletion rates were independently sampled from 0.2% to 1.75%. The AE and DLRT checkpoints were selected according to validation PSNR and byte-level Top-8 accuracy, respectively.

Unless otherwise specified, base error rates of 1–5% were evaluated with substitutions, insertions, and deletions at a 2:1:1 ratio<sup>[45]</sup>, and  $K$  and  $B$  were set to 8 and 256, respectively, in TCBS. Image quality was evaluated on images normalized to [0, 1]. For reporting, MSE was multiplied by 255<sup>2</sup> to express it on the 8-bit intensity scale. PSNR was calculated using a peak value of 1. The mean and standard error over 543 evaluation images were reported in this study. PELMI<sup>[22]</sup> and HELIX<sup>[31]</sup> were evaluated with the same images, error profiles, and random seeds as PALR-DNA. All experiments used the same random seed and were conducted on an NVIDIA GeForce RTX 5090 GPU with 32 GB of memory and 25 Intel Xeon Platinum 8470Q vCPUs.

## 3.2 Comparison with Existing DNA-Based Image Storage Methods

PALR-DNA was compared with PELMI and HELIX on the test set at error rates ranging from 1% to 5% to evaluate reconstruction performance under increasing DNA error rates. In PELMI, identifier bits locate erroneous sequences, and Reed-Solomon decoding corrects recoverable data. Remaining erroneous or missing positions are identified from the sequence addresses and restored using neighboring pixel values. In HELIX, an inner Levenshtein-marker code corrects errors within sequences; an outer Reed-Solomon code addresses sequence loss. Blocks failing the CRC check are identified for context-based image repair. Candidate latent vectors whose indices are valid and whose CRC fields pass verification are selected for PALR-DNA. When no candidate is accepted, the unresolved latent vector is estimated according to the remaining byte probabilities.

As shown in Fig. 2a, PELMI preserved the overall image contours but exhibited substantial loss of fine details and spatially irregular loss of RGB-channel information. HELIX correctly reconstructed some pixel blocks, while other blocks remained missing. The visual quality of both baseline methods deteriorated progressively as the error rate increased. In contrast, PALR-DNA preserved the complete image structure and favorable visual fidelity across all evaluated error rates. For the representative Confucius portrait, the error-free reconstruction achieved an MSE of 12.2688, a PSNR of 37.24 dB, an SSIM of 0.9377, and an MS-SSIM of 0.9915. At a 1% error rate, the reconstructed image and all four evaluation metrics remained identical to those obtained under the error-free condition. As the error rate increased further, degradation was largely limited to a small number of missing or blurred pixel blocks and did not substantially affect the overall image content.

The quantitative results in Fig. 2b show that the reconstruction quality of both PELMI and HELIX deteriorated substantially as the error rate increased, with higher MSE and lower PSNR, SSIM, and MS-SSIM than PALR-DNA across all evaluated error rates. In contrast, PALR-DNA maintained consistently and markedly higher reconstruction quality. In the absence of base errors, it achieved a mean MSE of 32.3692, a mean PSNR of 34.94 dB, a mean SSIM of 0.9164, and a mean MS-SSIM of 0.9882 across all 543 evaluation images. Although its reconstruction quality gradually decreased as the error rate increased, its advantage over both baseline methods became more pronounced. At an error rate of 5%, PALR-DNA retained a PSNR of 27.45 dB and an MS-SSIM of 0.918, exceeding PELMI and HELIX by 16.83 and 17.58 dB in PSNR and by 0.756 and 0.799 in MS-SSIM, respectively.

## 3.3 DLRT output probability analysis

DLRT processes each variable-length noisy DNA sequence and predicts a probability distribution over the 256 possible values at each of the 22 output byte positions, comprising two index bytes, 16 payload bytes, and four CRC bytes. Whether these distributions retained reliable candidate information with increasing DNA errors was evaluated by examining the rank assigned to each ground-truth byte at different error rates (Fig. 3). As the error

rate rose from 1% to 5%, the overall Top-1 proportion dropped from 99.79% to 96.56%, whereas the Top-8 proportion remained high, decreasing only from 99.994% to 99.680%. At an error rate of 5%, the Top-8 candidates contained the ground-truth byte in 99.68% of predictions, which was 3.12 percentage points higher than the Top-1 proportion. The payload bytes were the most stable, with Top-1 and Top-8 proportions of 97.58% and 99.89%, respectively, followed by the index bytes at 96.73% and 99.61%. The CRC bytes were more sensitive to noise; nevertheless, their Top-8 proportion remained 98.86%, despite a Top-1 proportion of 92.37%. Thus, increasing DNA error rates primarily changed the ranking of candidate byte values rather than eliminating the

correct value from the high-probability candidates, supporting the use of Top-K probability distributions in subsequent TCBS decoding.

### 3.4 Parameter Sensitivity of TCBS Decoding

In TCBS, the  $K$  most probable DLRT predictions were retained at each byte position, and a beam size of  $B$  was used to preserve the highest-scoring partial candidate sequences. The highest-scoring complete candidate with a valid index and matching CRC was accepted. Reconstruction quality became more sensitive to both search parameters at higher error rates. With  $B$  fixed at 256, increasing  $K$  from 1 to 8 at a 5% error rate improved

**Figure 4. Sensitivity analysis of TCBS parameters.** (a) Results for  $K$  values of 1, 2, 4, 8, and 16 at  $B = 256$ . (b) Results for  $B$  values of 32, 64, 128, 256, and 512 at  $K = 8$ .

PSNR and MS-SSIM from 20.59 to 27.45 dB and from 0.696 to 0.918, respectively (Fig. 4a). At  $K = 16$ , PSNR and MS-SSIM reached 28.81 dB and 0.939, respectively. In contrast, with  $K$  fixed at 8, increasing  $B$  from 32 to 256 raised PSNR from 26.44 to 27.45 dB and MS-SSIM from 0.896 to 0.918 at the same error rate (Fig. 4b). At  $B = 512$ , the corresponding values reached 27.66 dB and 0.922, respectively. Overall, reconstruction quality was more strongly affected by  $K$  than by  $B$ , as increasing  $K$  produced substantially larger improvements in both PSNR and MS-SSIM over the evaluated ranges. Increasing either parameter improved reconstruction quality, although the additional gains became progressively smaller at higher parameter values.

### 3.5 Image Reconstruction Quality under Different DNA Error Compositions

The robustness to different DNA error compositions was evaluated by reconstructing 543 evaluation images at 1–5% error rates under five substitution:insertion:deletion ratios of 8:1:1, 4:1:1, 2:1:1, 1:1:1, and 1:2:2. Reconstruction quality declined as the error rate increased under every composition. At a 1% error rate, differences among compositions were small, with PSNR and MS-SSIM varying by only 0.11 dB and 0.0002, respectively (Fig. 5). These differences became more pronounced at higher error rates. At 5%, the 8:1:1 condition yielded a PSNR of 27.94 dB, an SSIM of 0.879, and an MS-SSIM of 0.927. The corresponding values under the 1:2:2 condition, which contained higher proportions of insertions and deletions, were 26.86 dB, 0.858, and 0.904, respectively. The results for the intermediate mixtures fell between these conditions. Thus, PALR-DNA maintained comparable performance across the tested compositions, even though a higher proportion of insertion and deletion errors moderately reduced reconstruction quality.

### 3.6 Ablation study

The contribution of each recovery component was assessed by comparing complete PALR-DNA with variants excluding DLRT, TCBS, or BMPE. In the w/o DLRT variant, neural byte-probability prediction was replaced with edit-distance-based beam-search decoding of DNA codewords followed by CRC checking. In the w/o TCBS variant, decoding was restricted to Top-1 byte predictions, while BMPE remained enabled. In the w/o BMPE variant, latent positions unresolved by TCBS were left at their initial values. All variants were evaluated at error rates ranging from 1% to 5%, with the results at 5% presented in Table 1 and those at 1–4% provided in Supplementary Table S1.

Complete PALR-DNA achieved the highest reconstruction quality at every evaluated error rate, demonstrating larger differences from the ablated variants with the increasing error rate. At 5%, it reached a PSNR of 27.45 dB and an MS-SSIM of 0.918, compared with 11.11 dB and 0.200 without DLRT, 20.59 dB and 0.696 without TCBS, and 19.28 dB and 0.660 without BMPE, respectively. Replacing DLRT with hard decoding caused the largest performance decline, consistent with the loss of byte-level probability information required for subsequent recovery. Restricting decoding to Top-1 predictions lessened the benefit of searching for CRC-valid candidates. The contribution of BMPE became more apparent at higher error rates. Specifically, omitting it reduced PSNR relative to complete PALR-DNA by 1.65 dB at 1% and 8.17 dB at 5%. The visual results suggested the same overall trend, with the complete framework producing the fewest visible artifacts (Supplementary Fig. S3). Thus, (1) DLRT provides effective byte-level probability distributions for noisy DNA sequences; (2) TCBS improves latent-vector recovery by examining multiple candidate sequences and retaining candidates that have valid indices and pass the CRC check; (3) BMPE

improves image reconstruction quality by approximating unresolved latent vectors from the remaining probabilities for the index and payload bytes.

Table 1  
Comparison among PALR-DNA variants at 5% error rate. “w/o” denotes “without”. “↓” and “↑” arrows indicate that lower and higher values are better.

| Variant  | MSE ↓           | PSNR (dB) ↑  | SSIM ↑          | MS-SSIM ↑       |
|----------|-----------------|--------------|-----------------|-----------------|
| w/o DLRT | 5392.58 ± 84.00 | 11.11 ± 0.07 | 0.1409 ± 0.0015 | 0.1998 ± 0.0035 |
| w/o TCBS | 612.92 ± 11.15  | 20.59 ± 0.07 | 0.7088 ± 0.0024 | 0.6959 ± 0.0034 |
| w/o BMPE | 888.18 ± 23.13  | 19.28 ± 0.10 | 0.6993 ± 0.0029 | 0.6604 ± 0.0049 |
| PALR-DNA | 136.90 ± 4.32   | 27.45 ± 0.10 | 0.8707 ± 0.0026 | 0.9182 ± 0.0011 |

## 4. Discussion

Under the matched evaluation protocol, PALR-DNA consistently achieved better reconstruction performance than PELMI and HELIX. Its advantage increased with the DNA error rate (Fig. 2). This performance difference is closely associated with the use of byte-level probability distributions during retrieval. PELMI combines parity protection with local mean iteration. HELIX reconstructs image blocks that cannot be recovered through sequence-level decoding. PELMI becomes less effective when reliable neighboring pixels or color-channel values are unavailable, leading to the loss of fine details. Similarly, as more blocks remain unresolved, less surrounding context is available for HELIX to repair them, resulting in greater structural degradation. In contrast, PALR-DNA retains the probability distribution over possible byte values instead of immediately converting each noisy sequence into a single hard prediction. TCBS retains multiple high-probability candidate sequences and selects the highest-ranked candidate that has a valid index and passes the CRC check. Subsequently, BMPE uses the remaining probabilities for the index and payload bytes when TCBS finds no acceptable candidate. This probability-aware strategy preserves information that would otherwise be discarded as the error rate increases and more sequences become ambiguous, accounting for the increasingly pronounced advantage of PALR-DNA over the baseline methods.

This interpretation is further supported by the analyses of DLRT outputs and TCBS parameters. The ground-truth bytes remained highly concentrated among the top-ranked DLRT predictions even when they were not ranked first (Fig. 3). Hence, prediction errors typically involved changes in candidate ranking rather than the complete loss of the correct value. TCBS benefits from this property by examining lower-ranked candidates and retaining combinations whose decoded indices are valid and whose CRC fields match the recalculated CRCs. At a 5% error rate, increasing  $K$  from 1 to 8 with  $B$  fixed at 256 raised latent-vector recovery accuracy from 66.72% to 93.61%. Concomitantly, PSNR and MS-SSIM increased from 20.59 dB and 0.696 to 27.45 dB and 0.918, respectively. These results indicate that enhanced latent-vector recovery translates into higher reconstruction quality. The greater influence of  $K$  than of  $B$  is consistent with this mechanism:  $K$  determines whether the correct byte remains available during the search, whereas  $B$  controls only the number of retained candidate sequences. However, expanding the TCBS search space imposed a substantial computational cost. At  $K = 8$  and  $B = 256$ , the mean decoding time was 31.13 s per image. Increasing  $K$  to 16 while holding  $B = 256$  increased the decoding time to 133.76 s per image, whereas increasing  $B$  to 512 while holding  $K = 8$  increased it to 114.18 s per image (Supplementary Fig. S5). These results suggest that  $K = 8$  and  $B = 256$  provide a practical balance between recovery performance and decoding efficiency. Moreover, DLRT processes entire variable-length sequences without relying on fixed codeword boundaries, contributing to the high reconstruction quality of PALR-DNA at high error rates under error compositions containing higher proportions of insertions and deletions (Fig. 5). Taken together, the superior performance of PALR-DNA arises from retaining byte-level probability distributions and using them for candidate generation, CRC-based selection, and bitwise maximum-probability estimation.

The ablation study confirms the importance of each recovery component (Table 1, Supplementary Table S1, and Supplementary Fig. S3). Removing DLRT caused the largest performance degradation. In this variant, neural byte prediction was replaced by edit-distance-based decoding of DNA codewords. Although edit distance identifies locally similar codewords, it does not learn contextual error patterns across the complete variable-length sequence and is more sensitive to codeword-boundary shifts caused by insertions and deletions. Hard decoding also removes the byte-level probability distributions required by TCBS and

BMPE. Thus, predictions across all output byte positions and their associated byte-level probability distributions establish the foundation of the recovery process. Without TCBS, decoding was restricted to top-1 predictions, excluding lower-ranked but potentially correct byte values from candidate selection and CRC checking. Removing BMPE left the latent positions unresolved after TCBS, and the resulting performance gap became larger at higher error rates. This result indicates that the residual index and payload probabilities remain useful when TCBS finds no acceptable candidate. The complete framework achieved the best performance because DLRT provides byte-level probability distributions, TCBS selects candidates with valid indices that pass the CRC check, and BMPE reduces information loss by approximating unresolved latent vectors when no candidate is accepted. Hence, the three modules address different stages of the recovery process and make complementary contributions to error-resilient image reconstruction.

The present findings should be interpreted within the evaluated conditions. The experiments were based on simulated substitution, insertion, and deletion errors. Physical DNA storage may also involve strand dropout, sequence-dependent errors, amplification bias, and uneven sequencing coverage<sup>[46–48]</sup>. Additionally, BMPE produces approximate latent estimates without CRC checking; TCBS incurs increasing computational cost as its search space expands. Thus, PALR-DNA should be validated in future work through synthesized DNA and real sequencing data, followed by the evaluation of more diverse image datasets and storage configurations, as well as improvements to the reliability and efficiency of approximate latent recovery.

## 5. Conclusion

In conclusion, PALR-DNA, a probability-aware latent recovery framework, was developed for error-resilient image storage in DNA. PALR-DNA features a spatially organized latent image representation with DLRT, TCBS, and BMPE to recover images from noisy DNA sequences. By retaining and exploiting byte-level probability distributions, the framework combines the selection of candidates passing the CRC check with BMPE-based approximation when no such candidate is available. The simulation results revealed that PALR-DNA consistently outperformed the baseline methods, with a more pronounced advantage at higher error rates, while maintaining robust reconstruction performance across different error compositions. The complementary contributions of the three recovery components were further confirmed by the ablation results. Collectively, this probability-aware recovery strategy shows strong potential for robust DNA-based image storage, while further validation using synthesized DNA and real sequencing data remains necessary.

## Declarations

## Competing interests

The authors declare no competing interests.

## Author Contribution

S.C.: conceptualization, methodology, investigation, data analysis, and writing—original draft. R.Z., T.W., T.H. and X.H.: investigation and data analysis. Q.S.: conceptualization, supervision, project administration, funding acquisition, and writing—review and editing. All authors reviewed and approved the final manuscript.

## Acknowledgements

This work is funded by National Key Research and Development Program (2023YFF1206102) and National Natural Science Fund of China (21976028).

## Data Availability

The datasets used in this study are publicly available and properly cited as follows: (1) The WikiART Face dataset is available on Hugging Face: <https://huggingface.co/datasets/asahi417/wikiart-face>. (2) The portrait of Confucius is available from Wikimedia Commons at <https://upload.wikimedia.org/wikipedia/commons/b/b4/%E5%AD%94%E5%AD%90%E7%87%95%E5%B1%85%E5%83%8F.png>.

## Code Availability

The code for PALR-DNA is available at <https://github.com/csq2001/PALR-DNA>.

## References

1. Wang, Y. et al. Advances and challenges in non-canonical nucleic acids data storage. *Nat. Commun.* **17**, 2354 (2026).
2. Rebimbas, R. et al. DNA as a data storage medium. *J. Biotechnol.* **414**, 19–35 (2026).
3. Bar-Lev, D., Sabary, O. & Yaakobi, E. The zettabyte era is in our DNA. *Nat. Comput. Sci.* **4**, 813–817 (2024).
4. Noor, J. et al. Secure processing-aware media storage and archival (SPMSA). *Future Gener Comput. Syst.* **159**, 290–306 (2024).
5. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
6. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew Chem. Int. Ed.* **54**, 2552–2555 (2015).
7. Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **20**, 456–466 (2019).
8. Zhang, J. et al. Gungnir codec enabling high error-tolerance and low-redundancy DNA storage through substantial computing power. *Nat. Commun.* **17**, 4828 (2026).
9. Wang, S., Mao, X., Wang, F., Zuo, X. & Fan, C. Data storage using DNA. *Adv. Mater.* **36**, 2307499 (2024).
10. Dong, Y., Sun, F., Ping, Z., Ouyang, Q. & Qian, L. DNA storage: research landscape and future prospects. *Natl. Sci. Rev.* **7**, 1092–1107 (2020).
11. Matange, K., Tuck, J. M. & Keung, A. J. DNA stability: a central design consideration for DNA data storage systems. *Nat. Commun.* **12**, 1358 (2021).
12. Song, L. et al. Robust data storage in DNA by de Bruijn graph-based de novo strand assembly. *Nat. Commun.* **13**, 5361 (2022).
13. Wang, K. et al. Storing images in DNA via Base128 encoding. *J. Chem. Inf. Model.* **64**, 1719–1729 (2024).
14. Doricchi, A. et al. Emerging approaches to DNA data storage: challenges and prospects. *ACS Nano*. **16**, 17552–17571 (2022).
15. Wang, Y., Noor-A-Rahim, M., Gunawan, E., Guan, Y. L. & Poh, C. L. Modelling, characterization of data-dependent and process-dependent errors in DNA data storage. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **20**, 2147–2158 (2023).
16. Yeom, H. et al. Highly accurate sequence- and position-independent error profiling of DNA synthesis and sequencing. *ACS Synth. Biol.* **12**, 3567–3577 (2023).
17. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
18. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
19. Press, W. H., Hawkins, J. A., Jones, S. K. Jr., Schaub, J. M. & Finkelstein, I. J. HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. *Proc. Natl. Acad. Sci. USA* **117**, 18489–18496 (2020).
20. Welzel, M. et al. DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage. *Nat. Commun.* **14**, 628 (2023).
21. Yan, Z., Qu, G., Chen, X., Zheng, G. & Wu, H. DNA StairLoop: enabling high-fidelity data recovery and robust error correction in DNA-based data storage. *Nat. Commun.* **16**, 9191 (2025).
22. Cao, B. et al. PELMI: realize robust DNA image storage under general errors via parity encoding and local mean iteration. *Brief. Bioinform.* **25**, bbae463 (2024).

23. Xu, Q., Ma, Y., Lu, Z. & Bi, K. DP-ID: interleaving and denoising to improve the quality of DNA storage image. *Interdiscip Sci. Comput. Life Sci.* **17**, 306–320 (2025).
24. Xu, Q. et al. DNA-CTMF: reconstruct high quality image from lossy DNA storage via Pixel-Base codebook and median filter. *Synth. Syst. Biotechnol.* **10**, 925–935 (2025).
25. Cao, B. et al. Efficient data reconstruction: the bottleneck of large-scale application of DNA storage. *Cell. Rep.* **43**, 113699 (2024).
26. Zheng, X. & Xie, R. A generative adversarial network for multiple reads reconstruction in DNA storage. *Sci. Rep.* **14**, 32071 (2024).
27. Zheng, Y. et al. INNSE: invertible neural network-based DNA image storage with self-correction encoding. *Comput. Struct. Biotechnol. J.* **27**, 2492–2502 (2025).
28. Qin, Y., Zhu, F., Xi, B. & Song, L. Robust multi-read reconstruction from noisy clusters using deep neural network for DNA storage. *Comput. Struct. Biotechnol. J.* **23**, 1076–1087 (2024).
29. Bar-Lev, D., Orr, I., Sabary, O., Etzion, T. & Yaakobi, E. Scalable and robust DNA-based storage via coding theory and deep learning. *Nat. Mach. Intell.* **7**, 639–649 (2025).
30. Pan, C. et al. Rewritable two-dimensional DNA-based data storage with machine learning reconstruction. *Nat. Commun.* **13**, 2984 (2022).
31. Qu, G., Yan, Z., Chen, X. & Wu, H. DNA data storage for biomedical images using HELIX. *Nat. Comput. Sci.* **5**, 397–404 (2025).
32. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
33. Castagnoli, G., Brauer, S. & Herrmann, M. Optimization of cyclic redundancy-check codes with 24 and 32 parity bits. *IEEE Trans. Commun.* **41**, 883–892 (1993).
34. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* **30**, 5998–6008 (2017).
35. Wang, Z. & Bovik, A. C. Mean squared error: love it or leave it? A new look at signal fidelity measures. *IEEE Signal. Process. Mag.* **26**, 98–117 (2009).
36. Zhao, H., Gallo, O., Frosio, I. & Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging.* **3**, 47–57 (2017).
37. Charrier, C., Knoblauch, K., Maloney, L. T., Bovik, A. C. & Moorthy, A. K. Optimizing multiscale SSIM for compression via MLDS. *IEEE Trans. Image Process.* **21**, 4682–4694 (2012).
38. Löchel, H. F., Welzel, M., Hattab, G., Hauschild, A. C. & Heider, D. Fractal construction of constrained code words for DNA storage systems. *Nucleic Acids Res.* **50**, e30 (2022).
39. Huynh-Thu, Q. & Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **44**, 800–801 (2008).
40. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
41. Ushio, A. WikiART Face [Data set]. *Hugging Face*, (2024). <https://huggingface.co/datasets/asahi417/wikiart-face>
42. Wikimedia Commons. Portrait of Confucius at Leisure. (2025). <https://upload.wikimedia.org/wikipedia/commons/b/b4/%E5%AD%94%E5%AD%90%E7%87%95%E5%B1%85%E5%83%8F.png>
43. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations* (2015).
44. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations* (2019).
45. Heckel, R., Mikutis, G. & Grass, R. N. A characterization of the DNA data storage channel. *Sci. Rep.* **9**, 9663 (2019).
46. Gimpel, A. L., Stark, W. J., Heckel, R. & Grass, R. N. A digital twin for DNA data storage based on comprehensive quantification of errors and biases. *Nat. Commun.* **14**, 6026 (2023).
47. Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* **11**, 616 (2020).
48. Chen, Y. J. et al. Quantifying molecular bias in DNA data storage. *Nat. Commun.* **11**, 3264 (2020).

# Figures

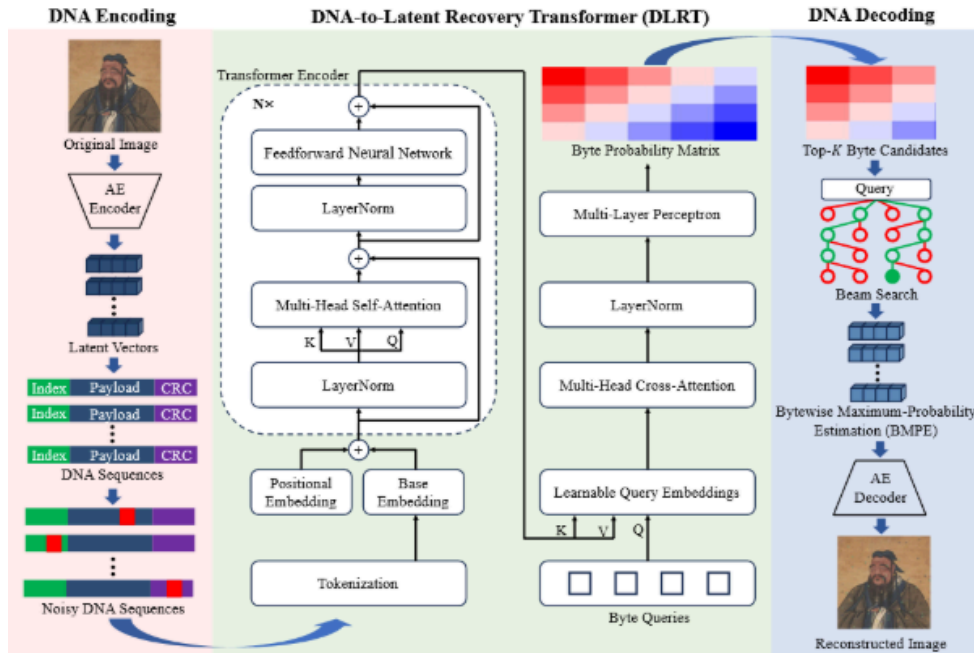


Figure 1

Overview of the PALR-DNA framework for DNA-based image storage.

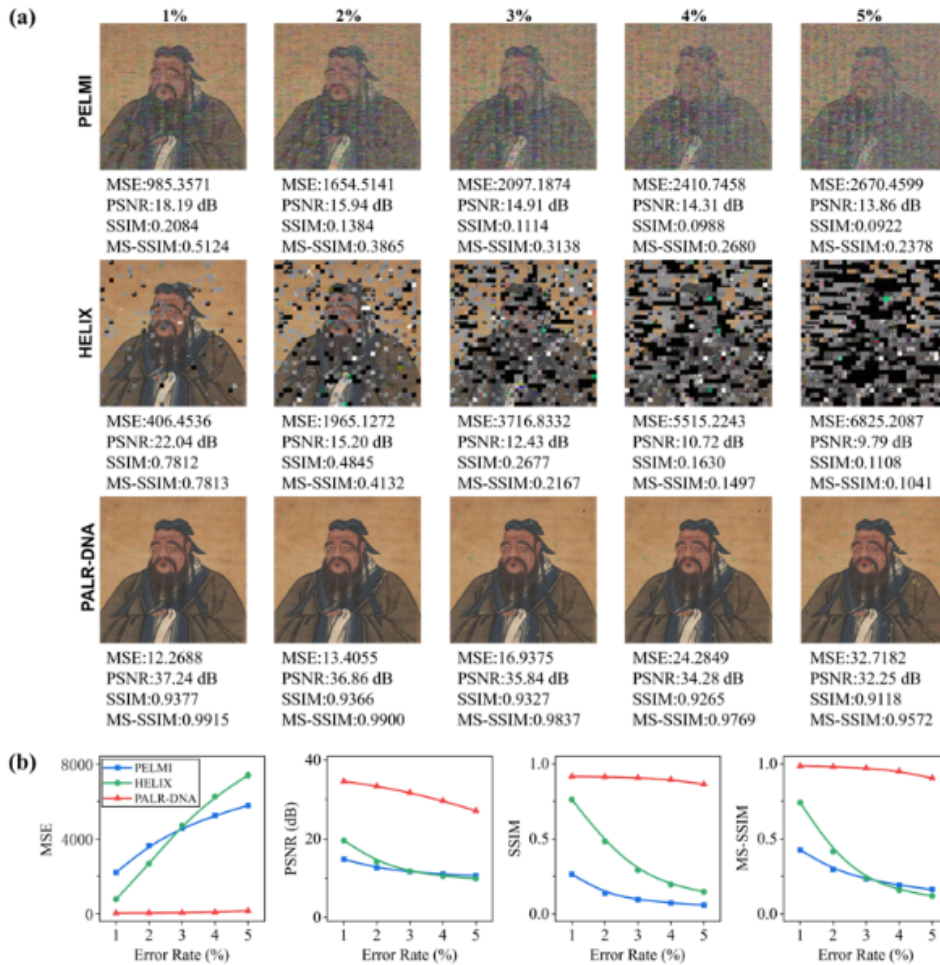


Figure 2

**Image reconstruction by PELMI, HELIX, and PALR-DNA at 1–5% error rates.** (a) Confucius portrait reconstructed by the three methods at 1–5% error rates. (b) Mean evaluation metrics over 543 evaluation images.

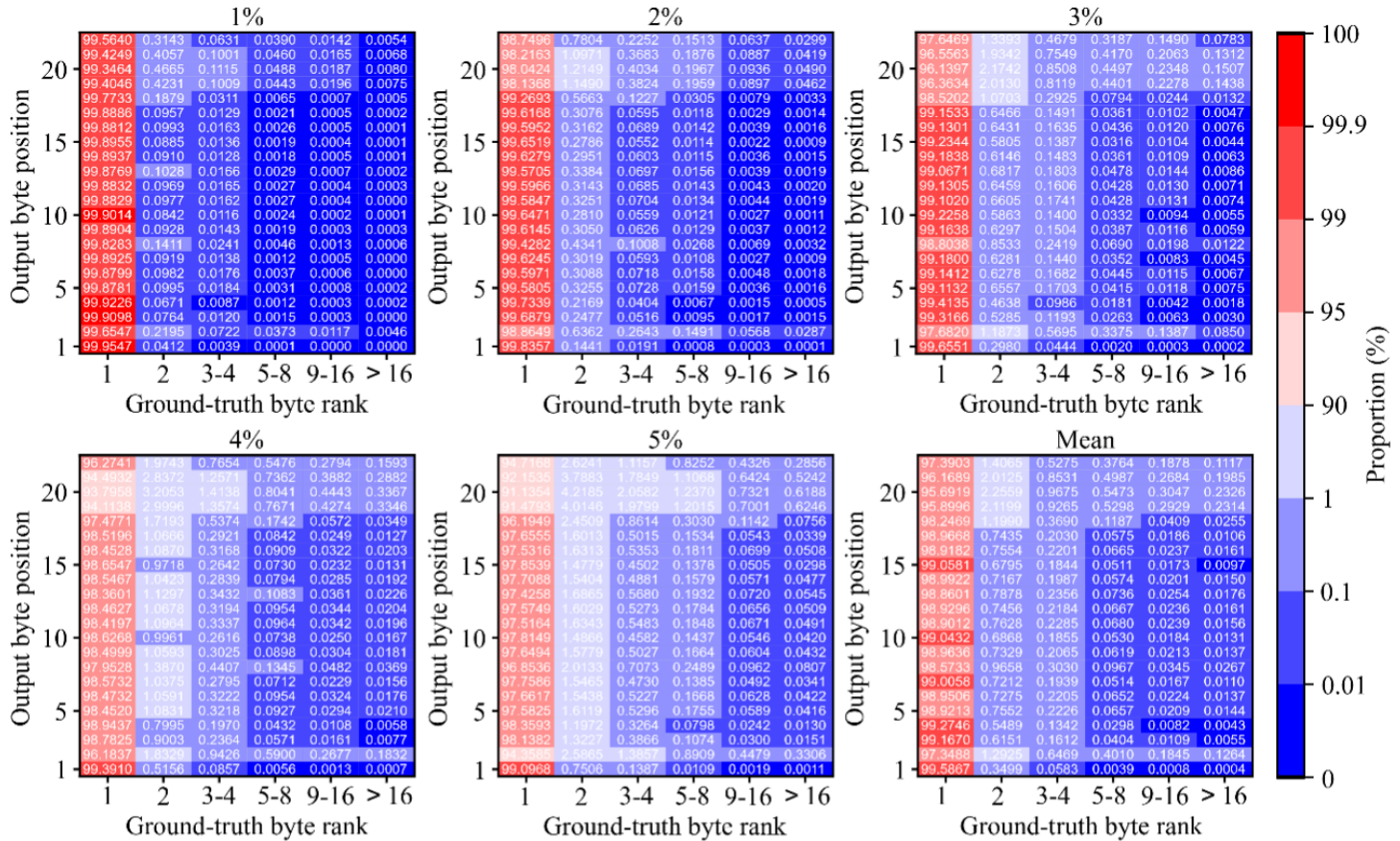


Figure 3

**Rank distribution of ground-truth bytes in the DLRT outputs.** The heatmaps show the percentages of ground-truth byte values at ranks 1, 2, 3–4, 5–8, 9–16, and greater than 16 across the 22 output positions. Positions 1–2, 3–18, and 19–22 correspond to the index, payload, and CRC bytes, respectively. Results are shown for 1–5% error rates and their mean, based on 2,224,128 noisy DNA sequences per error condition from 543 evaluation images.

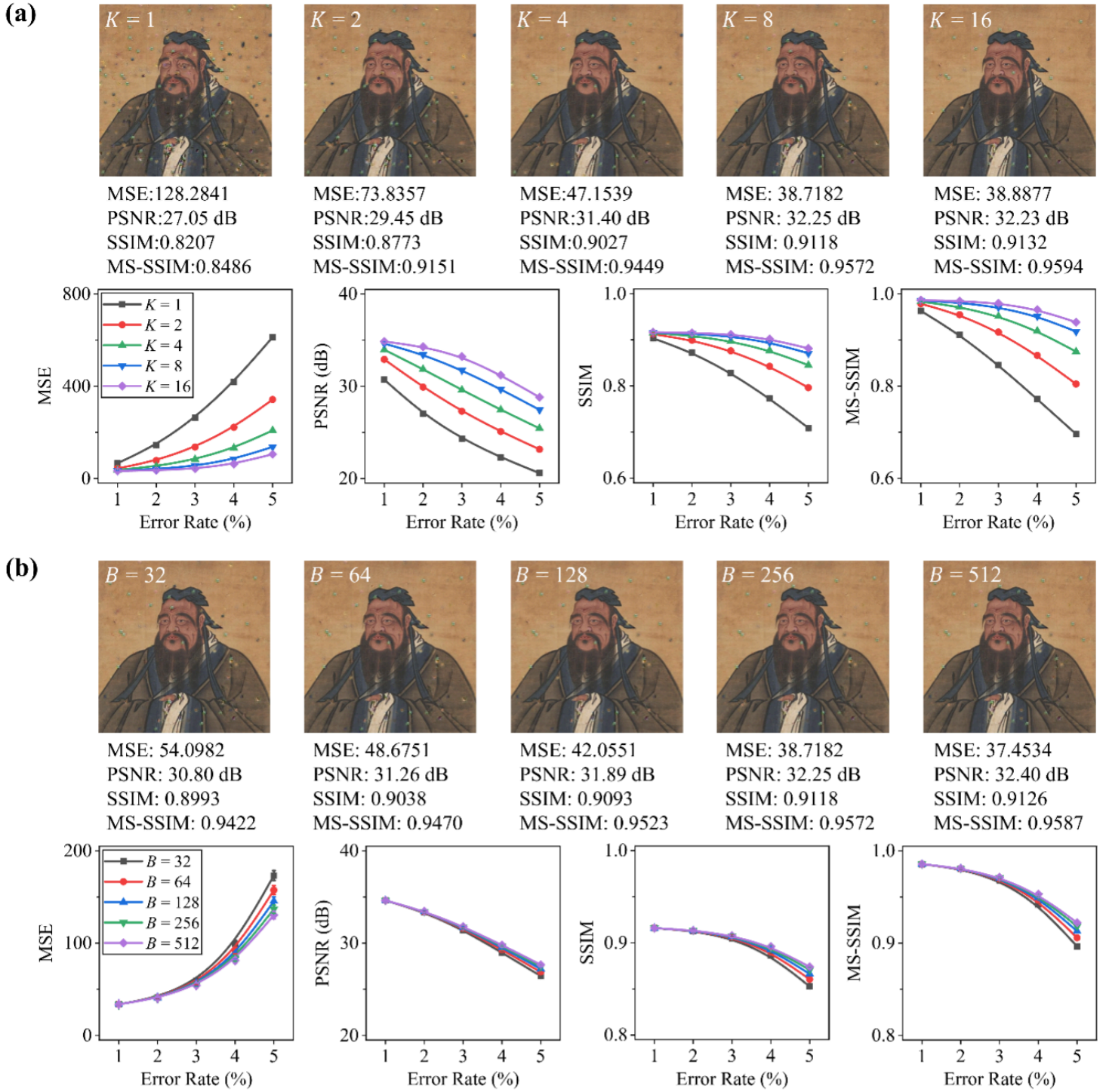


Figure 4

Sensitivity analysis of TCBS parameters. (a) Results for  $K$  values of 1, 2, 4, 8, and 16 at  $B = 256$ . (b) Results for  $B$  values of 32, 64, 128, 256, and 512 at  $K = 8$ .

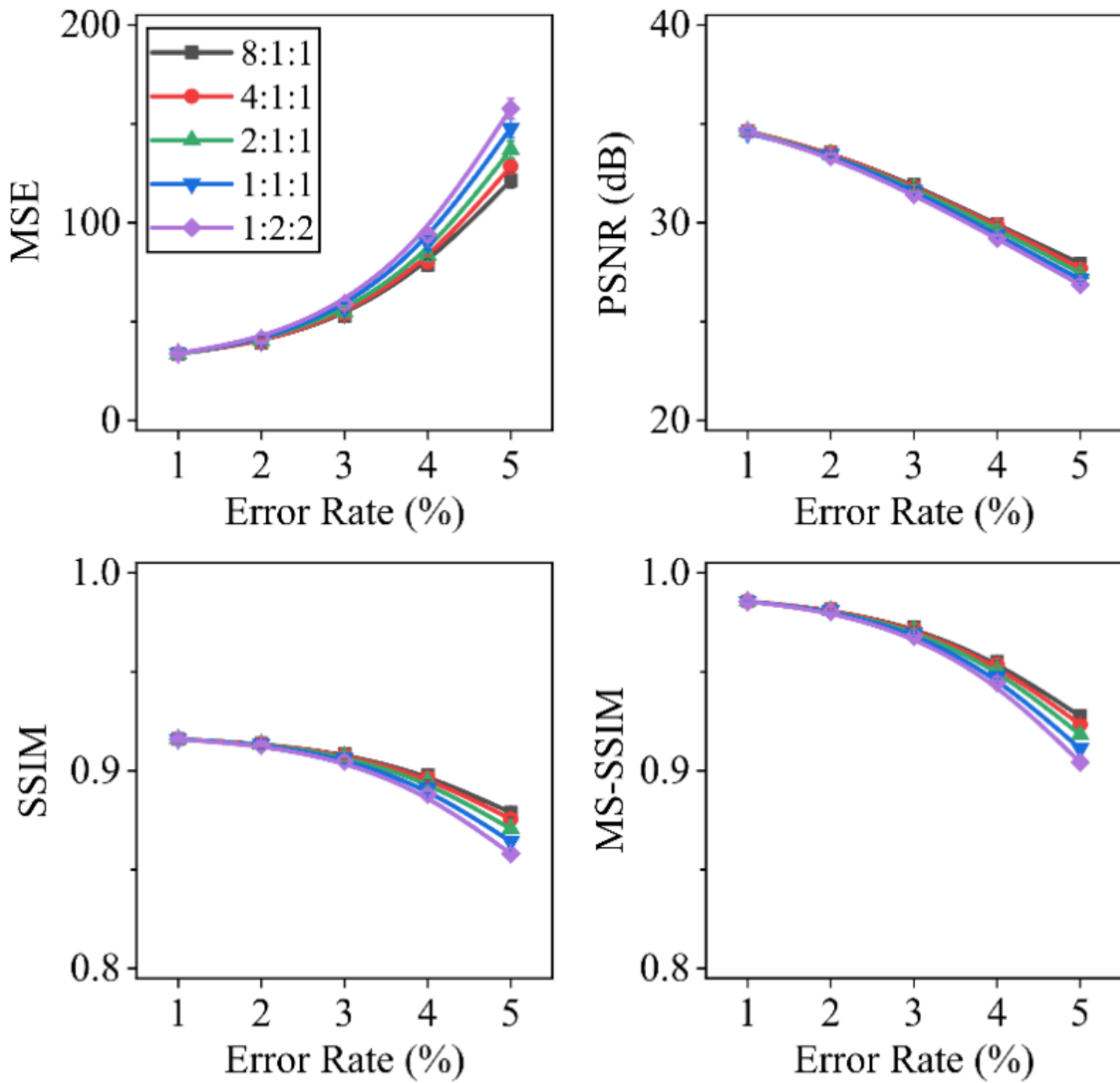


Figure 5

**Image reconstruction quality under different DNA error compositions.** MSE, PSNR, SSIM, and MS-SSIM are shown for substitution:insertion:deletion ratios of 8:1:1, 4:1:1, 2:1:1, 1:1:1, and 1:2:2 at 1–5% error rates.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation0929.docx](#)