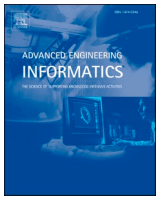
Advanced Engineering Informatics 77 (2027) 105296



Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



DNA-FRSC: A robust system for DNA-based image storage via deep visual feature representations and error-resilient sequence coding

Zhuangzhuang Zang [a], Shi Jia [a,*], Shiqian Jia [a], Qifan Zhang [a], Tianhao Zhang [a], Lianmeng Wu [a], Penghao Wang [a,b], Shufang Zhang [a,b], Huaming Wu [b,c], Jinlong Yu [a]

[a] School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China
[b] State Key Laboratory of Synthetic Biology, Tianjin University, Tianjin 300072, China
[c] Center for Applied Mathematics, KL-AAGDM, Tianjin University, Tianjin 300072, China

ARTICLE INFO

Keywords:
DNA-based image storage
Deep visual feature representation
Error-resilient coding
Feature imputation
Image reconstruction
Biological constraints

ABSTRACT

DNA data storage is emerging as a promising technology for the long-term preservation of large-scale data, and images are among the most common and data-intensive modalities. However, existing DNA-based image storage methods remain limited in terms of storage density, reconstruction quality, and the joint satisfaction of multiple biological constraints. To address these issues, we propose DNA-FRSC, a robust DNA-based image storage method integrating deep visual feature representations with error-resilient sequence coding. An encoder from a Swin Transformer-based deep joint source-channel coding model first extracts compact latent feature representations, which are then linearly quantized and mapped to DNA sequences to improve the storage density. A constrained codebook generator produces DNA sequences with both biological feasibility and error tolerance. During decoding, a two-stage error-correction module combines a top-*k* dynamic programming list decoding strategy with an equal-distance candidate refinement scheme to correct insertion, deletion, and substitution errors, and a training-free feature imputation method compensates for missing feature blocks caused by sequence losses. The experimental results confirm that the generated sequences satisfy biological constraints. Extensive tests on the Kodak and General-100 datasets at a 5 % base error rate show that DNA-FRSC outperforms existing methods, reducing the MSE by more than 82.7 %, improving the PSNR by at least 8.23 dB, and increasing the SSIM and MS-SSIM by more than 83.7 % and 50.4 %, respectively. These results indicate that DNA-FRSC provides a low-cost, highly robust, and biologically feasible solution for large-scale DNA-based image storage.

1. Introduction

In the digital era, the global data volume is increasing exponentially. Existing storage technologies, such as magnetic and optical media, are increasingly challenged by the rapidly growing demand for long-term data storage [1]. Compared with conventional storage technologies, DNA as a storage medium offers high storage density and long-term stability without requiring a continuous power supply. In theory, DNA data storage systems can reach a storage density of up to 455 EB/g [2] and preserve information for thousands of years under suitable conditions [3]. Owing to these unique advantages, DNA has emerged as a promising medium for next-generation data storage, especially for large-scale data management [4]. However, DNA data storage systems rely on biochemical processes such as synthesis, amplification, and sequencing [5–7], which inevitably introduce base insertion, deletion, and substitution (IDS) errors [8–10]. Therefore, a DNA storage system can usually be regarded as a quaternary sequence channel over the alphabet {A, T, G, C} [11]. In large-scale chip synthesis and high-throughput sequencing scenarios, base errors substantially affect reliable data recovery. Therefore, encoding and decoding methods with error-correction capabilities play key roles in DNA data storage [12,13]. Classic error- correcting codes, including the low-density parity check, Levenshtein-Marker (LM), and Reed-Solomon (RS) codes [14–16], have been used to increase data recovery reliability in DNA data storage.

As important carriers of information, images exhibit inherent perceptual tolerance to small local distortions. The human visual system can tolerate slight noise and local pixel variations; thus, a moderate number of errors does not usually significantly affect the perception of image content. Furthermore, with advances in image processing technologies, image denoising and the restoration of local distortions are

* Corresponding author.
E-mail address: tjujession@tju.edu.cn (S. Jia).

https://doi.org/10.1016/j.aei.2026.105296
Received 12 July 2026; Received in revised form 24 August 2026; Accepted 14 September 2026
Available online 25 September 2026
1474-0346/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

possible [17,18]. Therefore, considering the error patterns of DNA storage, images and DNA data storage show a high degree of compatibility. Storing images in DNA not only has important research significance but also offers a new technical approach for the long-term preservation of large-scale cold data, such as medical images, remote sensing images, and multimedia archives [19]. Motivated by these properties, recent studies have explored various DNA-based image storage methods, aiming to balance reconstruction quality, storage cost, and biological constraint satisfaction.

The IMG-DNA method [20] stores JPEG-based images by separating image coefficients and adding barriers to reduce error propagation. The HL-DNA method [21] uses a hybrid lossy and lossless coding scheme. The influence of insertions and deletions (indels) is mitigated through a rotational coding scheme and a barrier design, and the SSIM is reported to reach 0.55 at a 0.5% error rate. Furthermore, the DNA-base128 method [22] converts image pixels into DNA sequences through probability statistics and a base128 coding table and performs internal error correction through threshold judgment and drift comparison schemes. At a 1% error rate, the reconstructed images have a PSNR of 32.43 dB and an SSIM of 0.34. Additionally, the PELMI method [23] adopts parity encoding and local mean iteration schemes to process image data under base errors. At a 1% error rate, its SSIM and PSNR are approximately 0.79 and 34 dB, respectively. Moreover, the MMBP-based DNA information storage system [24] encoded four Southeast University images using DNA Fountain codes with RS coding and achieved complete image recovery through a microfluidic magnetic bead polymerase chain reaction scheme. These methods reduce the impact of DNA errors at the pixel or coefficient level. However, their reliance on handcrafted coding rules and local recovery strategies limits their generalizability across diverse image scenarios and their robustness under high base error rates and high sequence loss rates.

The DP-ID [25] and DNA-CTMF [26] methods further transform the influence of DNA base errors into salt-and-pepper noise in reconstructed images. The DP-ID method combines dynamic programming compression with interleaving to map compressed pixel bitstreams to DNA sequences, while the DNA-CTMF method maps pixels to DNA sequences through a pixel-base codebook and adjusts base groups shifted by indels before performing median filtering. At a 1% error rate, the DNA-CTMF method achieved a PSNR above 30 dB and an SSIM above 0.93. Although these methods can reduce local discrete noise, owing to their reliance on image postprocessing, they have a limited ability to recover information lost to burst errors or large-scale sequence losses and tend to oversmooth image details.

Since data writing in DNA storage relies on sequence synthesis, its cost remains an important factor limiting the large-scale application of DNA storage methods. At an estimated array synthesis cost of approximately 10^{-4} USD per base, writing one terabyte of data to DNA may cost approximately 800 million USD, which is 7–8 orders of magnitude greater than the cost of tape storage [27]. One estimate suggests that writing 1 MB of encoded data to DNA costs approximately 800–5000 USD, whereas tape storage costs approximately 16 USD per TB of data [28]. Therefore, reducing the number of bases to be synthesized is critical for lowering the cost of DNA-based image storage. Existing pixel-level and coefficient-level DNA-based image storage methods make limited use of compact representations, requiring a large number of synthesized bases and leading to high synthesis costs for large-scale cold data preservation.

To reduce the amount of data written to DNA, conventional image compression methods have been introduced into DNA-based image storage. For example, the JPEG-DNA method [29] adapts the JPEG workflow to DNA by introducing a quaternary variable-length coding scheme with biological constraints embedded into the JPEG encoder. Additionally, a wavelet-based method decomposes images into subbands, applies uniform scalar quantization to each subband, and uses nucleotide allocation to control the compression ratio and DNA sequence length [30]. Another method combines the discrete wavelet transform with vector quantization and optimizes the codebook size and vector dimension of each subband through nucleotide allocation [31]. The HELIX method [19] introduces a compression strategy by dividing images into fixed-size blocks, transforming RGB components into YUV space, applying discrete cosine transform (DCT), and quantizing transformed coefficients. It further exploits repeated image regions through matrix-based quantization and duplicate removal to reduce redundant information and the number of DNA sequences. Despite these compression strategies reducing the synthesis volume, the resulting compressed representations remain highly sensitive to the coding structure. A base error at a bitstream boundary or codeword can propagate and amplify into large-scale image distortion.

In addition, deep learning-based compression methods have been used to encode images into compact latent representations. The generative DNA method [32] employs a variational autoencoder (VAE)-based generative model to compress images into compact latent representations that are stored in DNA and reconstructs images through a generative decoder, which tolerates DNA storage errors and reduces both synthesis and sequencing costs. Moreover, the DJSCC-DNA method [33] uses a convolutional neural network-based deep joint source-channel coding framework, embeds the DNA channel into the network and adds biological constraints to the loss function to realize end-to-end optimization. The VAE-based DNA storage system [34] compresses images via a VAE and singular value decomposition and stores the latent blocks using a rotational coding scheme with an RS code. The DNA-QLC method [35] employs a quantized ResNet VAE (QRes-VAE) to compress images into entropy-coded bitstreams and applies a Levenshtein code (LC) for error correction. It achieved an SSIM of 0.917 at a 2% error rate. The INNSE method [36] uses an invertible neural network to downsample image data and encodes the resulting low-dimensional vectors with a constrained 6-nt codebook. At a 1% error rate, the PSNR of most of the reconstructed images remains above 25 dB. A common limitation, however, is that these methods focus mainly on representation learning and reconstruction, while feature block losses caused by sequence losses or unrecoverable base errors remain largely unaddressed. When such blocks cannot be recovered, the decoder receives an incomplete feature tensor, leading to a substantial degradation in reconstruction quality.

In the DNA-ELMR method [37], a Transformer autoencoder is used to construct a DNA-based image storage system, and IDS errors are introduced into the generated sequences during training to improve robustness to sequencing noise. The model jointly optimizes image reconstruction and multiple biological constraint losses to regulate sequence properties during encoding. However, DNA-ELMR is not primarily designed to minimize the amount of data written to DNA, resulting in relatively high synthesis cost, and its reconstruction quality may degrade at high base error rates. More broadly, existing methods generally address only subsets of biological constraints and rarely jointly consider GC balance, homopolymer length, secondary structure stability, error-prone motifs, and undesired motifs.

To address these limitations, this paper proposes DNA-FRSC, a DNA-based image storage method integrating deep visual feature representations with error-resilient sequence coding. DNA-FRSC employs the encoder and decoder of the Swin Transformer-based deep joint source-channel coding (SwinJSCC) model [38] for compact feature extraction and image reconstruction, respectively. The extracted features are quantized and encoded using a constrained codebook designed to balance biological constraint satisfaction and error distinguishability. During decoding, a two-stage sequence recovery procedure integrates top-k dynamic programming list decoding, cyclic redundancy check (CRC) verification, and equal-distance candidate refinement to recover reads corrupted by IDS errors. Any residual missing feature values associated with unrecovered data sequences are estimated through training-free feature imputation before image reconstruction. By coupling sequence-level error recovery with feature-level compensation, DNA-FRSC reduces the amount of data written to DNA and enables reconstruction quality to degrade gradually as the base error rate

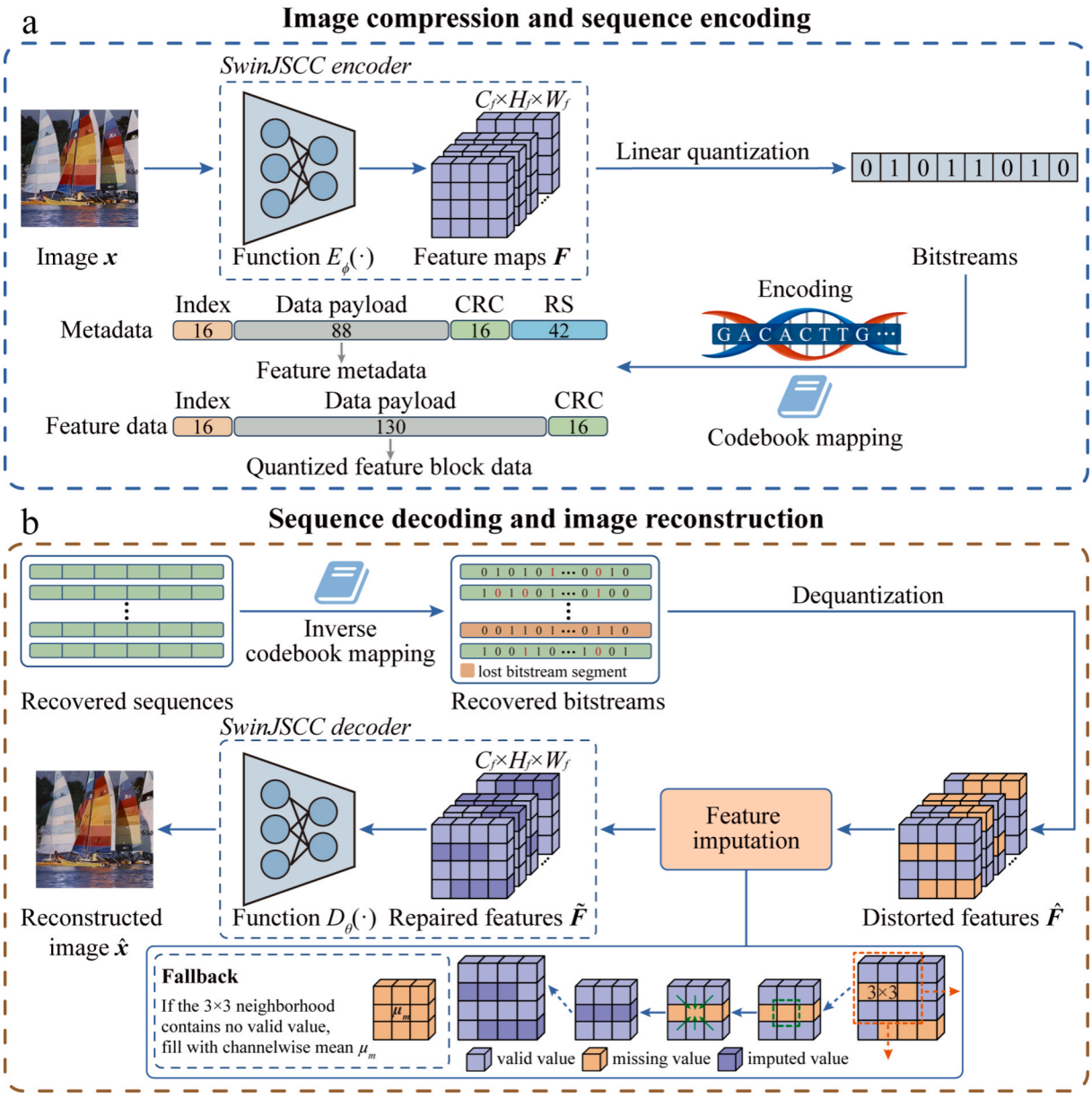


Fig. 1. Flowcharts of (a) the image compression and sequence encoding module and (b) the sequence decoding and image reconstruction module of DNA-FRSC.

increases. The main contributions of this study are summarized as follows:

1) We propose a coding method based on deep feature representations for low-cost DNA-based image storage. DNA-FRSC uses a SwinJSCC encoder to extract compact latent feature representations and converts them into encodable data through a linear quantization scheme. This design substantially reduces the amount of data written to DNA, thereby improving storage density and lowering the storage cost.
2) We propose a robust sequence recovery method to handle DNA reads with high error rates. A top-k dynamic programming (DP) list decoding method is constructed using a precomputed nearest-neighbor table. An equal-distance candidate refinement strategy is applied to resolve ambiguities among equidistant candidates. CRC verification is used to select reliable candidates. In addition, a training-free feature imputation method compensates for missing feature blocks caused by sequence losses and improves image reconstruction quality.
3) We propose a codebook generator for biologically constrained and error-resilient sequence encoding. The generated sequences satisfy key biological constraints in terms of GC balance, homopolymer control, secondary structure stability, and the suppression of error-prone and undesired motifs.

2. Methods

DNA-FRSC consists of three modules, as shown in Figs. 1 and 2. The first module involves image compression and sequence encoding, as shown in Fig. 1(a). This module serves as the front end of the DNA-FRSC system. It encodes the input image into a compact latent feature representation and maps the resulting features to biologically constrained DNA sequences for synthesis and storage in DNA. The second module is the error-correction module, as shown in Fig. 2. This module performs robust sequence recovery from DNA reads containing IDS errors and provides reliable recovered sequences for the subsequent feature reconstruction process. The third module involves sequence decoding and image reconstruction, as shown in Fig. 1(b). This module serves as

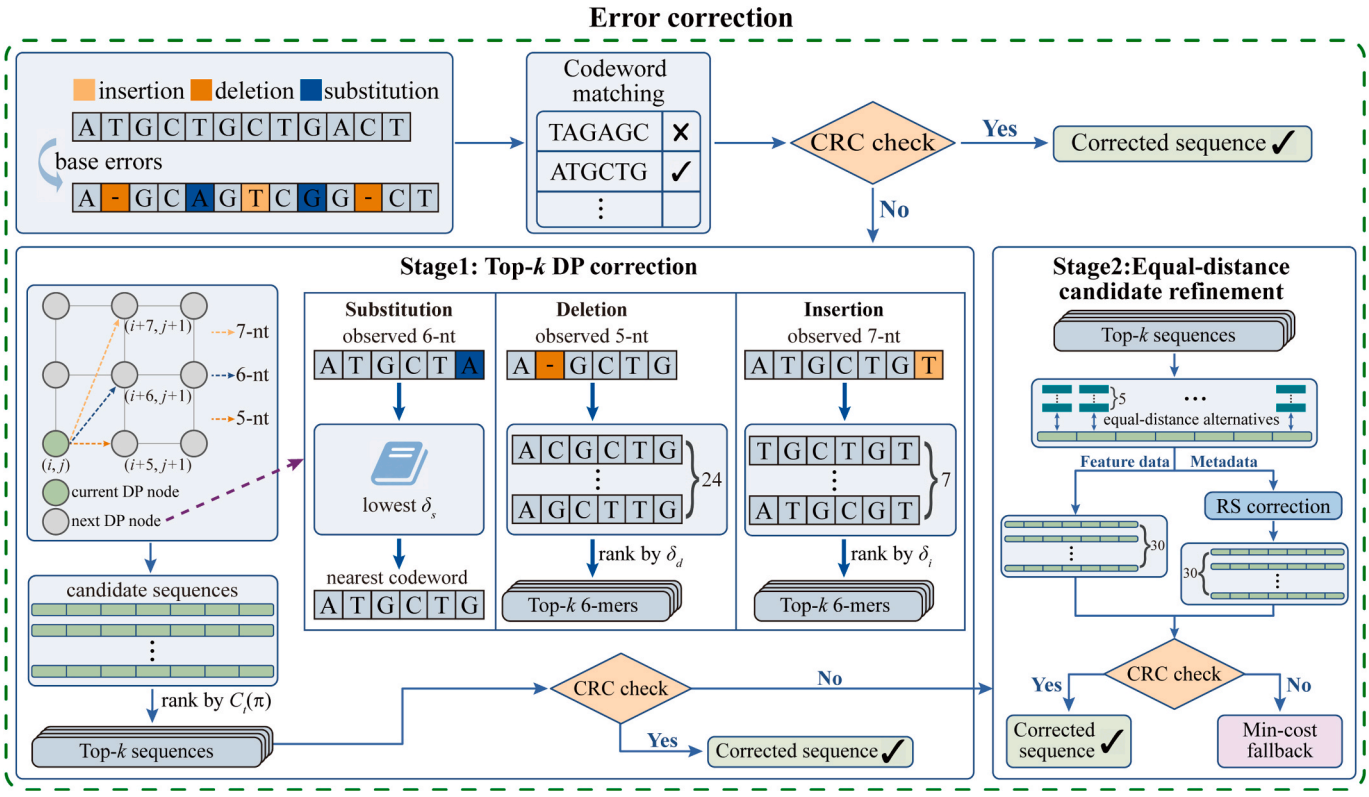


Fig. 2. Flowchart of the error correction module of DNA-FRSC.

the back end of the DNA-FRSC system. It decodes the recovered sequences to generate feature representations, imputes missing feature blocks, and reconstructs the final image using the SwinJSCC decoder.

2.1. SwinJSCC-based visual feature representation

SwinJSCC is a deep joint source-channel coding model constructed using a hierarchical Swin Transformer backbone [38]. Its four-stage encoder first divides the input image into non-overlapping 2 × 2 patches and converts them into patch embeddings. The four stages contain 2, 2, 6, and 2 Swin Transformer blocks, with embedding dimensions of 128, 192, 256, and 320, respectively. Between successive stages, patch merging halves both spatial dimensions while increasing the embedding dimension. The resulting spatial resolutions of the four-stage outputs are $\frac{H}{2} \times \frac{W}{2}$, $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, and $\frac{H}{16} \times \frac{W}{16}$, respectively. Within each stage, successive blocks alternate between window-based and shifted-window self-attention, capturing local structures and facilitating information exchange between neighboring windows. The decoder mirrors this hierarchy, using Swin Transformer blocks to refine the feature representations and patch division operations to progressively restore the spatial resolution.

Given an input image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$, where H and W denote the height and width of the image, respectively, the encoder $E_\phi(\cdot)$ with parameters ϕ maps the image to a compact continuous latent representation:

$$\boldsymbol{y} = E_\phi(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{y}$ denotes the latent representation. In DNA-FRSC, this latent representation corresponds to the feature tensor $\boldsymbol{F}$ defined in Section 2.2. In the original communication setting, $\boldsymbol{y}$ is power-normalized and transmitted over a noisy channel. For the additive white Gaussian noise channel, the received latent representation is expressed as follows:

$$\hat{\boldsymbol{y}} = \boldsymbol{y} + \boldsymbol{n}, \boldsymbol{n} \sim \mathcal{N}(0, \sigma_n^2 \boldsymbol{I}), \tag{2}$$

where $\hat{\boldsymbol{y}}$ denotes the received noisy latent representation, $\boldsymbol{n}$ denotes the Gaussian noise vector, σ_n^2 is the noise power determined by the channel signal-to-noise ratio, and $\boldsymbol{I}$ denotes the identity matrix. The decoder $D_\theta(\cdot)$ with parameters θ then reconstructs the image from the perturbed latent representation as $\hat{\boldsymbol{x}} = D_\theta(\hat{\boldsymbol{y}})$. The encoder and decoder are jointly optimized in an end-to-end manner, with the channel noise retained in the training loop. The overall loss function is defined as the expected reconstruction distortion:

$$(\phi^*, \theta^*) = \underset{\phi,\theta}{\operatorname{argmin}} \mathbb{E}_{\boldsymbol{x}\sim p_x} \mathbb{E}_{\hat{\boldsymbol{y}}\sim p_{\hat{y}|x}} [d(\boldsymbol{x}, \hat{\boldsymbol{x}})], \tag{3}$$

where ϕ^* and θ^* are the optimized encoder and decoder parameters, respectively; p_x denotes the source image distribution; $p_{\hat{y}|x}$ is the conditional distribution of the noisy latent representation given the input image; and $d(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2$ is the mean squared error loss. The released checkpoint of the SwinJSCC w/o SA&RA model, trained on randomly cropped 256 × 256 patches from DIV2K using Adam with an initial learning rate of 1×10^{-4} and a batch size of 16, was adopted. DNA-FRSC incorporates the SwinJSCC encoder and decoder while excluding the wireless transmission channel. The four encoder stages use 4, 6, 8, and 10 attention heads, respectively, with an attention window size of 8 and a multilayer perceptron ratio of 4. The decoder uses the corresponding configuration in reverse order.

The DNA storage process can be regarded as a noisy channel in which errors may corrupt feature values or cause them to be missing. Channel perturbations introduced during SwinJSCC training enable the model to learn noise-tolerant latent representations and tolerate residual feature errors and deviations introduced by feature imputation. Experiments in Section 3 further show that DNA-FRSC reconstruction quality degrades gradually, rather than abruptly, as the base error rate increases. The adopted nonadaptive encoder-decoder architecture contains 18.34 million parameters. The other components of DNA-FRSC introduce no additional learnable parameters. Section 3.4 analyzes the theoretical encoding and decoding complexities of DNA-FRSC.

2.2. Image compression and sequence encoding

This module encodes an input image and maps it to DNA sequences that satisfy biological constraints. As shown in Fig. 1(a), the input image is first encoded into a compact feature tensor by the SwinJSCC encoder. The feature values are then linearly quantized into a binary stream, which is further divided into fixed-length bitstream blocks. Next, a constrained 6-nt codebook maps the bitstreams to DNA sequences. The generated sequences include feature metadata and quantized feature blocks, which support subsequent feature recovery and image reconstruction.

Given an input image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$, the SwinJSCC encoder maps it to a compact feature tensor:

$$\boldsymbol{F} = E_{\phi}(\boldsymbol{x}),\ \boldsymbol{F} \in \mathbb{R}^{C_f\times H_f\times W_f}, \tag{4}$$

where $\boldsymbol{F}$ denotes the extracted feature tensor, C_f is the number of feature channels, and H_f and W_f denote the height and width of the feature maps, respectively. The overall spatial downsampling factor of the SwinJSCC encoder is 16. Based on the evaluation in Section 3.1, C_f is set to 96. Therefore, $H_f = \frac{H}{16}$ and $W_f = \frac{W}{16}$, and the resulting feature tensor has dimensions of $96 \times \frac{H}{16} \times \frac{W}{16}$.

After feature extraction, the continuous feature tensor is discretized using 8-bit linear quantization before DNA sequence encoding is performed. The quantized value is calculated as follows:

$$q_i = \lfloor 255\cdot\frac{f_i - f_{\min}}{f_{\max} - f_{\min}} + \frac{1}{2}\rfloor, q_i \in \{0, 1, ..., 255\}, \tag{5}$$

where f_i is the i-th element of $\boldsymbol{F}$, $f_{\min}$ and $f_{\max}$ are the minimum and maximum values of $\boldsymbol{F}$, respectively, and q_i is the corresponding 8-bit quantized value. The quantized feature tensor is denoted as $\boldsymbol{Q} \in \{0,\ 1,\ ...,\ 255\}^{C_f\times H_f\times W_f}$. Since each element of $\boldsymbol{Q}$ is represented by 8 bits, the quantized tensor is serialized into a feature bitstream $\boldsymbol{B} \in \{0,\ 1\}^{8C_fH_fW_f}$. The feature bitstream is further divided into fixed-length bitstream blocks, each of which is used as the data payload of a DNA data sequence.

To map the quantized feature bitstream to DNA sequences, an error-resilient constrained codebook generator is designed for DNA-FRSC. Several sequence patterns should be avoided during DNA sequence design. Long single-nucleotide repeats may lead to a higher error rate during synthesis or sequencing [39]. Because of complementary base pairing, sequences containing reverse-complementary regions may fold into stable secondary structures such as hairpins, which can hinder synthesis and sequencing [40]. DNA sequences with excessively low or high GC contents are often difficult to synthesize, and a GC content between 40% and 60% is generally considered suitable for DNA storage sequences [41]. Therefore, the codebook generator considers four types of constraints: the homopolymer length, reverse-complementary pairs, the GC content balance, and the codeword distance.

The generator first enumerates all possible L-nt codewords and filters them according to single-codeword biochemical constraints. Each retained codeword has a GC content of 50%, corresponding to exactly $L/2$ G or C nucleotides, and a maximum internal homopolymer length of $h_{\max} = 3$. The resulting candidate pool is denoted by Φ. A randomized greedy procedure then constructs an initial codebook of K codewords by sequentially selecting feasible candidates subject to a minimum Hamming distance of $d_{\min} = 3$, a maximum homopolymer length of 3 nucleotides across codeword boundaries, and the exclusion of reverse-complementary pairs such as ACGAAC and GTTCGT. The Hamming distance between two codewords c_i and c_j is defined as follows:

$$d_H(c_i, c_j) = \sum_{l=1}^{L} \mathbb{I}(c_{i,l} \neq c_{j,l}), \tag{6}$$

where $c_{i,l}$ and $c_{j,l}$ denote the bases at the l-th position of c_i and c_j, respectively, L is the codeword length, and $\mathbb{I}(\cdot)$ is the indicator function.

Because sequential selection is path-dependent, the greedy procedure may produce a feasible but suboptimal codebook. The codebook is refined using simulated annealing (SA) [42], which reduces insertion- and deletion-induced ambiguity while preserving the codebook size and predefined constraints. To quantify such ambiguity, the generator constructs the deletion and insertion neighborhoods of each candidate codeword. For a codeword $c \in \Sigma^L$ with $\Sigma = \{A, C, G, T\}$, the deletion neighborhood $\mathcal{N}_{\text{del}}(c)$ is defined as the set of all $(L-1)$-mers obtained by deleting one base from c, and the insertion neighborhood $\mathcal{N}_{\text{ins}}(c)$ is defined as the set of all $(L+1)$-mers obtained by inserting one base into c. For a candidate codebook $C = \{c_1, ..., c_K\}$ containing K codewords, the deletion neighborhood statistics are expressed as follows:

$$A_{ij}^{\text{del}} = \left|\mathcal{N}_{\text{del}}(c_i) \cap \mathcal{N}_{\text{del}}(c_j)\right|, \tag{7}$$

$$O_{\text{del}}(C) = \sum_{1\leqslant i<j\leqslant K} A_{ij}^{\text{del}}, \tag{8}$$

$$P_{\text{del}}(C) = \sum_{1\leqslant i<j\leqslant K} \mathbb{I}\left(\mathcal{N}_{\text{del}}(c_i) \cap \mathcal{N}_{\text{del}}(c_j) \neq \varnothing\right). \tag{9}$$

The corresponding insertion neighborhood statistics are defined as follows:

$$A_{ij}^{\text{ins}} = \left|\mathcal{N}_{\text{ins}}(c_i) \cap \mathcal{N}_{\text{ins}}(c_j)\right|, \tag{10}$$

$$O_{\text{ins}}(C) = \sum_{1\leqslant i<j\leqslant K} A_{ij}^{\text{ins}}, \tag{11}$$

$$P_{\text{ins}}(C) = \sum_{1\leqslant i<j\leqslant K} \mathbb{I}\left(\mathcal{N}_{\text{ins}}(c_i) \cap \mathcal{N}_{\text{ins}}(c_j) \neq \varnothing\right), \tag{12}$$

where K is set to 64, L is set to 6, and $\boldsymbol{A}^{\text{del}}$ and $\boldsymbol{A}^{\text{ins}}$ denote the precomputed pairwise overlap matrices for the deletion and insertion neighborhoods, respectively, with the elements A_{ij}^{del} and A_{ij}^{ins} representing the corresponding neighborhood overlap sizes between codewords c_i and c_j. $O_{\text{del}}(C)$ and $O_{\text{ins}}(C)$ represent the total deletion-neighborhood and insertion-neighborhood overlap of the codebook, respectively, while $P_{\text{del}}(C)$ and $P_{\text{ins}}(C)$ denote the number of codeword pairs with nonempty deletion and insertion neighborhood intersections, respectively. The fitness function of the SA algorithm is expressed as follows:

$$F(C) = w_1O_{\text{del}}(C) + w_2P_{\text{del}}(C) + w_3O_{\text{ins}}(C) + w_4P_{\text{ins}}(C), \tag{13}$$

where w_1, w_2, w_3, and w_4 are the corresponding weights and are set to 100, 10, 30, and 3, respectively. The SA algorithm is employed to minimize $F(C)$ and search for the final codebook. The parameters N_r, N_s, T_0, and α denote the number of restarts, the number of iterations per restart, the initial temperature, and the cooling factor, respectively, and are set to 20, 1000, 4, and 0.9992, respectively, in this work. After all the restarts are completed, the codebook C^* that achieves the lowest fitness value is retained for subsequent DNA sequence encoding. The overall procedure is summarized in Algorithm 1.

Algorithm 1 Error-resilient constrained codebook generation.

Require: Alphabet $\Sigma = \{A, C, G, T\}$, codeword length L, codebook size K, homopolymer bound $h_{\max}$, minimum Hamming distance $d_{\min}$, weights w_1–w_4, SA parameters N_r, N_s, T_0, α
Ensure: Optimized codebook $C^* = \{c_1, ..., c_K\}$
$\Phi \leftarrow \{\ c \in \Sigma^L$: GC $(c) = L/2$, MaxHomo$(c) \leq h_{\max}\ \}$
Precompute d_H, the reverse-complementary pairs, the boundary homopolymer lengths, and the overlap matrices $\boldsymbol{A}^{\text{del}}$ and $\boldsymbol{A}^{\text{ins}}$
$F^* \leftarrow +\infty$
for $r = 1$ **to** N_r **do**
  $C \leftarrow$ GreedyInit(Φ, K)
  $F_{\text{cur}} \leftarrow F(C)$; $T \leftarrow T_0$; $C_{\text{best}} \leftarrow C$; $F_{\text{best}} \leftarrow F_{\text{cur}}$
  for $n = 1$ **to** N_s **do**

(continued on next page)

(*continued*)

Algorithm 1 Error-resilient constrained codebook generation.

```
        c_out ← RandomSelect(C)
        c_in ← FeasibleCandidate(Φ \ C, C \ {c_out})
        C' ← (C \ {c_out}) ∪ {c_in}
        ΔF ← F(C') − F_cur
        if ΔF ≤ 0 or rand(0, 1) < exp(−ΔF / T) then
          C ← C'; F_cur ← F_cur + ΔF
          if F_cur < F_best then C_best ← C; F_best ← F_cur
        end if
        T ← α · T
      end for
      if F_best < F* then C* ← C_best; F* ← F_best
    end for
    return C*
```

To enable robust storage and retrieval, DNA-FRSC generates two types of sequences: metadata sequences and data sequences. Each sequence consists of an index component, a payload component and a CRC component. The index component is set to 16 bits and is used to identify the sequence position, whereas the CRC component is set to 16 bits for sequence verification. For metadata sequences, the payload component is 88 bits long and records the feature shape, the image size, and the quantization range. The resulting 120-bit information component is protected by an RS(27,20) code over GF(2^6). For data sequences, the payload component is 130 bits long and stores a fixed segment of the quantized feature bitstream. The resulting codebook comprises 64 distinct 6-nt codewords, providing a one-to-one mapping for all 64 values representable by a 6-bit symbol. Each encoded sequence contains 27 such symbols, yielding a final length of 162 nt. The encoding overhead (the ratio of non-payload bits to payload bits) is 24.62%. As synthesis errors become more difficult to control with increasing sequence length, oligonucleotide synthesis is generally limited to approximately 200 nt or less [1,43]. Thus, the resulting 162-nt sequences remain within this practical synthesis range.

2.3. Sequence decoding and image reconstruction

Following error correction, each recovered sequence is first partitioned into consecutive 6-nt codewords. Each codeword is mapped back to its corresponding 6-bit symbol using the inverse codebook, and the binary representations of the recovered symbols are concatenated in sequence order to obtain the binary representation of the sequence. Based on the sequence format defined in Section 2.2, the sequence index and feature payload are subsequently extracted. The recovered feature payloads are reordered according to their indices and concatenated to reconstruct the quantized feature bitstream. The reconstructed quantized feature bitstream $\widehat{\boldsymbol{B}}$ is first grouped into 8-bit unsigned integers:

$$\widehat{Q}_{m,h,w} = \sum_{t=0}^{7} 2^{7-t}\widehat{B}_{8(mH_fW_f+hW_f+w)+t}, \tag{14}$$

where $\widehat{\boldsymbol{Q}} \in \{0, 1, \ldots, 255\}^{C_f\times H_f\times W_f}$ denotes the quantized feature tensor. The variables m, h, and w denote the channel index, row index, and column index of the feature tensor, respectively, where $m \in \{0, \ldots, C_f-1\}$, $h \in \{0, \ldots, H_f-1\}$, and $w \in \{0, \ldots, W_f-1\}$. The distorted feature tensor is recovered by inverse quantization as follows:

$$\widehat{F}_{m,h,w} = f_{\min} + \frac{\widehat{Q}_{m,h,w}}{255}(f_{\max} - f_{\min}), \tag{15}$$

where $\widehat{F}_{m,h,w}$ denotes the dequantized feature value at the m-th channel and spatial position (h, w) and $\widehat{\boldsymbol{F}} \in \mathbb{R}^{C_f\times H_f\times W_f}$ denotes the distorted feature tensor.

At high base error rates, sequence loss may occur. Sequence loss is defined as the failure to recover any sequence corresponding to a given expected index after error correction. Because each indexed data sequence carries a fixed segment of the quantized feature bitstream, sequence loss renders the corresponding segment unavailable. Consequently, the associated feature values cannot be restored, resulting in contiguous missing regions in the distorted feature tensor, which may produce spatially structured artifacts in the reconstructed image. To address this issue, a training-free feature imputation module is introduced in DNA-FRSC. A binary mask $\boldsymbol{M} \in \{0, 1\}^{C_f\times H_f\times W_f}$ is constructed according to the recovery status of the sequence indices. Specifically, $M_{m,h,w} = 1$ indicates that the feature value at (m, h, w) is valid, whereas $M_{m,h,w} = 0$ indicates that it is missing. On the basis of the mask, the repaired feature tensor is obtained as follows:

$$\widetilde{F}_{m,h,w} = \begin{cases} \widehat{F}_{m,h,w}, & M_{m,h,w} = 1 \\ \dfrac{\sum_{(u,v)\in\Omega_{h,w}} M_{m,u,v}\widehat{F}_{m,u,v}}{\sum_{(u,v)\in\Omega_{h,w}} M_{m,u,v}}, & M_{m,h,w} = 0, \sum_{(u,v)\in\Omega_{h,w}} M_{m,u,v} > 0, \\ \mu_{m,} & M_{m,h,w} = 0, \sum_{(u,v)\in\Omega_{h,w}} M_{m,u,v} = 0 \end{cases} \tag{16}$$

$$\Omega_{h,w} = \{(u,v) \mid |u-h| \leqslant 1,\ |v-w| \leqslant 1,\ 0 \leqslant u < H_f,\ 0 \leqslant v < W_f\}, \tag{17}$$

$$\mu_m = \frac{\sum_{h=0}^{H_f-1}\sum_{w=0}^{W_f-1} M_{m,h,w}\widehat{F}_{m,h,w}}{\sum_{h=0}^{H_f-1}\sum_{w=0}^{W_f-1} M_{m,h,w}}, \tag{18}$$

where $\widetilde{F}_{m,h,w}$ denotes the repaired feature value at the m-th channel and spatial position (h, w), $\widetilde{\boldsymbol{F}} \in \mathbb{R}^{C_f\times H_f\times W_f}$ denotes the repaired feature tensor, u and v denote the spatial indices of neighboring feature values, $\Omega_{h,w}$ denotes the 3×3 spatial neighborhood centered at (h, w), and μ_m denotes the mean of all valid feature values in the m-th channel.

This module preserves valid feature values and fills missing positions using the local spatial mean within a 3×3 neighborhood. If no valid value exists in the local neighborhood, the channelwise mean is used. The combination of local and channelwise averaging strategies exploits the spatial correlations among neighboring feature values to estimate missing features without introducing additional trainable parameters or requiring extra training. The resulting smooth estimates provide locally consistent replacements for contiguous missing regions caused by sequence loss. The ablation results in Section 3.2 show that feature imputation helps DNA-FRSC preserve recognizable image structures and colors. After feature imputation, the repaired feature tensor is fed into the SwinJSCC decoder to reconstruct the image:

$$\widehat{\boldsymbol{x}} = D_\theta\left(\widetilde{\boldsymbol{F}}\right),\ \widehat{\boldsymbol{x}} \in \mathbb{R}^{3\times H\times W}, \tag{19}$$

where $\widetilde{\boldsymbol{F}}$ denotes the repaired feature tensor and $\widehat{\boldsymbol{x}}$ denotes the final reconstructed image.

2.4. Error correction

DNA synthesis and sequencing are major sources of sequence errors during the writing and reading stages of DNA data storage, respectively [4]. During writing, oligonucleotide synthesis converts encoded target sequences into physical DNA molecules, and incomplete reactions or side reactions can introduce IDS errors [5]. During reading, sequencing instruments infer nucleotide identities from physical or chemical signals, and signal noise or base-calling inaccuracies can introduce IDS errors [7]. Although error rates and spectra vary across synthesis and sequencing technologies, errors within individual sequences primarily manifest as indels and substitutions. Insertions add extra nucleotides, deletions remove existing nucleotides, and substitutions replace one nucleotide with another. Unlike substitutions, indels shift subsequent nucleotides relative to the original codeword boundaries and disrupt their positional correspondence with the target sequence. They are

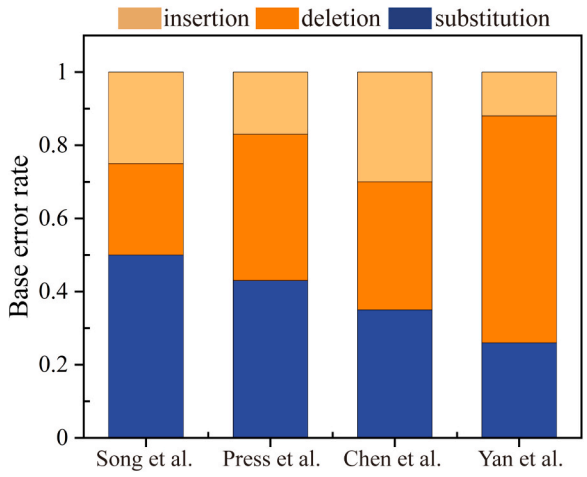


Fig. 3. The error distributions of Song et al. [44], Press et al. [45], Chen et al. [46], and Yan et al. [47].

therefore generally more difficult for decoders based on fixed-length codewords to correct.

To address these error characteristics, DNA-FRSC employs the two-stage error-correction mechanism shown in Fig. 2. Given an observed read $\boldsymbol{r}$, DNA-FRSC first divides the read into consecutive 6-nt segments and directly matches each segment with the codebook. If the recovered sequence passes the CRC check, it is accepted as a corrected sequence. Otherwise, the read is further processed by the two-stage correction module.

Stage 1: Top-k DP correction.

The DP state (i, j) represents the set of partial paths that have consumed the first i bases of $\boldsymbol{r}$ and recovered j symbols. At each state, three transition types are considered.

For an observed 6-nt segment $\boldsymbol{s} \in \Sigma^6$, the nearest codeword symbol is selected as follows:

$$a^*(\boldsymbol{s}) = \underset{a\in\mathscr{A}}{\operatorname{argmin}} d_H(\boldsymbol{s}, \boldsymbol{c}_a) \tag{20}$$

and the substitution cost is as follows:

$$\delta_s(\boldsymbol{s}) = w_s d_H\left(\boldsymbol{s}, \boldsymbol{c}_{a^*(\boldsymbol{s})}\right), \tag{21}$$

where $\mathscr{A} = \{0, 1, \ldots, 63\}$ is the symbol set, $\boldsymbol{c}_a \in \Sigma^6$ is the codeword of symbol a, and w_s is the substitution weight, which is set to 1.1. For an observed 5-nt segment $\boldsymbol{s}_5 \in \Sigma^5$, one base is inserted at each of the six positions and with each of the four bases, yielding up to 24 candidate 6-mers. The deletion recovery cost is as follows:

$$\delta_d(\boldsymbol{s}_5) = w_i + \min_{\boldsymbol{s}\in\mathscr{D}(\boldsymbol{s}_5)} \delta_s(\boldsymbol{s}), \tag{22}$$

where w_i denotes the penalty weight of the indels, which is set to 1.3, and $\mathscr{D}(\boldsymbol{s}_5)$ denotes the candidate set generated from $\boldsymbol{s}_5$. The candidate 6-mers in $\mathscr{D}(\boldsymbol{s}_5)$ are ranked by δ_d, and the top-k candidates with the lowest cost are retained. For an observed 7-nt segment $\boldsymbol{s}_7 \in \Sigma^7$, one base is deleted at each of the seven positions, yielding up to 7 candidate 6-mers. The insertion recovery cost is as follows:

$$\delta_i(\boldsymbol{s}_7) = w_i + \min_{\boldsymbol{s}\in\mathscr{I}(\boldsymbol{s}_7)} \delta_s(\boldsymbol{s}), \tag{23}$$

where $\mathscr{I}(\boldsymbol{s}_7)$ denotes the candidate set generated from $\boldsymbol{s}_7$. The candidate 6-mers in $\mathscr{I}(\boldsymbol{s}_7)$ are ranked by δ_i, and the top-k candidates with the lowest cost are retained.

The total cost of a decoding path π is defined as follows:

$$C_t(\pi) = \sum_{j=1}^{M} \delta_j, \tag{24}$$

where M denotes the number of codeword positions in each sequence, which is set to 27, and δ_j denotes the local cost at the j-th codeword position, which is computed using Eq. (21), Eq. (22), or Eq. (23) according to the transition type. At each state (i, j), only the k partial paths with the lowest cumulative cost are retained. The DP outputs the top-k complete paths ranked by $C_t(\pi)$, and each candidate is verified by the CRC scheme.

Stage 2: Equal-distance candidate refinement.

If no top-k candidate passes the CRC check, DNA-FRSC further applies an equal-distance candidate refinement module. This stage addresses the case in which an observed 6-mer shares the same minimum Hamming distance with multiple codewords. The equal-distance candidate set of $\boldsymbol{s}$ is defined as follows:

$$\mathscr{T}(\boldsymbol{s}) = \left\{ a \in \mathscr{A} | d_H(\boldsymbol{s}, \boldsymbol{c}_a) = \min_{b\in\mathscr{A}} d_H(\boldsymbol{s}, \boldsymbol{c}_b) \right\} \tag{25}$$

where $\mathscr{T}(\boldsymbol{s})$ denotes the set of symbols that share the same minimum Hamming distance to $\boldsymbol{s}$. To control the search complexity, a bounded beam search is adopted in Stage 2 for each DP alignment from Stage 1 that fails the CRC check. At each codeword position, each retained partial candidate is expanded using at most 5 equal-distance alternatives. The expanded candidates are then pruned to retain at most 30 partial candidate sequences in the beam. After all positions are traversed, at most 30 refined candidate sequences are verified by the CRC scheme for each DP alignment. A refined candidate that passes the CRC check is accepted as the corrected sequence. If none of the refined candidates passes the CRC check, DNA-FRSC does not discard the read directly. Instead, the candidate sequence with the lowest $C_t(\pi)$ is retained as the fallback result. This process is defined as the minimum-cost fallback strategy.

2.5. Reconstructed image quality metrics

To comprehensively evaluate the quality of the reconstructed images, six full-reference image quality metrics are used, namely, the mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), multiscale structural similarity index measure (MS-SSIM), universal quality index (UQI), and feature similarity index measure (FSIM). The MSE is defined as the average squared difference between the original and reconstructed images and directly reflects the pixel-level reconstruction error. The PSNR is derived from the MSE on a logarithmic scale and is used to quantify the signal fidelity of the reconstructed images. Compared with error metrics computed at the pixel level, the SSIM evaluates image quality on the basis of three components, namely, luminance, contrast, and structure, which makes it more consistent with human visual perception. The MS-SSIM extends the SSIM by measuring structural similarity at multiple scales, allowing fine details and global structures to be evaluated simultaneously. The UQI decomposes image distortion into correlation, luminance, and contrast components and captures the overall similarity between the original and reconstructed images. The FSIM evaluates perceptual similarity using phase congruency and the gradient magnitude, making it sensitive to structural and edge information. Lower MSE values indicate higher reconstruction fidelity, whereas higher values of the other five metrics indicate better quality.

3. Results

3.1. Experimental setup and sequence encoding performance

To comprehensively evaluate DNA-FRSC, simulations were conducted at base error rates ranging from 1% to 5% using a base-wise independent and identically distributed error model. For each nucleotide, an error indicator was independently sampled from a Bernoulli distribution with the corresponding base error probability. Conditional

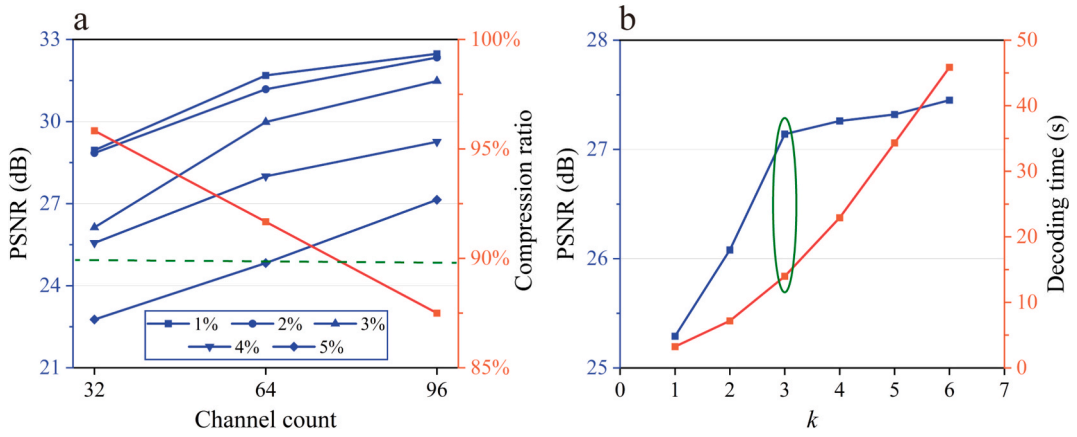


Fig. 4. (a) Average reconstruction performance with respect to the number of feature channels. (b) Average reconstruction PSNR and decoding time with respect to the top-k value at a 5% base error rate.

Table 1
Comparison of the general performance of different DNA-based image storage methods.

Method	Image compression	Image data representation	Error type	Single-read decoding	Sequence length excluding primer (nt)	Image data payload (bits)	Net information density (bits/nt)
DNA-base128 [22]	No	Pixel data	IDS	Yes	160	255	1.59
PELMI [23]	No	Pixel data	IDS and sequence loss	Yes	177	256	1.45
DNA-CTMF [26]	No	Pixel data	IDS	Yes	160	224	1.50
DNA-ELMR [37]	No	Latent representations	IDS	Yes	240	480	2.00
DP-ID [25]	Yes	Pixel data	IDS	Yes	160	304	2.15
JPEG-DNA [29]	Yes	DCT coefficients	Deletion	Yes	200–300	–	–
INNSE [36]	Yes	Pixel data	IDS	Yes	175	128	11.69
DNA-QLC [35]	Yes	Latent representations	IDS	Yes	206	290	47.72
HELIX [19]	Yes	DCT coefficients	IDS and sequence loss	No	148	88	9.07
DNA-FRSC	Yes	Latent representations	IDS and sequence loss	Yes	162	130	6.41

on an error event, the error type was sampled from a categorical distribution according to the specified IDS ratio. A substitution replaced the original nucleotide with one of the other three nucleotides selected uniformly at random. A deletion removed the original nucleotide, whereas an insertion retained it and appended one nucleotide selected uniformly from {A, C, G, T}. Each combination of base error rate and IDS ratio was evaluated over 500 independent Monte Carlo runs. The plotted values represent the means, and the error bars indicate the corresponding 95% confidence intervals. As shown in Fig. 3, four IDS ratios were considered: 25%:25%:50%, 17%:40%:43%, 30%:35%:35%, and 12%:62%:26% [44–47].

We selected six BMP color images, each with a resolution of 256 × 256 pixels and a file size of 192 KB, from the USC-SIPI dataset as test samples. We evaluated the reconstruction performance of DNA-FRSC at different values of C_f. Fig. 4(a) shows the variation in the average PSNR of the six images with C_f at base error rates from 1% to 5%, where the IDS ratio is fixed at 17%:40%:43%. In this experiment, the top-k parameter was fixed at $k = 3$, and the selection of this value is further evaluated in Fig. 4(b). The results show that as C_f increases from 32 to 96, the PSNR increases monotonically at all base error rates. For example, at a 5% base error rate, the PSNR increases from 22.76 dB to 27.14 dB. These results indicate that more channels can retain richer feature information, thereby improving the quality of the reconstructed images. However, the improvement in the reconstruction quality comes at the cost of higher storage overhead, reflecting the trade-off between feature representation capability and storage efficiency.

In general, a higher PSNR indicates better image quality. Reconstructed images with a PSNR higher than 25 dB are usually considered to have acceptable visual quality [48]. Since image fidelity is critical for DNA-based image storage systems, $C_f = 96$ is selected as the default configuration of DNA-FRSC. Under this setting, the PSNR remains higher than 25 dB over the entire base error rate range from 1% to 5%, indicating that this configuration satisfies the basic reconstruction quality requirement for practical image storage. The influence of the top-k parameter is further evaluated in Fig. 4(b). The experiments were conducted at a 5% base error rate with an IDS ratio of 17%:40%:43%. The reported PSNR and decoding time are averaged over the six images. As k increases from 1 to 3, the average PSNR increases from 25.29 dB to 27.14 dB, while the average decoding time increases from 3.25 s to 13.98 s. When k is further increased from 3 to 6, the average PSNR improves by only 0.31 dB, whereas the average decoding time increases from 13.98 s to 45.86 s. These results indicate that retaining more than three candidate paths provides only limited gains in reconstruction quality but substantially increases the decoding cost. Therefore, $k = 3$ is selected as the default setting to achieve a favorable trade-off between reconstruction quality and decoding efficiency.

The general characteristics of the representative DNA-based image storage methods are summarized in Table 1. Image data representation refers to the form in which image information is expressed before DNA encoding. This classification is based solely on the underlying

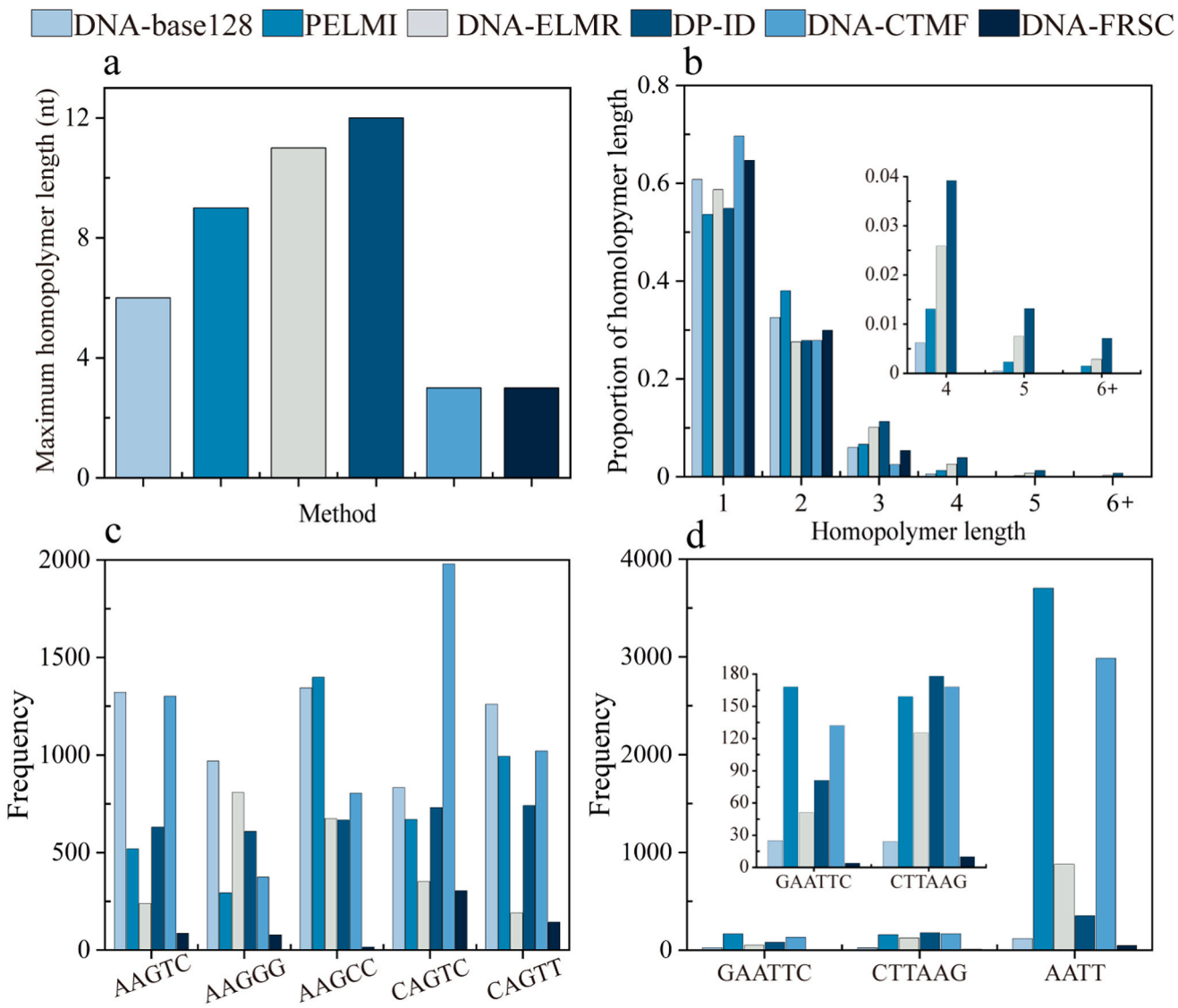


Fig. 5. Encoding results of different DNA-based image storage methods on the Airplane image. (a) Maximum homopolymer length. (b) Homopolymer length distribution. (c) Frequency of error-prone motifs. (d) Frequency of undesired motifs.

representation, irrespective of subsequent processing steps such as quantization, entropy coding, and error-correction coding. For consistency, net information density was evaluated for each method using the same 192 KB RGB image and calculated as the original image size in bits divided by the total number of encoded nucleotides. To evaluate the performance of DNA-FRSC, we compared it with four representative non-compression-based methods and one compression-based method. In addition, a broader comparison with compression-based DNA image storage methods is provided in Section 3.4. The Airplane image from the USC-SIPI dataset was used for the encoding analysis presented in this subsection. All simulations in this study were performed on a platform equipped with an Intel Core Ultra 9 275HX CPU, 64 GB RAM, and an NVIDIA GeForce RTX 5080 Laptop GPU with 16 GB VRAM.

The encoding results of the different methods are shown in Fig. 5. DNA-FRSC and DNA-CTMF produce the shortest maximum homopolymer runs of only 3 nt, whereas DP-ID and DNA-ELMR generate runs of up to 12 nt and 11 nt, respectively. Long homopolymer runs are known to increase synthesis and sequencing error rates. This advantage of DNA-FRSC is confirmed by the results shown in Fig. 5(b), where more than 95% of the homopolymer runs in DNA-FRSC sequences are no longer than 2 nt, and runs of length ≥ 4 are eliminated entirely. The occurrence of five representative error-prone motifs, namely, AAGTC, AAGGG, AAGCC, CAGTC, and CAGTT [37], is shown in Fig. 5(c). Compared with the best baseline for each motif, DNA-FRSC reduces the motif counts by 13.6% to 97.7%, with an average reduction of 54.7%. Specifically, DNA-FRSC reduces the counts of AAGTC, AAGGG, AAGCC, CAGTC, and CAGTT by 63.9%, 73.4%, 97.7%, 13.6%, and 24.7%, respectively. Moreover, undesired motifs are short DNA subsequences that may interfere with DNA synthesis, sequencing, and restriction-enzyme processing [49,50]. In this study, three representative motifs are counted: GAATTC, CTTAAG, and AATT. GAATTC and CTTAAG constitute the two complementary strands of the EcoRI recognition site, and AATT is the central tetranucleotide of the relaxed EcoRI* recognition site [51]. As shown in Fig. 5(d), DNA-FRSC achieves the lowest counts for all three motifs among the compared methods. Compared with the best baseline for each motif, DNA-FRSC reduces the counts of GAATTC, CTTAAG, and AATT by 84.0%, 58.3%, and 61.0%, respectively, with an average reduction of 67.8%. These results demonstrate that the constrained codebook generator effectively suppresses biologically unfavorable sequence patterns.

The GC content and minimum free energy (MFE) of the encoded DNA sequences are shown in Fig. 6. In Fig. 6(a), the GC content of each sequence generated by DNA-base128 or DNA-FRSC is fixed at 50%, whereas the other methods show varying deviations from this target.

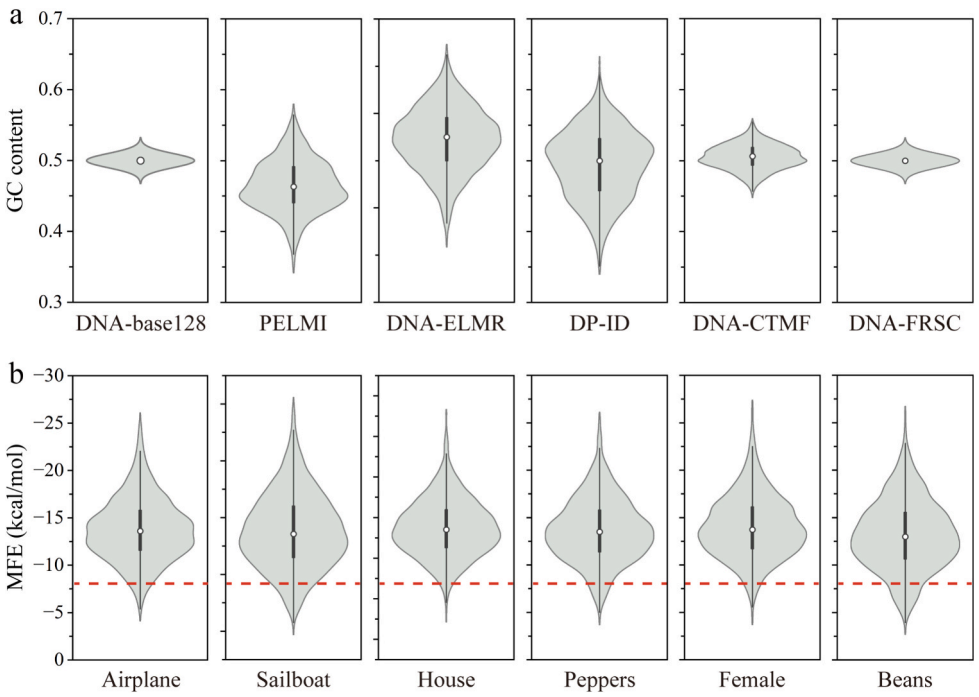


Fig. 6. (a) Guanine-cytosine content and (b) minimum free energy of DNA sequences generated by various DNA-based image storage methods. The white dot represents the median, and the thick centerlines represent the quartiles.

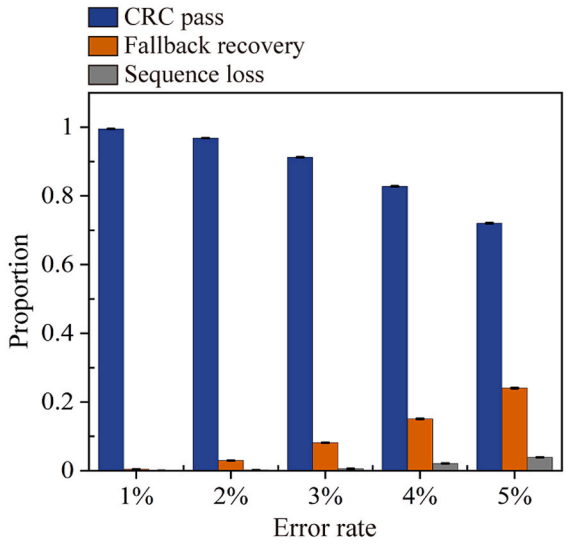


Fig. 7. Sequence decoding results of DNA-FRSC on the Sailboat image under different base error rates.

These results demonstrate that DNA-FRSC maintains strict GC balance during sequence encoding. The MFE is an important indicator of the tendency of a sequence to form stable secondary structures, as DNA sequences with overly stable secondary structures may form hairpins that interfere with DNA synthesis and sequencing. In this study, the MFE was calculated using ViennaRNA version 2.7.2 [52] to assess the folding potential of the encoded DNA sequences. In Fig. 6(b), the −8 kcal/mol reference represents a relatively stringent folding-potential criterion used in previous sequence design [53], whereas an MFE above −30 kcal/mol is recommended to avoid excessively stable secondary structures [52]. The MFEs of the sequences encoded from the six test images are distributed mainly around −13.81 ± 3.37 kcal/mol and remain well above –30 kcal/mol, indicating that the encoded sequences avoid excessively stable secondary structures.

3.2. *Error-resilient decoding and ablation experiments*

To analyze the sequence decoding behavior of DNA-FRSC under different base error rates, we measured the proportions of three sequence categories: sequences that passed the CRC check, sequences recovered by the minimum-cost fallback strategy, and lost sequences. The Sailboat image was evaluated under an IDS ratio of 17%:40%:43%, and the results are shown in Fig. 7. At a 1% base error rate, 99.5% of the sequences pass the CRC check, 0.4% are recovered via the fallback strategy, and essentially no sequences are lost. As the base error rate increases to 2% and 3%, the CRC pass ratio decreases to 96.8% and 91.3%, the fallback recovery ratio increases to 2.9% and 8.2%, and the sequence loss ratio remains below 1%. At 4% and 5% base error rates, the CRC pass ratio further decreases to 82.7% and 72.1%, respectively, while the fallback recovery ratio increases sharply to 15.1% and 24.1%, respectively. This shift indicates that under high base error rates, the top-k DP and equal-distance candidate refinement strategies can no longer identify candidates that pass the CRC check for a portion of sequences. However, the minimum-cost fallback strategy still recovers these sequences as the closest approximations to the original codewords, preventing them from being directly discarded. Even at a 5% base error rate, the combined proportion of CRC-passed and fallback-recovered sequences remains above 95%, with only ~4% of the sequences lost. These results indicate that the error-correction module of DNA-FRSC preserves most of the encoded information at high base error rates.

To determine the contributions of the main components of DNA-FRSC, we evaluated the complete system and six ablated variants. (1) The complete DNA-FRSC system served as the reference configuration. (2) The feature imputation module was removed, and missing feature values were filled with zeros to assess its role in compensating for missing feature blocks. (3) The minimum-cost fallback module was removed, and sequences that failed CRC verification were directly discarded to assess its contribution to limiting sequence loss. According to the decoding statistics in Fig. 7, removing this module increased the sequence loss rate to approximately 28% at a base error rate of 5%. (4)

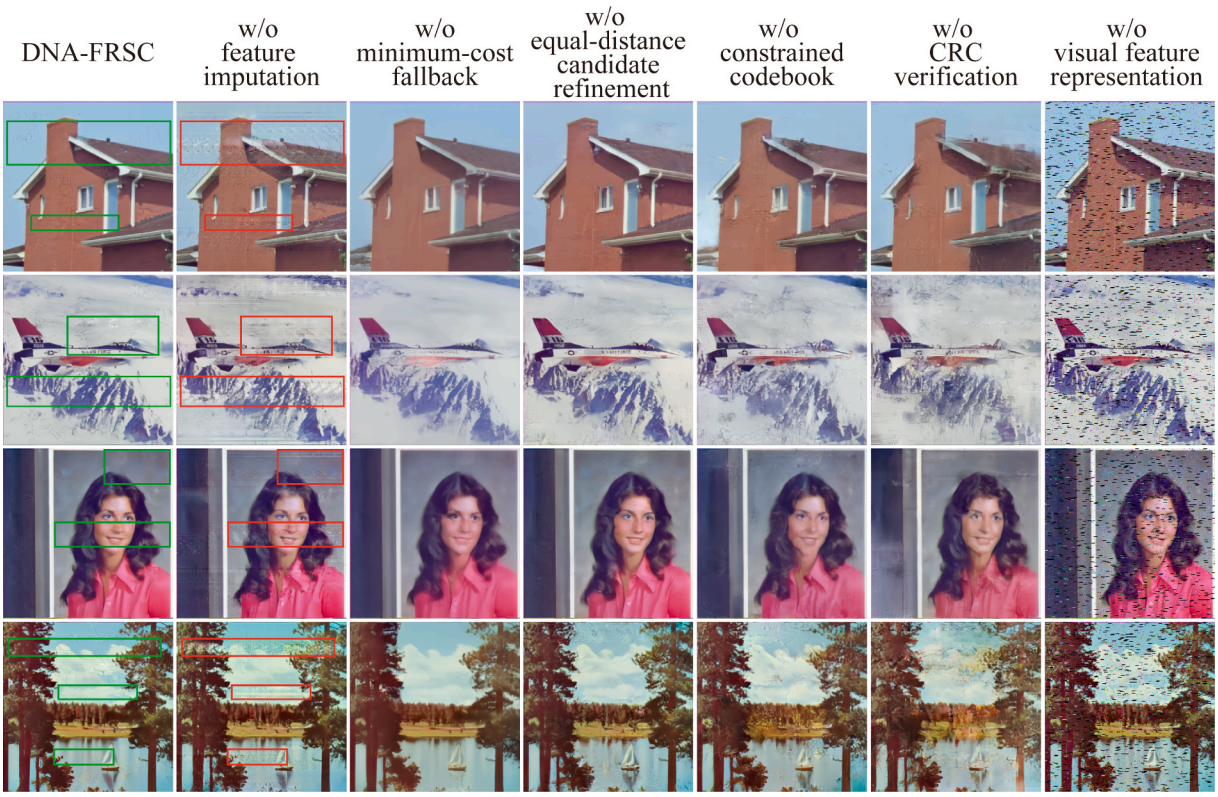


Fig. 8. Visual reconstruction results of DNA-FRSC under different ablation configurations at a 5% base error rate with an IDS ratio of 17%:40%:43%.

Table 2
Quantitative reconstruction results of DNA-FRSC under different ablation configurations.

Method	Image	MSE	PSNR (dB)	SSIM	MS-SSIM	UQI	FSIM
DNA-FRSC	House	63.20	30.12	0.83	0.96	0.99	0.93
	Airplane	98.61	28.19	0.86	0.95	0.97	0.91
	Female	46.82	31.43	0.90	0.98	0.99	0.95
	Sailboat	148.86	26.40	0.83	0.95	0.98	0.90
w/o feature imputation	House	198.29	25.16	0.74	0.88	0.96	0.88
	Airplane	343.48	22.77	0.74	0.83	0.90	0.85
	Female	110.87	27.68	0.82	0.93	0.98	0.91
	Sailboat	331.07	22.93	0.79	0.92	0.96	0.89
w/o minimum-cost fallback	House	197.43	25.18	0.78	0.89	0.96	0.88
	Airplane	404.47	22.06	0.80	0.75	0.87	0.88
	Female	105.32	27.91	0.87	0.95	0.98	0.93
	Sailboat	412.32	21.98	0.76	0.90	0.95	0.85
w/o equal-distance candidate refinement	House	109.65	27.73	0.81	0.92	0.98	0.91
	Airplane	192.85	25.28	0.82	0.91	0.95	0.89
	Female	93.30	28.42	0.88	0.96	0.98	0.94
	Sailboat	309.89	23.22	0.76	0.91	0.96	0.87
w/o constrained codebook	House	242.21	24.29	0.78	0.89	0.95	0.90
	Airplane	411.91	21.98	0.75	0.77	0.87	0.86
	Female	177.74	25.63	0.83	0.93	0.96	0.92
	Sailboat	502.97	21.12	0.69	0.83	0.94	0.84
w/o CRC verification	House	321.78	23.06	0.76	0.86	0.94	0.89
	Airplane	573.09	20.55	0.69	0.71	0.82	0.85
	Female	516.16	21.00	0.81	0.89	0.90	0.92
	Sailboat	637.98	20.13	0.65	0.68	0.78	0.81
w/o visual feature representation	House	1461.73	16.48	0.38	0.66	0.78	0.77
	Airplane	2146.34	14.81	0.37	0.60	0.63	0.78
	Female	1118.24	17.65	0.40	0.74	0.83	0.79
	Sailboat	1398.47	16.67	0.49	0.75	0.86	0.82

The equal-distance candidate refinement module was removed to assess its role in resolving ambiguities among codewords at the same minimum distance. (5) The constrained codebook was replaced with a randomly generated codebook of 64 distinct, unconstrained 6-nt codewords to assess the contribution of the proposed codebook design to sequence recovery. (6) CRC verification was disabled to assess the contribution of CRC-based validation to sequence recovery. (7) The visual feature representation was replaced with direct encoding of image pixels to assess the contribution of learned feature representations. All seven configurations were evaluated on several images from the USC-SIPI dataset at a base error rate of 5% and an IDS ratio of 17%:40%:43%. The corresponding visual and quantitative results are shown in Fig. 8 and Table 2,

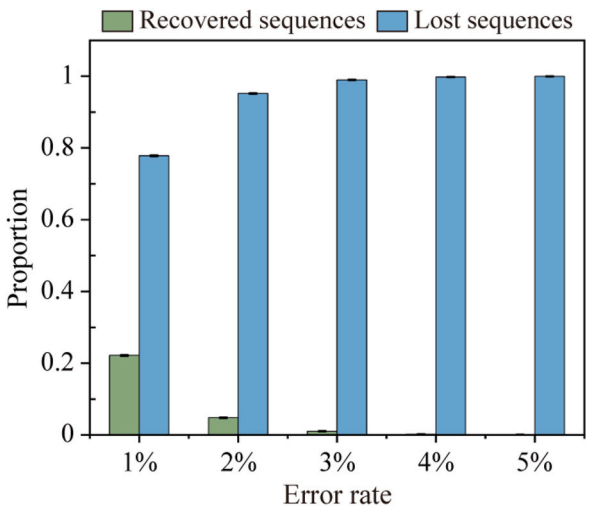


Fig. 9. Proportions of recovered and lost sequences for the Sailboat image without CRC verification and two-stage error correction at different base error rates.

respectively.

The six ablated configurations produce distinct visual degradation patterns, as shown in Fig. 8. Without feature imputation, the boxed regions exhibit horizontal banding, rectangular artifacts, and local color distortions, particularly on the airplane fuselage, house roof, the facial and surrounding regions in Female, and the sky and tree line in Sailboat. These regions remain clean in the complete system but become corrupted under zero filling. The zero-filled missing values are propagated through the decoder during upsampling, producing spatially structured artifacts. Without minimum-cost fallback, the sequence loss increases to approximately 28%. Although feature imputation largely preserves the global structure and color, imputation over large missing regions suppresses high-frequency information, resulting in oversmoothing, loss of fine textures, and occasional color deviations. Removing equal-distance candidate refinement causes milder degradation, with some loss of sharpness and local detail, particularly in Sailboat, because this module resolves only a limited number of ambiguities among codewords at the same minimum distance. Replacing the constrained codebook with a randomly generated, unconstrained codebook causes blurring, texture loss, and local color deviations while preserving the main scene structures. The random codebook does not enforce a minimum Hamming distance or minimize insertion- and deletion-neighborhood overlaps. Consequently, IDS-corrupted segments are more likely to produce ambiguous or incorrect matches, leading to missing or erroneous feature values and degraded reconstruction quality. Without CRC verification, the overall image content remains recognizable, but blurring, fine-detail loss, and local structural distortions become more pronounced. CRC verification rejects unreliable decoding candidates; disabling it allows erroneous feature values to enter the decoder, thereby degrading reconstruction quality. Finally, direct pixel encoding without visual feature representation produces dense, short-line impulse artifacts throughout the images while preserving their main contours. In this variant, residual sequence errors directly corrupt pixel values instead of being mitigated in the noise-tolerant latent space.

Table 2 quantitatively supports the visual comparisons in Fig. 8. Across all four images, the complete DNA-FRSC system achieves the lowest MSE and the highest values for the other five metrics, with PSNR ranging from 26.40 dB to 31.43 dB and SSIM from 0.83 to 0.90. Direct pixel encoding without visual feature representation causes the greatest degradation, reducing PSNR to 14.81–17.65 dB and SSIM to 0.37–0.49. Its net information density is only 0.80 bits/nt, compared with 6.41 bits/nt for DNA-FRSC, demonstrating the advantages of latent representation in both storage density and reconstruction robustness. Among the remaining ablations, disabling CRC verification causes the most pronounced performance loss, reducing PSNR to 20.13–23.06 dB and SSIM to 0.65–0.81. The next-largest degradation is observed when the constrained codebook is replaced with a randomly generated unconstrained codebook, reducing PSNR to 21.12–25.63 dB and SSIM to 0.69–0.83. The random codebook also shows weaker biological constraint satisfaction. For House, the resulting sequences have a maximum homopolymer length of 6 nt and an average GC content of 48%, compared with 3 nt and 50%, respectively, for DNA-FRSC. Removing minimum-cost fallback or feature imputation also substantially reduces reconstruction fidelity, whereas removing equal-distance candidate refinement has a smaller effect.

These ablations demonstrate that the different components of DNA-FRSC play complementary roles. Visual feature representation transforms correlated pixels into a compact, noise-tolerant latent space, preventing residual sequence errors from directly corrupting individual pixels. The constrained codebook improves sequence distinguishability and biological constraint satisfaction by enforcing a minimum Hamming distance and reducing insertion- and deletion-neighborhood overlaps. CRC verification rejects unreliable decoding candidates, while equal-distance candidate refinement resolves ambiguities among codewords at the same minimum distance. Minimum-cost fallback limits sequence loss by retaining the lowest-cost candidate when CRC verification fails, and feature imputation repairs the remaining missing feature blocks. Among the components specifically addressing sequence loss, feature imputation is particularly important. It preserves recognizable structures and colors even when removing minimum-cost fallback increases sequence loss to approximately 28%. In contrast, replacing feature imputation with zero filling causes substantial degradation at a sequence loss rate of only approximately 4%. Overall, these complementary components enable DNA-FRSC to maintain high reconstruction fidelity, while compact visual feature representation substantially reduces storage overhead.

To evaluate the contribution of the error-correction module, CRC verification and the two-stage error-correction procedure were disabled. Reads retaining the original length were decoded by exact codebook matching, whereas length-altered reads were discarded. The Sailboat image was used for evaluation, and each expected sequence was classified as recovered or lost according to whether it could be decoded. As shown in Fig. 9, at a base error rate of 1%, only 22.2% of the sequences are recovered, whereas 77.8% are lost. At base error rates of 3% or higher, nearly all sequences are lost. This extensive sequence loss prevents reliable recovery of the feature tensor. Consequently, even at a 1% base error rate, the reconstructed image is severely corrupted and contains no recognizable content.

3.3. Robustness under the indel-dominated IDS channel

To investigate the effect of IDS composition on the reconstruction performance of DNA-FRSC, six quality metrics were evaluated for the reconstructed Peppers image under four representative IDS ratios. As shown in Fig. 10, all metrics indicate monotonic but gradual degradation as the base error rate increases. In the low error range from 1% to 3%, the curves for the four IDS ratios nearly overlap, with the PSNR remaining above 31 dB and the SSIM remaining above 0.90. Differences among the ratios become more pronounced at base error rates of 4% and 5%. Even at a 5% base error rate, the PSNR remains above 26.81 dB and the SSIM remains above 0.85 for all ratios, indicating stable reconstruction across the evaluated error compositions. Comparison across the four ratios further shows that reconstruction quality is closely associated with the combined proportion of indels. Under the ratio with the lowest proportion of indels, 25%:25%:50%, the best results are achieved across the six metrics, whereas the worst metric values are obtained under the ratio with the highest proportion of indels, 12%:62%:26%. At a 5% base error rate, the PSNR and SSIM differences between these two ratios are approximately 1.3 dB and 0.02,

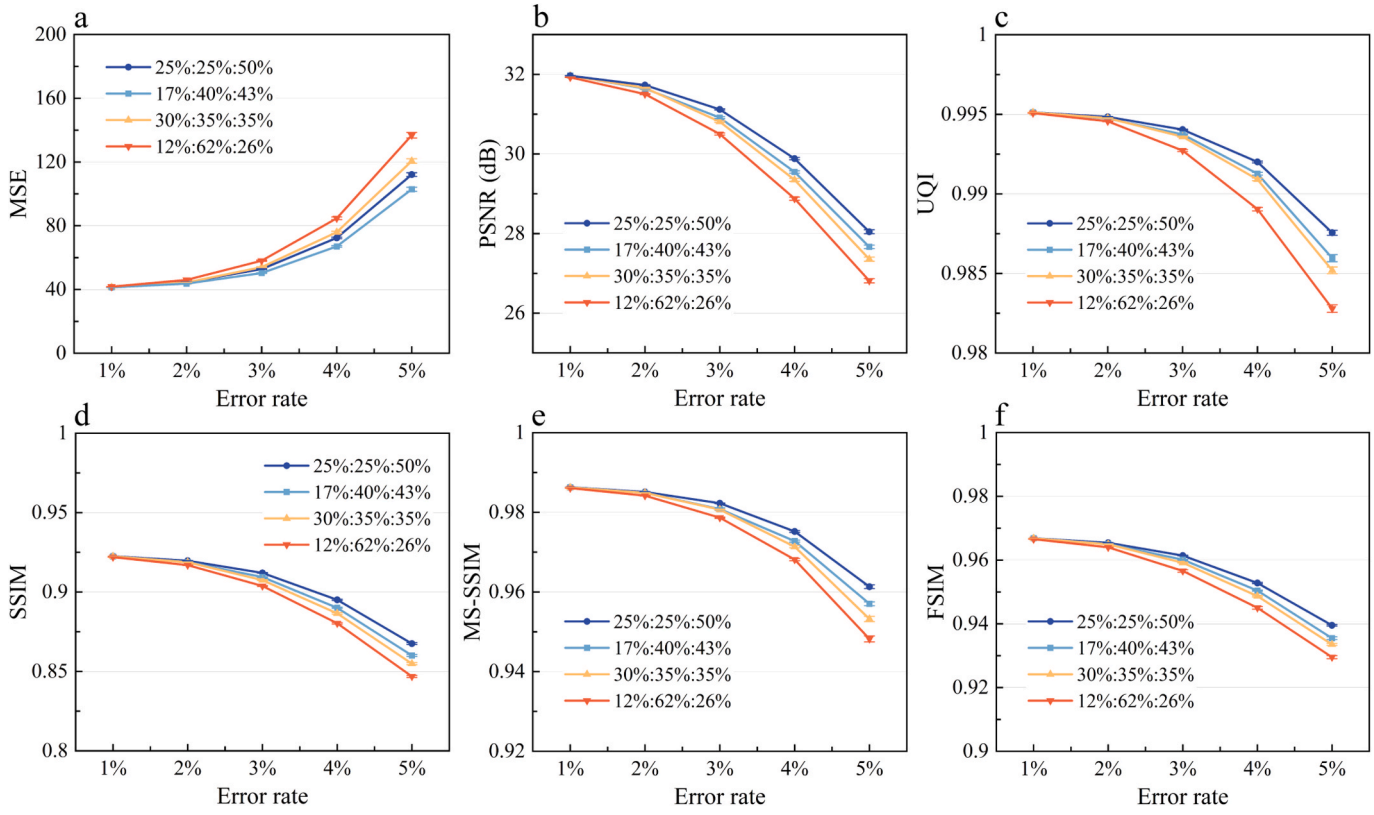


Fig. 10. Reconstruction quality of the Peppers image under different base error rates and IDS ratios: (a) MSE, (b) PSNR, (c) UQI, (d) SSIM, (e) MS-SSIM, and (f) FSIM.

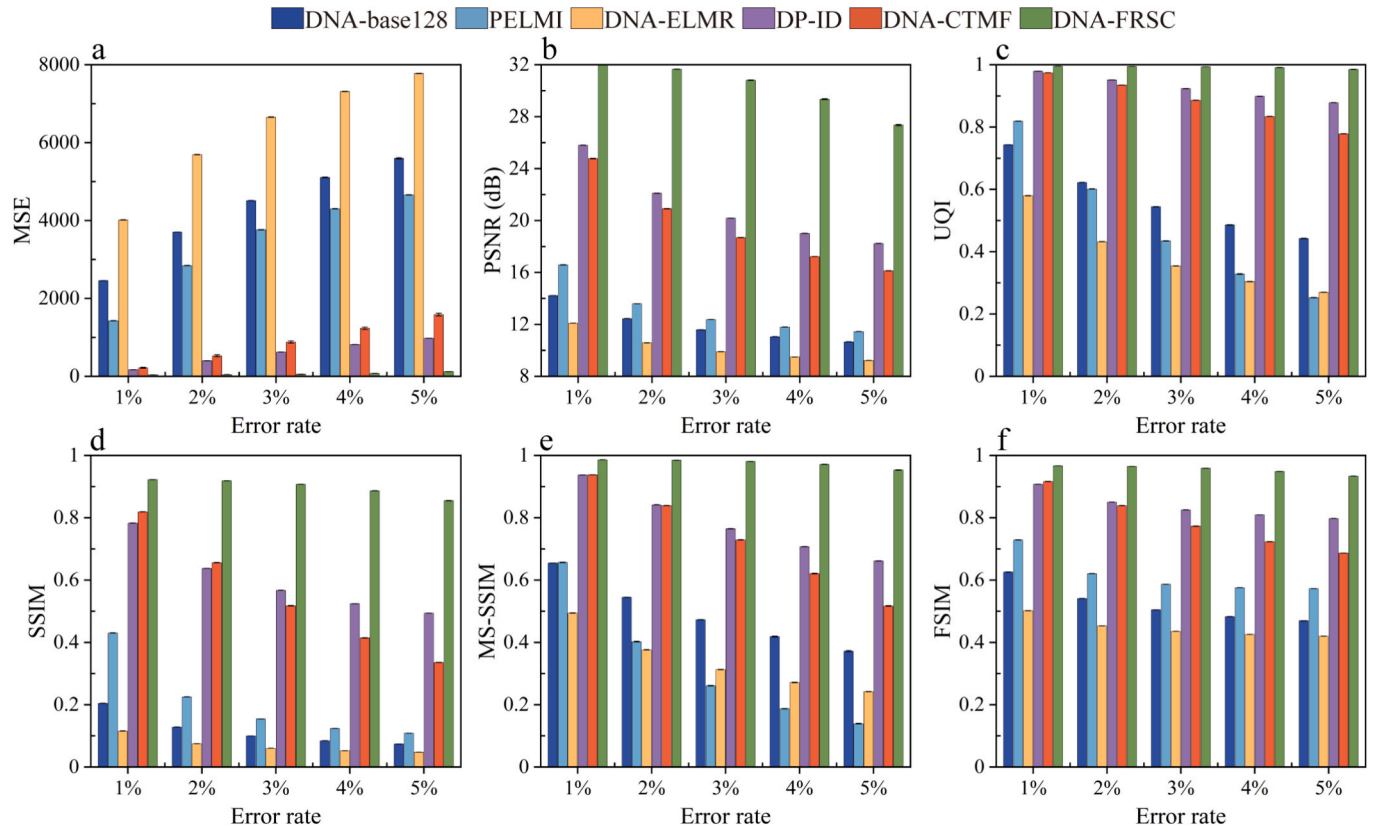


Fig. 11. Quantitative comparison of different DNA-based image storage methods on the Peppers image under different base error rates with an IDS ratio of 30%:35%:35%: (a) MSE, (b) PSNR, (c) UQI, (d) SSIM, (e) MS-SSIM, and (f) FSIM.

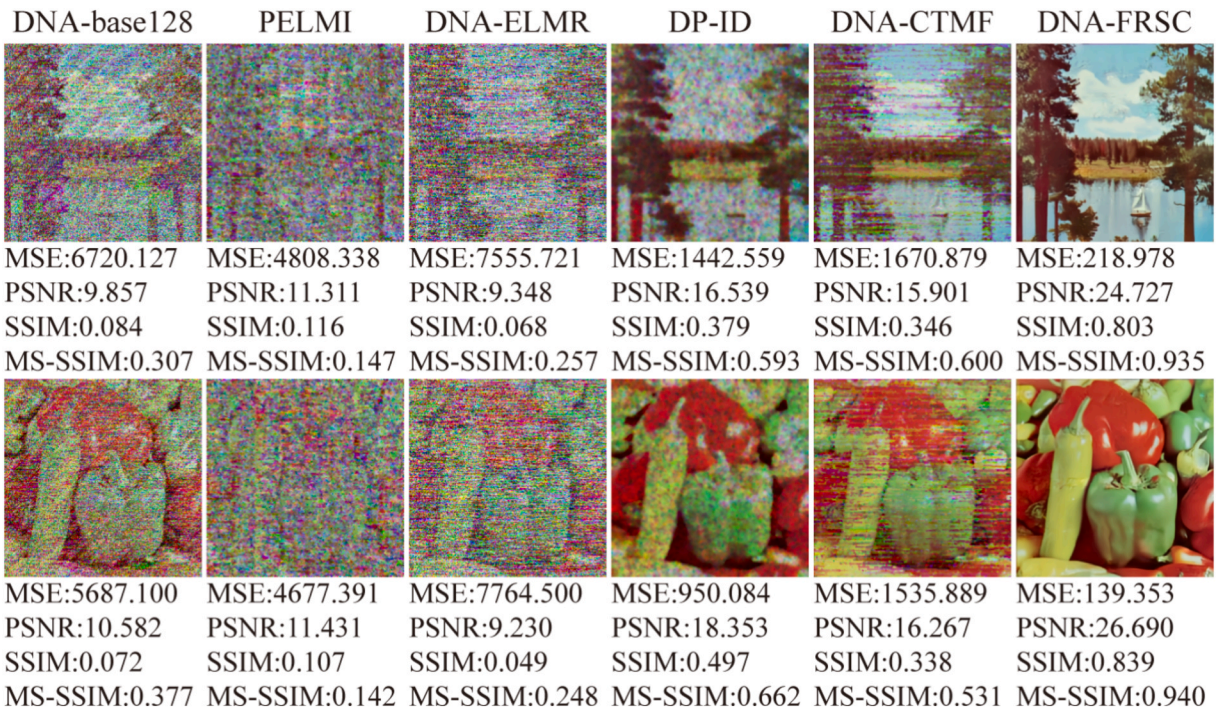


Fig. 12. Visual reconstruction results of Sailboat and Peppers using different DNA-based image storage methods at a 5% base error rate with an IDS ratio of 30%:35%:35%.

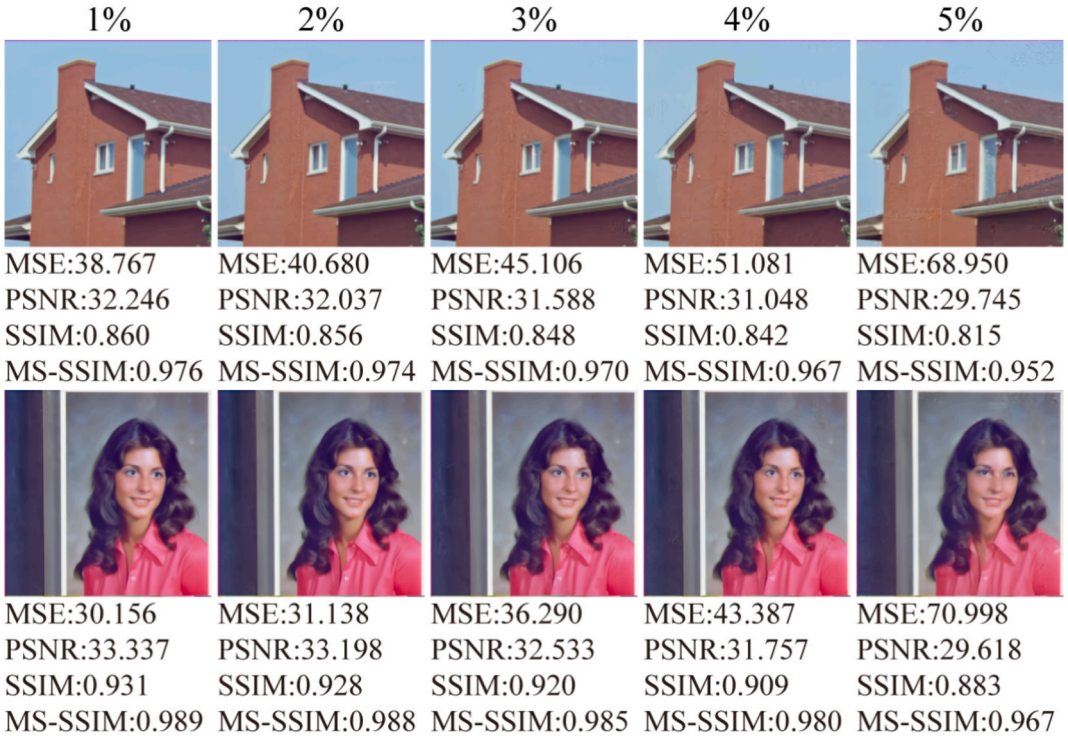


Fig. 13. Visual reconstruction results of House and Female under different base error rates with an IDS ratio of 12%:62%:26%.

respectively. This trend reflects the distinct effects of the three error types. Substitutions affect individual bases, whereas indels disrupt the alignment of subsequent bases and cause greater decoding errors. Overall, DNA-FRSC maintains robust performance across different IDS ratios and can maintain acceptable reconstruction quality even when the proportion of indels is high.

To verify the advantages of DNA-FRSC over existing methods under a high proportion of indels, different DNA-based image storage methods were evaluated on the Peppers image with an IDS ratio of 30%:35%:35%, where indels account for approximately 2/3 of the errors. As shown in Fig. 11, DNA-FRSC consistently outperforms the other five methods across all base error rates and all six metrics. At a 1% base error rate, DNA-FRSC achieves a PSNR of approximately 31.96 dB and an SSIM of 0.92, compared with 25.80 dB and 0.82 for DP-ID, the best-

Table 3
Comparison of different compression-based DNA image storage methods.

Method	Compression method	Compression ratio	Error-correction method	Post-decoding compensation
JPEG-DNA [29]	JPEG-based DCT, scalar quantization	11.25:1	–	–
INNSE [36]	Invertible neural network-based downsampling	16.00:1	Hamming/edit-distance self-correction	–
DNA-QLC [35]	QRes-VAE-based latent compression, quantization	33.90:1	Inner LC	–
HELIX [19]	Block-based DCT, quantization, duplicate removal	15.46:1	Inner LM, outer RS	Deep learning-based image restoration
DP-ID [25]	Dynamic programming-based pixel compression	1.01:1	Inner RS, interleaving /deinterleaving	Median filtering
DNA-FRSC	SwinJSCC-based feature extraction, linear quantization	8.00:1	Inner RS, CRC, top-k DP, candidate refinement	Training-free feature imputation

performing baseline. The performance gap widens as the base error rate increases. From 1% to 5%, the PSNR of DNA-FRSC decreases from approximately 31.96 dB to 27.36 dB, whereas that of DP-ID decreases from approximately 25.80 dB to 18.23 dB. At a 5% base error rate, DNA-FRSC outperforms DP-ID by approximately 9.13 dB in PSNR, 0.36 in SSIM, and 0.29 in MS-SSIM. In contrast, DNA-base128, PELMI, and DNA-ELMR show much lower reconstruction quality under this IDS ratio, with PSNRs of less than 12 dB and SSIMs of less than 0.11 at a 5% base error rate. This result may be related to their direct encoding at the pixel level, as base shifts caused by indels can propagate along the sequence and lead to large-scale decoding misalignment errors. These results show that DNA-FRSC can still reconstruct images with high visual quality under error conditions dominated by indels.

Visual reconstruction results for the Sailboat and Peppers images at a 5% base error rate and an IDS ratio of 30%:35%:35% are shown in Fig. 12. Consistent with the quantitative results, the six methods exhibit

Table 4
Encoding and decoding times and computational complexities of different DNA-based image storage methods.

Method	Encoding time (ms)	Decoding time (s)	Encoding complexity	Decoding complexity
DNA-base128 [22]	447.20 ± 2.06	31.83 ± 0.14	$O(N)$	$O(N)$
PELMI [23]	3137.20 ± 4.87	23.12 ± 0.10	$O(N)$	$O(N\log N)$
DNA-ELMR [37]	1377.99 ± 4.51	0.82 ± 0.04	$O(N)$	$O(N)$
DNA-CTMF [26]	33.01 ± 0.06	2.52 ± 0.07	$O(N)$	$O(N)$
DP-ID [25]	1432.18 ± 3.62	1.16 ± 0.05	$O(N+S)$	$O(N+S)$
JPEG-DNA [29]	167.28 ± 1.59	1.78 ± 0.02	$O(N+S)$	$O(N+S)$
HELIX [19]	149.25 ± 1.34	0.15 ± 0.02	$O(N+S)$	$O(N+S^2)$
DNA-QLC [35]	284.35 ± 1.89	45.32 ± 0.24	$O(N+S\log S)$	$O(N+S^2)$
INNSE [36]	275.36 ± 0.92	6.58 ± 0.17	$O(N\log N)$	$O(N^2)$
DNA-FRSC	702.14 ± 0.68	9.16 ± 0.01	$O(N)$	$O(N)$

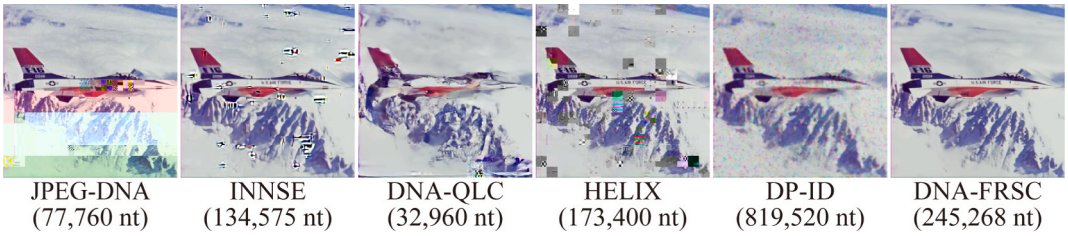


Fig. 14. Visual comparison of different compression-based DNA image storage methods at a 1% base error rate.

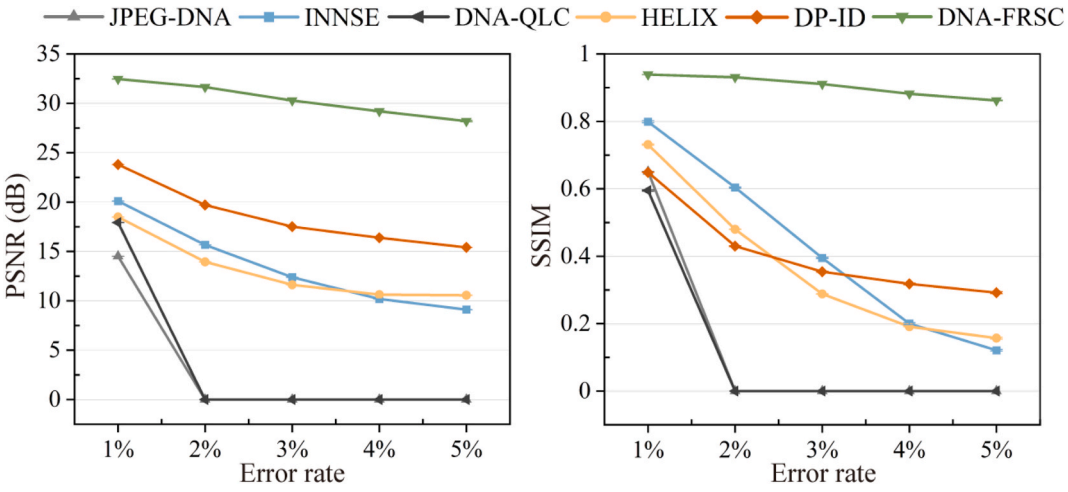


Fig. 15. Quantitative comparison of reconstructed image quality among different compression-based DNA image storage methods under different base error rates.

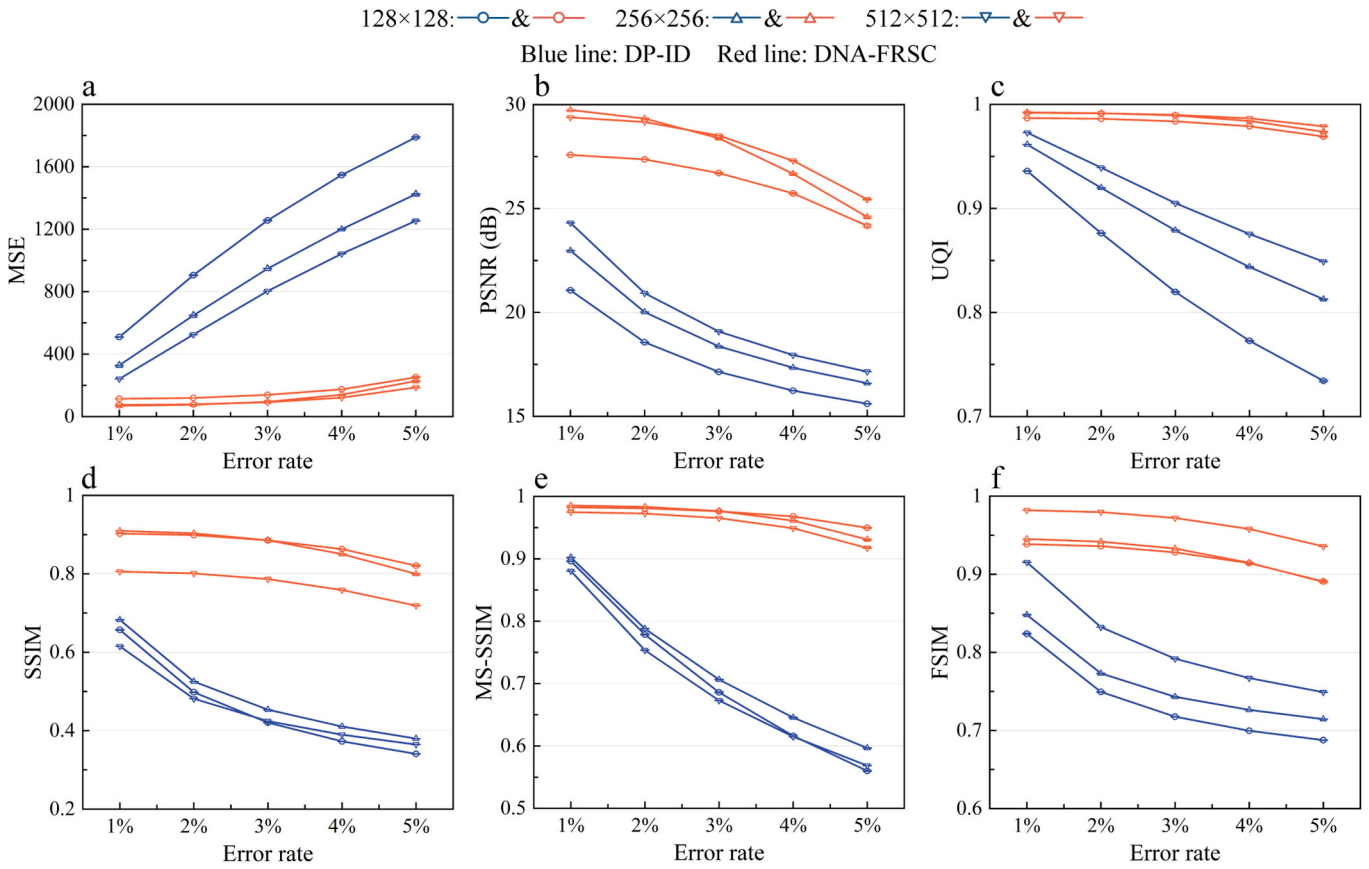


Fig. 16. Reconstruction quality of the Sailboat image using DP-ID and DNA-FRSC at different resolutions and base error rates with an IDS ratio of 17%:40%:43%: (a) MSE, (b) PSNR, (c) UQI, (d) SSIM, (e) MS-SSIM, and (f) FSIM.

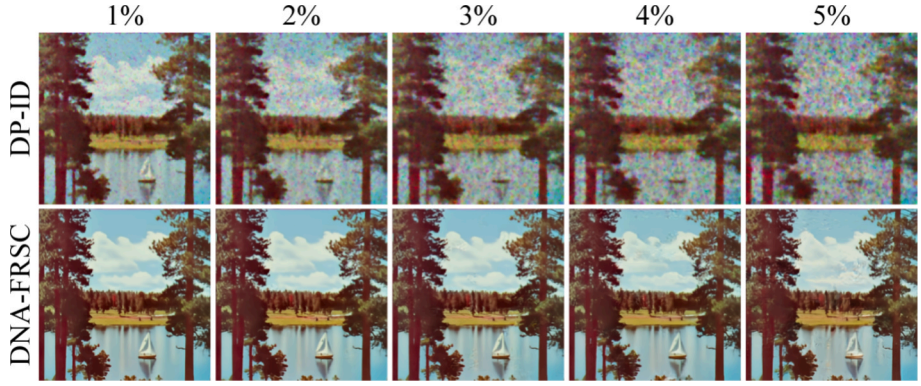


Fig. 17. Visual reconstruction results of the 256 × 256 Sailboat image using DP-ID and DNA-FRSC at different base error rates.

three distinct levels of reconstruction quality. The images reconstructed by DNA-base128, PELMI, and DNA-ELMR are severely degraded and contain substantial color noise, making the original structure and main content difficult to recognize. For these methods, the PSNR is generally less than 12 dB, and the SSIM is less than 0.12. This degradation occurs primarily because these methods are sensitive to base shifts caused by indels. Such shifts can propagate along the sequence and cause large-scale decoding misalignment, leading to severe image distortion. DP-ID and DNA-CTMF produce better reconstructions than the first three methods and retain the rough outlines of the images. However, the reconstructed images still show pronounced horizontal stripes, local blurring, and block-like color artifacts. For example, the sky region in the Sailboat image and the surface region in the Peppers image both show noticeable structural distortion. The corresponding metrics also indicate that the PSNRs of these two methods mainly range between 16 dB and 18 dB, while the SSIMs remain below 0.50. These results indicate that image postprocessing methods such as median filtering can suppress discrete noise to some extent. Under a high proportion of indels, continuous shift errors result in more structural damage; thus, post-processing alone is insufficient to fully recover image details.

In contrast, the reconstructions produced by DNA-FRSC are visually closest to the original images. For Sailboat, DNA-FRSC preserves a relatively complete scene structure and natural colors, achieving PSNR, SSIM, and MS-SSIM of 24.73 dB, 0.80, and 0.94, respectively. For Peppers, DNA-FRSC preserves the object contours and color distribution, achieving PSNR, SSIM, and MS-SSIM of 26.96 dB, 0.84, and 0.94, respectively. These results indicate that DNA-FRSC mitigates sequence misalignment and local feature loss through compact feature-domain

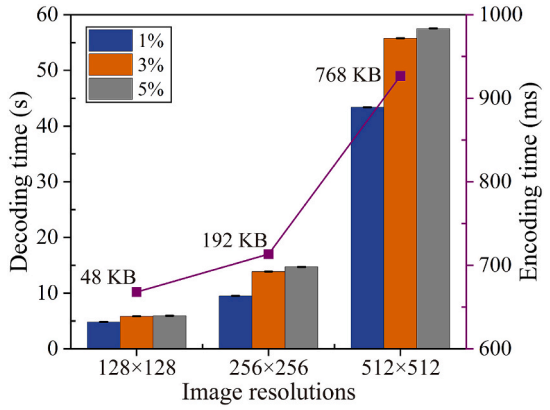


Fig. 18. Encoding times of DNA-FRSC for different image resolutions and decoding times for the same resolutions at base error rates of 1%, 3%, and 5%.

Table 5
Peak RAM and GPU memory usage during DNA-FRSC encoding and decoding.

Image resolution	Encoding peak RAM (MiB)	Encoding peak GPU memory (MiB)	Decoding peak RAM (MiB)	Decoding peak GPU memory (MiB)
128 × 128	655.43 ± 0.12	232.44 ± 0.76	1089.39 ± 0.95	542.43 ± 2.68
256 × 256	657.27 ± 0.24	232.64 ± 1.38	1076.21 ± 0.97	732.06 ± 2.86
512 × 512	663.02 ± 0.17	233.24 ± 1.44	1078.40 ± 1.30	1382.29 ± 2.25

encoding, robust sequence recovery, and feature imputation, preserving visual quality under an error distribution with a high proportion of indels.

The reconstruction performance of DNA-FRSC was further evaluated on the House and Female images using an IDS ratio of 12%:62%:26%, with indels accounting for approximately three quarters of all errors. This setting simulates an extreme indel-dominated condition, and the corresponding reconstruction results are shown in Fig. 13. Across base error rates ranging from 1% to 5%, the visual quality of both images remains high, and the structures, textures, and colors of the original images are preserved in the reconstructed images without noticeable artifacts or distortions. With respect to objective evaluation metrics, the PSNR of House decreases from 32.25 dB at a 1% base error rate to 29.75 dB at a 5% base error rate, a decrease of 2.5 dB, whereas the PSNR of Female decreases from 33.34 dB to 29.62 dB at the same base error rates. The SSIM and MS-SSIM show similar trends. The SSIM of House decreases from 0.86 to 0.82, the SSIM of Female decreases from 0.93 to 0.88, and the MS-SSIMs remain above 0.95 throughout. Over the full error range, the PSNRs of both images remain at approximately 30 dB, and the SSIMs remain above 0.81. These results demonstrate that DNA-FRSC produces reconstructions with high visual quality even under an extreme error distribution with a high proportion of indels.

3.4. Comparison with compression-based DNA image storage methods

DNA-FRSC was further compared with five representative compression-based DNA image storage methods. The main characteristics are summarized in Table 3. For JPEG-DNA, the simulation employed JPEG-based 8 × 8 DCT and scalar quantization with a quality factor of 50, and the sequence length was set to 200 nt. The compression ratio was calculated as the ratio of the bit length of the original 256 × 256 Airplane image to that of the compressed bitstream. Fig. 14 presents reconstructed images obtained at a 1% base error rate. Fig. 15 compares the PSNRs and SSIMs of the different methods across different base error rates. All comparisons were conducted under an IDS ratio of 30%:35%:35%.

As shown in Figs. 14 and 15, DNA-FRSC consistently achieves the highest reconstruction quality across all base error rates. Although JPEG-DNA, INNSE, DNA-QLC, and HELIX require fewer encoded bases than DNA-FRSC, they exhibit distinct reconstruction distortions. JPEG-DNA employs variable-length quaternary coding, whereby an indel may shift subsequent codewords and propagate decoding errors, leading to severe structural distortion or reconstruction failure. INNSE stores downsampled image data generated by an invertible neural network. However, its self-correction relies mainly on local Hamming- and edit-distance matching without additional redundancy, limiting its effectiveness against indels. Residual errors in the low-resolution image may therefore be amplified during inverse reconstruction. DNA-QLC uses QRes-VAE to compress images into entropy-coded latent bitstreams. However, its Levenshtein coding is primarily effective against multiple errors of the same type and corrects only a limited set of substitution patterns. Under mixed IDS errors, residual errors may corrupt the latent bitstreams and be amplified during QRes-VAE decoding, resulting in severe color and structural distortion. HELIX limits error propagation by independently compressing fixed-size image blocks and combining cascaded error correction with image restoration. Nevertheless, unrecovered blocks may still produce local block and color artifacts. By comparison, DP-ID disperses indel-induced burst errors into salt-and-pepper noise through interleaving and then suppresses the noise using median filtering, preserving recognizable image structures more consistently, although residual noise and smoothing remain.

Together with the broader comparison in Fig. 11, the PSNRs and SSIMs in Fig. 15 indicate that DP-ID is the strongest overall baseline among the compared methods. Although JPEG-DNA, INNSE, DNA-QLC, and HELIX require fewer encoded bases, their reconstruction quality deteriorates markedly under IDS errors. In particular, when the base error rate exceeds 1%, the reconstructions produced by JPEG-DNA and DNA-QLC completely collapse, and the original image content can no longer be recovered. These methods are less suitable as the principal references for a joint assessment of reconstruction robustness and storage density. Therefore, DP-ID was selected as the principal comparator for the resolution study in Section 3.5 and as the reference method for the storage-overhead and synthesis-cost analysis in Section 3.6.

Encoding and decoding times were measured for the 256 × 256 Airplane image at a 1% base error rate with an IDS ratio of 30%:35%:35%. Table 4 reports the mean runtimes and 95% confidence intervals based on 500 repeated runs, together with the theoretical encoding and decoding complexities of the compared methods. Combined with the reconstruction results in Fig. 15, the results reveal a clear trade-off between computational efficiency and reconstruction quality. Although several competing methods require shorter decoding times, their reconstruction quality degrades substantially as the base error rate increases. For example, HELIX requires only 0.15 s for decoding, but its PSNR decreases to approximately 10 dB at a 5% base error rate, whereas DNA-FRSC maintains a PSNR of approximately 28 dB.

For the complexity analysis, N denotes the number of pixels in the input image, and S denotes the number of encoded DNA sequences. The non-compression-based methods generally scale linearly with N, with PELMI being the only exception because of its near-linear decoding complexity. Among the compression-based methods, DP-ID and JPEG-DNA exhibit linear dependence on both N and S. In contrast, HELIX, DNA-QLC, and INNSE introduce quadratic decoding terms, which depend on S for the former two methods and on N for INNSE. DNA-FRSC avoids these higher-order terms and maintains $O(N)$ complexity for both encoding and decoding under the fixed system configuration. Although asymptotic complexity does not directly predict runtime for a fixed input size, the absence of higher-order terms suggests favorable scalability with increasing image size. Overall, DNA-FRSC achieves a favorable balance among moderate runtime, linear computational complexity, and superior reconstruction quality.

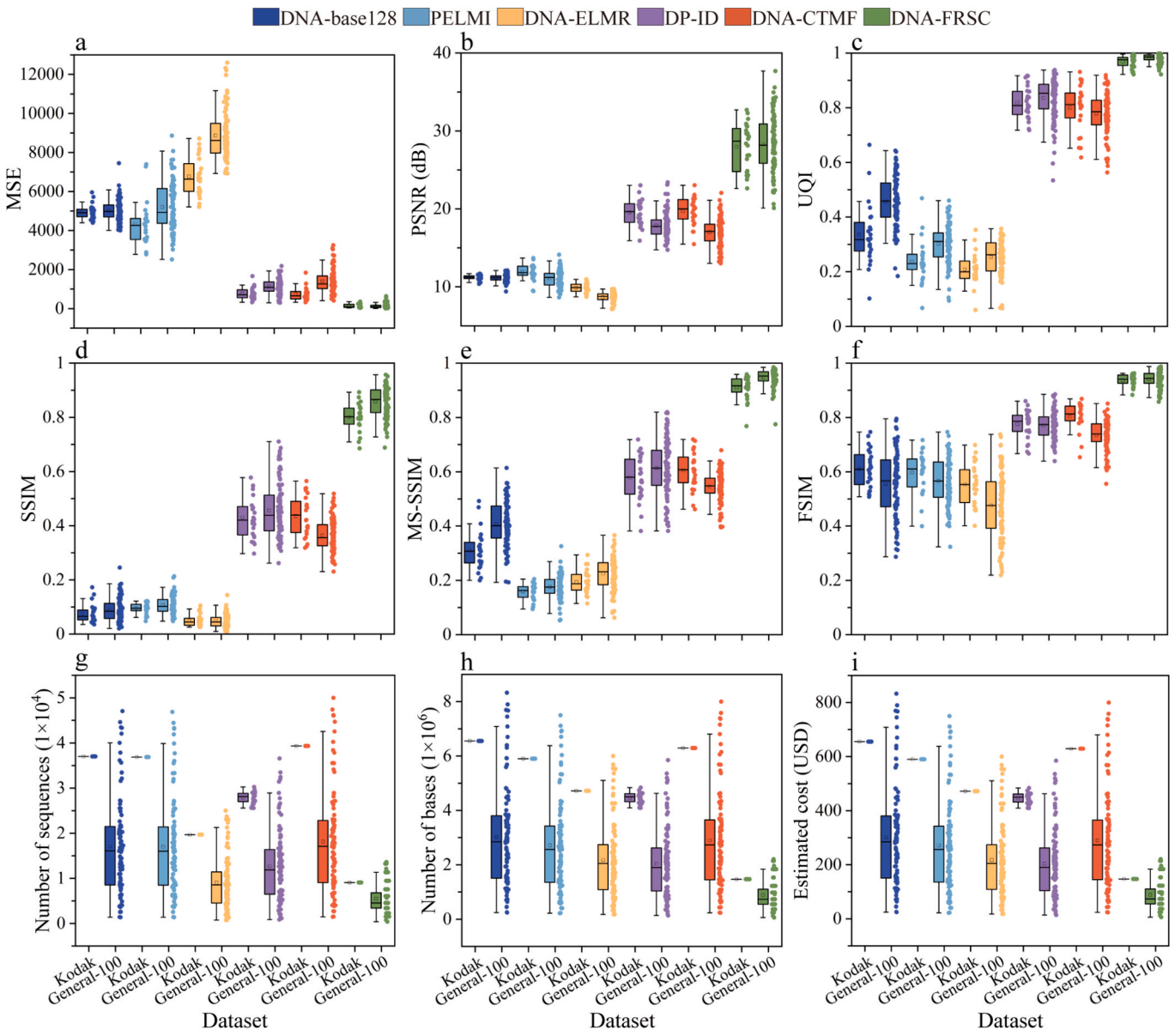


Fig. 19. Reconstruction quality and storage overhead of different DNA-based image storage methods on Kodak and General-100. The reconstruction quality results were obtained at a 5% base error rate with an IDS ratio of 25%:25%:50%: (a) MSE, (b) PSNR, (c) UQI, (d) SSIM, (e) MS-SSIM, (f) FSIM, (g) number of sequences, (h) number of bases, and (i) estimated costs.

Table 6
Comparison of the reconstruction quality of the different methods on Kodak and General-100.

Dataset	Method	MSE	PSNR (dB)	SSIM	MS-SSIM	UQI	FSIM
Kodak	DNA-base128 [22]	4946 ± 157	11.200 ± 0.135	0.075 ± 0.013	0.310 ± 0.030	0.336 ± 0.051	0.611 ± 0.030
	PELMI [23]	4328 ± 479	11.892 ± 0.439	0.095 ± 0.008	0.158 ± 0.013	0.238 ± 0.034	0.598 ± 0.034
	DNA-ELMR [37]	6766 ± 413	9.873 ± 0.258	0.050 ± 0.008	0.194 ± 0.017	0.208 ± 0.025	0.554 ± 0.034
	DP-ID [25]	774 ± 128	19.543 ± 0.693	0.430 ± 0.034	0.578 ± 0.034	0.818 ± 0.025	0.775 ± 0.021
	DNA-CTMF [26]	752 ± 146	19.741 ± 0.752	0.436 ± 0.034	0.605 ± 0.030	0.800 ± 0.038	0.804 ± 0.021
	DNA-FRSC	**130 ± 38**	**27.968 ± 1.263**	**0.801 ± 0.021**	**0.910 ± 0.017**	**0.969 ± 0.008**	**0.937 ± 0.008**
General-100	DNA-base128 [22]	5010 ± 104	11.155 ± 0.087	0.091 ± 0.008	0.409 ± 0.016	0.457 ± 0.018	0.554 ± 0.024
	PELMI [23]	5211 ± 244	11.082 ± 0.206	0.107 ± 0.006	0.176 ± 0.008	0.302 ± 0.014	0.567 ± 0.016
	DNA-ELMR [37]	8875 ± 242	8.687 ± 0.113	0.048 ± 0.004	0.225 ± 0.012	0.254 ± 0.012	0.476 ± 0.024
	DP-ID [25]	1128 ± 73	17.873 ± 0.319	0.456 ± 0.020	0.614 ± 0.020	0.835 ± 0.014	0.772 ± 0.010
	DNA-CTMF [26]	1401 ± 116	17.018 ± 0.349	0.364 ± 0.012	0.545 ± 0.010	0.778 ± 0.014	0.735 ± 0.010
	DNA-FRSC	**131 ± 22**	**28.292 ± 0.698**	**0.857 ± 0.012**	**0.947 ± 0.006**	**0.982 ± 0.004**	**0.940 ± 0.006**

Table 7
Comparison of the storage costs of the different methods on Kodak and General-100.

Dataset	Method	Number of sequences (1×10^4)	Number of bases (1×10^6)	Estimated cost (USD)
Kodak	DNA-base128 [22]	3.701	6.551	655.095
	PELMI [23]	3.686	5.898	589.824
	DNA-ELMR [37]	1.966	4.719	471.864
	DP-ID [25]	2.787 ± 0.051	4.459 ± 0.084	445.899 ± 8.381
	DNA-CTMF [26]	3.932	6.292	629.184
	DNA-FRSC	**0.908**	**1.470**	**147.031**
General-100	DNA-base128 [22]	1.705 ± 0.210	3.017 ± 0.373	301.729 ± 37.293
	PELMI [23]	1.698 ± 0.210	2.717 ± 0.335	271.682 ± 33.579
	DNA-ELMR [37]	0.906 ± 0.111	2.173 ± 0.268	217.340 ± 26.862
	DP-ID [25]	1.272 ± 0.159	2.035 ± 0.253	203.482 ± 25.398
	DNA-CTMF [26]	1.811 ± 0.224	2.898 ± 0.359	289.792 ± 35.819
	DNA-FRSC	**0.550 ± 0.062**	**0.890 ± 0.101**	**89.022 ± 10.080**

3.5. Reconstruction across image resolutions

To evaluate the effect of image resolution on the reconstruction performance of DNA-FRSC, the Sailboat image was tested at resolutions of 128 × 128, 256 × 256, and 512 × 512 under an IDS ratio of 17%:40%:43%. The results are shown in Fig. 16, where the red and blue lines represent the results of DNA-FRSC and DP-ID, respectively. Across all three resolutions and the full base error rate range from 1% to 5%, DNA-FRSC clearly outperforms DP-ID in terms of all six metrics, and the performance gap increases as the base error rate increases. For example, at a 5% base error rate, the PSNR of DNA-FRSC remains above 24.16 dB at all resolutions, whereas that of DP-ID remains below 17.15 dB, resulting in a minimum gap of approximately 7 dB.

A closer observation of the resolution effect reveals that the two methods present similar trends for most metrics. As the resolution increases from 128 × 128 to 512 × 512, the MSEs of both DP-ID and DNA-FRSC decrease, whereas the PSNR, UQI, and FSIM increase. These results indicate that a higher resolution does not weaken the overall pixel fidelity of either method. Compared with DP-ID, DNA-FRSC consistently achieves a lower MSE and higher PSNR, UQI, and FSIM, and its performance changes more smoothly as the base error rate increases. This may be related to the compact encoding scheme used by DNA-FRSC in the feature domain. Higher-resolution images provide richer spatial context and neighborhood redundancy, thereby providing more reliable local information for feature imputation.

Notably, the SSIM and MS-SSIM of DNA-FRSC decrease slightly as the resolution increases, which differs from the trends of the other four metrics. This is because DNA-FRSC uses the same downsampling ratio and quantization precision across all the image resolutions. Higher-resolution images contain richer textures and more local structural details, whereas the degree of feature-domain compression remains unchanged. As a result, a larger proportion of fine structural details are affected during compression. Metrics such as the MSE and PSNR reflect errors at the pixel level and are more sensitive to the recovery of large smooth regions, whereas the SSIM and MS-SSIM reflect the structural similarity and are more sensitive to the loss of fine details, which may explain the different trends. Nevertheless, the SSIM and MS-SSIM of DNA-FRSC at 512 × 512 are clearly higher than those of DP-ID at any resolution. The visual reconstruction results of the 256 × 256 Sailboat image at base error rates ranging from 1% to 5% are shown in Fig. 17. Compared with DP-ID, DNA-FRSC preserves clearer scene structures and more natural colors as the base error rate increases, which is consistent with the quantitative results in Fig. 16. These results demonstrate that DNA-FRSC can stably reconstruct images with high visual quality across different resolutions and maintain good overall fidelity as the resolution increases.

The encoding and decoding times of DNA-FRSC are shown in Fig. 18. Encoding time was measured for each image resolution, while decoding time was measured at base error rates of 1%, 3%, and 5% for each resolution. As the image resolution increases from 128 × 128 to 512 × 512, the pixel count increases 16-fold, whereas the encoding time increases by 38.6%, from 667.98 ms to 926.14 ms. In contrast, decoding time increases markedly with image resolution at all three base error rates. Specifically, it increases from 4.82 s to 43.37 s at a 1% base error rate, from 5.83 s to 55.76 s at a 3% base error rate, and from 5.93 s to 57.53 s at a 5% base error rate. At each resolution, the increase in decoding time is greater from 1% to 3% than from 3% to 5%. Higher base error rates cause more sequences to undergo top-k DP correction and equal-distance candidate refinement, thereby increasing the decoding time. However, because the number of retained candidate paths and the refinement search range are bounded, the additional decoding overhead grows more slowly from 3% to 5%. These results show that DNA-FRSC limits the growth of decoding overhead at high base error rates.

Peak RAM and GPU memory usage were measured during encoding and decoding at different image resolutions under a 3% base error rate. Values are reported as means with 95% confidence intervals based on 500 repeated runs. As shown in Table 5, despite the 16-fold increase in pixel count from 128 × 128 to 512 × 512, encoding RAM, encoding GPU memory, and decoding RAM vary by no more than 1.3%, indicating that these memory requirements remain effectively stable over the tested resolution range. In contrast, peak GPU memory during decoding increases from 542.43 to 1382.29 MiB. This 2.55-fold increase is substantially smaller than the increase in pixel count, and the maximum demand of approximately 1.35 GiB remains well within the 16 GB VRAM of the experimental platform. Together with the runtime results in Fig. 18, these findings indicate that decoding is the dominant computational bottleneck as image resolution increases, while the overall memory requirements remain moderate.

3.6. Generalization across natural image benchmarks

To further evaluate the reconstruction performance and adaptability of DNA-FRSC across different image contents, experiments were conducted on two public datasets, General-100 [54] and Kodak. Together, these datasets cover diverse scenes, textures, resolutions, and color distributions, thereby enabling a broader evaluation of content adaptability. The IDS ratio was set to 25%:25%:50%, and the base error rate was set to 5%. The General-100 dataset contains 100 uncompressed BMP color images with different resolutions, including scenes with components such as people, buildings, and natural landscapes. These high-quality images contain rich details. The Kodak dataset is a widely used standard test set in image processing. It contains 24 PNG color images and provides realistic color representations and rich texture details.

The reconstruction quality of the different methods on Kodak and General-100 was evaluated using the six abovementioned metrics. The boxplot distributions are shown in Fig. 19(a)–(f), and the corresponding means and 95% confidence intervals are reported in Table 6. DNA-FRSC performs best across all six metrics on both datasets. All improvements were calculated relative to the best-performing baseline for each metric and dataset. DNA-FRSC achieves MSEs of 130 and 131 on Kodak and General-100, corresponding to reductions of 82.7% and 88.4%, respectively. Its PSNRs reach 27.97 dB and 28.29 dB, exceeding those of the best baselines by 8.23 dB and 10.42 dB. DNA-FRSC also achieves

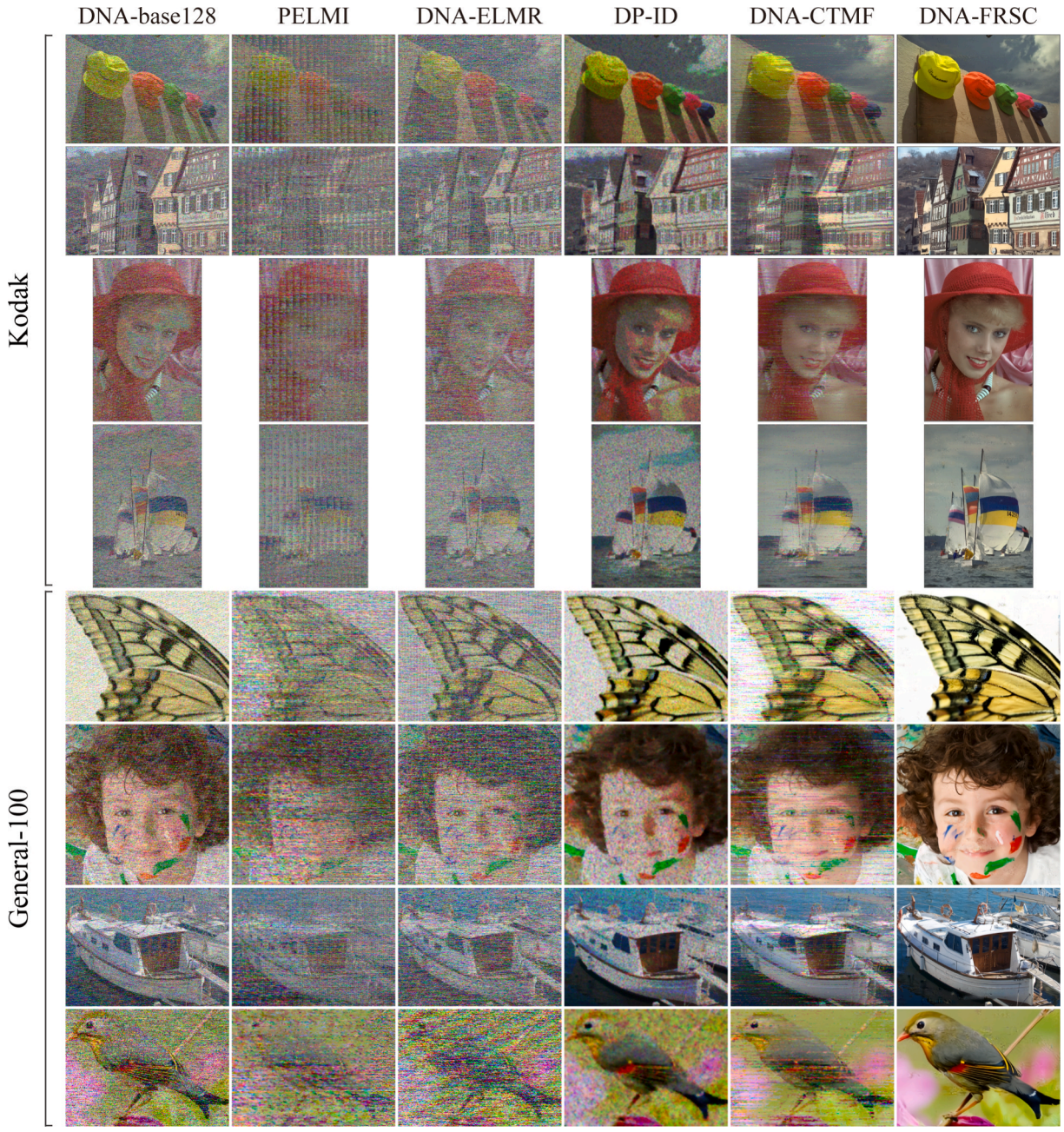


Fig. 20. Visual reconstruction results of representative images from Kodak and General-100 using different DNA-based image storage methods.

SSIMs of 0.801 and 0.857 and MS-SSIMs of 0.910 and 0.947, corresponding to improvements of 83.7%–87.9% and 50.4%–54.2%, respectively. Moreover, DNA-FRSC improves UQI by 17.6%–18.5% and FSIM by 16.5%–21.8% across the two datasets. In contrast, DNA-base128, PELMI, and DNA-ELMR yield PSNRs below 12 dB and SSIMs below 0.12 on both datasets. DP-ID and DNA-CTMF perform better than these three methods but remain inferior to DNA-FRSC across all six metrics. Overall, these results demonstrate the consistent reconstruction performance and cross-dataset generalization of DNA-FRSC.

The storage overhead and estimated synthesis cost of all six methods are compared in Fig. 19(g)–(i) and Table 7. The cost is estimated using an array-based synthesis price of approximately 10^{-4} USD per base [27,55]. On Kodak, DNA-FRSC requires 0.908×10^4 sequences and 1.470×10^6 bases, whereas DP-ID requires 2.787×10^4 sequences and 4.459×10^6 bases. Thus, DNA-FRSC reduces the numbers of sequences and bases by 67.4% and 67.0%, respectively, and lowers the estimated synthesis cost from 445 USD to 147 USD. On General-100, DNA-FRSC requires 0.550×10^4 sequences and 0.890×10^6 bases, compared with 1.272×10^4 sequences and 2.035×10^6 bases for DP-ID. These values correspond to reductions of 56.8% and 56.3% in the numbers of sequences and bases, respectively, while the estimated synthesis cost decreases from 203 USD to 89 USD. These results demonstrate the lower storage overhead and synthesis cost of DNA-FRSC.

Considering both reconstruction quality and storage cost, DNA-FRSC achieves a more favorable trade-off than the baseline methods. DNA-base128, PELMI, and DNA-ELMR exhibit severe degradation under high base error rates, whereas DP-ID and DNA-CTMF preserve more image information but remain inferior to DNA-FRSC across all six metrics. On the Kodak and General-100 datasets, DNA-FRSC also requires fewer bases and incurs a lower estimated synthesis cost than DP-ID. Its narrow 95% confidence intervals in SSIM, MS-SSIM, UQI, and FSIM further indicate stable performance across images with diverse

resolutions and scene contents. These results demonstrate the reconstruction robustness, storage efficiency, and adaptability of DNA-FRSC for practical DNA-based image storage.

The visual reconstruction results of the six methods on representative images from Kodak and General-100 are shown in Fig. 20. The images reconstructed by DNA-base128, PELMI, and DNA-ELMR are dominated by color noise and dense stripes, and the original content is barely recognizable. DP-ID and DNA-CTMF recover the main structure and color of the images, but color shifts and block artifacts remain visible in regions with fine details, such as building windows, facial features, and bird feathers. In contrast, DNA-FRSC produces images with the best visual quality across all the samples, with complete structures, natural colors, and clear details, making its reconstructions closest to the original images. These results are consistent with the objective metrics in Table 6 and further verify that DNA-FRSC maintains stable reconstruction quality across different image contents.

3.7. Discussion and implications

The experimental results demonstrate that DNA-FRSC improves storage efficiency and reconstruction robustness while satisfying multiple biological constraints. These advantages provide a basis for extending the system to the following application scenarios.

(1) **Ultra-high-resolution image storage.** The results in Section 3.5 show that DNA-FRSC maintains favorable reconstruction quality as image resolution increases, while its memory demand grows moderately. Section 3.6 further demonstrates stable reconstruction performance on General-100, which contains images with different resolutions. For images at still higher resolutions, image tiling and parallel processing could be incorporated to improve scalability [56,57].

(2) **Large-scale image library storage.** DNA-FRSC first compresses each image into a compact latent feature representation and converts it into DNA sequences through quantization and sequence encoding. The encoding and recovery of different images can be performed independently. The experiments in Section 3.6 show that the compact feature representation reduces the numbers of sequences and bases compared with the baseline methods. For large-scale image repositories, DNA-FRSC could be combined with established file-indexing and random-access mechanisms for DNA storage [58] to support hierarchical organization, independent retrieval, and parallel processing of images.

(3) **Physical archival storage.** The sequence design of DNA-FRSC accounts for constraints relevant to DNA synthesis and sequencing, including GC balance, homopolymer length, and oligonucleotide length, providing a basis for integration with existing DNA synthesis, sequencing, and random-access workflows [59,60]. These properties support the application of DNA-FRSC to the physical archival storage of long-term cold data such as multimedia archives. Its end-to-end performance can be further evaluated through future experiments involving DNA synthesis and sequencing.

4. Conclusions

In summary, we propose DNA-FRSC, a robust DNA-based image storage system that integrates deep visual feature representations with error-resilient sequence coding. DNA-FRSC employs a SwinJSCC encoder to extract compact latent feature representations, which are linearly quantized and mapped to DNA sequences using a biologically constrained codebook. During decoding, two-stage sequence recovery based on top-k dynamic programming list decoding and equal-distance candidate refinement is combined with training-free feature imputation to recover corrupted or missing feature values. Based on this design, DNA-FRSC achieves a net information density of 6.41 bits/nt while maintaining robust reconstruction quality under high base error rates. Experimental results on the Kodak and General-100 datasets demonstrate the effectiveness of DNA-FRSC. At a 5% base error rate, DNA-FRSC reduces the MSE by more than 82.7%, improves the PSNR by at least 8.23 dB, and increases the SSIM, MS-SSIM, UQI, and FSIM by more than 83.7%, 50.4%, 17.6%, and 16.5%, respectively, compared with the best-performing baseline. This research provides a practical solution for the long-term preservation of large-scale cold image data, such as medical images, remote sensing images, and multimedia archives. In future work, we plan to incorporate data encryption and privacy-preserving mechanisms into the DNA storage pipeline to enhance the privacy and security of stored data, thereby facilitating more reliable real-world applications of DNA data storage.

CRediT authorship contribution statement

Zhuangzhuang Zang: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Shi Jia:** Writing – review & editing, Visualization, Supervision, Resources, Investigation. **Shiqian Jia:** Formal analysis. **Qifan Zhang:** Methodology. **Tianhao Zhang:** Validation. **Lianmeng Wu:** Formal analysis. **Penghao Wang:** Methodology. **Shufang Zhang:** Writing – review & editing, Supervision. **Huaming Wu:** Writing – review & editing, Supervision. **Jinlong Yu:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Emerging Frontiers Cultivation Program of Tianjin University Interdisciplinary Center (3800903075102) and the National Natural Science Foundation of China (62505217).

Data availability

Data will be made available on request.

References

[1] A. Doricchi, C.M. Platnich, A.L. Gimpel, E. Horn, M. Earle, G. Lanzavecchia, A. Cortajarena, L. Liz-Marzán, N. Liu, R. Heckel, et al., Emerging approaches to DNA data storage: challenges and prospects, ACS Nano 16 (11) (2022) 17552–17571, https://doi.org/10.1021/acsnano.2c06748.

[2] L. Organick, Y.-J. Chen, S.D. Ang, R. Lopez, X. Liu, K. Strauss, L. Ceze, Probing the physical limits of reliable DNA data retrieval, Nat. Commun. 11 (1) (2020) 616, https://doi.org/10.1038/s41467-020-14319-8.

[3] R.N. Grass, R. Heckel, M. Puddu, D. Paunescu, W.J. Stark, Robust chemical preservation of digital information on DNA in silica with error-correcting codes, Angew. Chem. Int. Ed. 54 (8) (2015) 2552–2555, https://doi.org/10.1002/anie.201411378.

[4] L.C. Meiser, P.L. Antkowiak, J. Koch, W.D. Chen, A.X. Kohll, W.J. Stark, R. Heckel, R.N. Grass, Reading and writing digital data in DNA, Nat. Protoc. 15 (2020) 86–101, https://doi.org/10.1038/s41596-019-0244-5.

[5] H. Lee, D.J. Wiegand, K. Griswold, S. Punthambaker, H. Chun, R.E. Khaman, G. M. Church, Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage, Nat. Commun. 11 (1) (2020) 5246, https://doi.org/10.1038/s41467-020-18681-5.

[6] M. Kubista, J.M. Andrade, M. Bengtsson, A. Forootan, J. Jonák, K. Lind, R. Sindelka, R. Sjöback, B. Sjögreen, L. Strömbom, et al., The real-time polymerase chain reaction, Mol. Aspects Med. 27 (2–3) (2006) 95–125, https://doi.org/10.1016/j.mam.2005.12.007.

[7] J. Shendure, S. Balasubramanian, G.M. Church, W. Gilbert, J. Rogers, J.A. Schloss, R.H. Waterston, DNA sequencing at 40: past, present and future, Nature 550 (7676) (2017) 345–353, https://doi.org/10.1038/nature24286.

[8] Y. Erlich, D. Zielinski, DNA Fountain enables a robust and efficient storage architecture, Science 355 (6328) (2017) 950–954, https://doi.org/10.1126/science.aaj2038.

[9] M. Welzel, P.M. Schwarz, H.E. Löchel, T. Kabdullayeva, S. Clemens, A. Becker, B. Freisleben, D. Heider, DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage, Nat. Commun. 14 (1) (2023) 628, https://doi.org/10.1038/s41467-023-36297-3.
[10] H. Choi, Y. Choi, J. Choi, A.C. Lee, H. Yeom, J. Hyun, T. Ryu, S. Kwon, Purification of multiplex oligonucleotide libraries by synthesis and selection, Nat. Biotechnol. 40 (1) (2022) 47–53, https://doi.org/10.1038/s41587-021-00988-3.
[11] A.J.X. Guo, M. Wei, Y. Dai, Y. Wei, P. Zhang, Disturbance-based discretization, differentiable IDS channel, and an IDS-correcting code for DNA-based storage, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2026, pp. 21423–21431, https://doi.org/10.1609/aaai.v40i26.39289.
[12] Y. Zhang, L. Kong, F. Wang, B. Li, C. Mao, D. Chen, K. Liu, C. Fan, H. Zhang, Information stored in nanoscale: encoding data in a single DNA strand with Base64, Nano Today 33 (2020) 100871, https://doi.org/10.1016/j.nantod.2020.100871.
[13] L. Ding, S. Wu, Z. Hou, A. Li, Y. Xu, H. Feng, W. Pan, J. Ruan, Improving error-correcting capability in DNA digital storage via soft-decision decoding, Natl. Sci. Rev. 11 (2) (2024) nwad229, https://doi.org/10.1093/nsr/nwad229.
[14] X. Lu, J. Jeong, J.-W. Kim, J.-S. No, H. Park, A. No, Error rate-based log-likelihood ratio processing for low-density parity-check codes in DNA storage, IEEE Access 8 (2020) 162892–162902, https://doi.org/10.1109/ACCESS.2020.3021700.
[15] Z. Yan, C. Liang, H. Wu, A segmented-edit error-correcting code with re-synchronization function for DNA-based storage systems, IEEE Trans. Emerg. Top. Comput. 11 (3) (2023) 605–618, https://doi.org/10.1109/TETC.2022.3225570.
[16] L. Organick, S.D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, et al., Random access in large-scale DNA data storage, Nat. Biotechnol. 36 (3) (2018) 242–248, https://doi.org/10.1038/nbt.4079.
[17] C. Ruan, L. Yang, R. Han, S. Gao, H. Wu, N. Ling, Robust DNA image storage decoding with residual CNN, IEEE International Symposium on Circuits and Systems (2024) 1–5, https://doi.org/10.1109/ISCAS58744.2024.10558127.
[18] H. Jung, S. Lee, J. Yoon, Q. Huang, S. Kim, Single image-based Gaussian splatting for 3D reconstruction of movable articulated objects, Adv. Eng. Informatics 70 (2026) 104191, https://doi.org/10.1016/j.aei.2025.104191.
[19] G. Qu, Z. Yan, X. Chen, H. Wu, DNA data storage for biomedical images using HELIX, Nat. Comput. Sci. 5 (5) (2025) 397–404, https://doi.org/10.1038/s43588-025-00793-x.
[20] B. Li, L. Ou, D.H.C. Du, IMG-DNA: approximate DNA storage for images, in: Proceedings of the 14th ACM International Conference on Systems and Storage, 2021, pp. 1–9, https://doi.org/10.1145/3456727.3463771.
[21] Y. Li, D.H.C. Du, L. Ou, B. Li, HL-DNA: a hybrid lossy/lossless encoding scheme to enhance DNA storage density and robustness for images, in: Proceedings of the IEEE 40th International Conference on Computer Design, 2022, pp. 434–442, https://doi.org/10.1109/ICCD56317.2022.00071.
[22] K. Wang, B. Cao, T. Ma, Y. Zhao, Y. Zheng, B. Wang, S. Zhou, Q. Zhang, Storing images in DNA via base128 encoding, J. Chem. Inf. Model. 64 (5) (2024) 1719–1729, https://doi.org/10.1021/acs.jcim.3c01592.
[23] B. Cao, K. Wang, L. Xie, J. Zhang, Y. Zhao, B. Wang, P. Zheng, PELMI: realize robust DNA image storage under general errors via parity encoding and local mean iteration, Brief. Bioinform. 25 (5) (2024) bbae463, https://doi.org/10.1093/bib/bbae463.
[24] Y. Zhou, K. Bi, Q. Xu, Q. Liu, X. Zhao, Q. Ge, Z. Lu, Ultrafast and accurate DNA storage and reading integrated system via microfluidic magnetic beads polymerase chain reaction, ACS Nano 19 (7) (2025) 7306–7316, https://doi.org/10.1021/acsnano.4c17817.
[25] Q. Xu, Y. Ma, Z. Lu, K. Bi, DP-ID: interleaving and denoising to improve the quality of DNA storage image, Interdiscip. Sci. 17 (2) (2025) 306–320, https://doi.org/10.1007/s12539-024-00671-6.
[26] Q. Xu, Y. Zhou, Q. Sun, X. Zhao, Z. Lu, K. Bi, DNA-CTMF: reconstruct high quality image from lossy DNA storage via pixel-base codebook and median filter, Synth. Syst. Biotechnol. 10 (3) (2025) 925–935, https://doi.org/10.1016/j.synbio.2025.04.015.
[27] L. Ceze, J. Nivala, K. Strauss, Molecular digital data storage using DNA, Nat. Rev. Genet. 20 (8) (2019) 456–466, https://doi.org/10.1038/s41576-019-0125-3.
[28] Y. Dong, F. Sun, Z. Ping, Q. Ouyang, L. Qian, DNA storage: research landscape and future prospects, Natl. Sci. Rev. 7 (6) (2020) 1092–1107, https://doi.org/10.1093/nsr/nwaa007.
[29] M. Dimopoulou, E. Gil San Antonio, M. Antonini, A JPEG-based image coding solution for data storage on DNA, in: Proceedings of the 29th European Signal Processing Conference, 2021, pp. 786–790, https://doi.org/10.23919/EUSIPCO54536.2021.9616020.
[30] M. Dimopoulou, M. Antonini, P. Barbry, R. Appuswamy, A biologically constrained encoding solution for long-term storage of images onto synthetic DNA, in: Proceedings of the 27th European Signal Processing Conference, 2019, pp. 1–5, https://doi.org/10.23919/EUSIPCO.2019.8902583.
[31] M. Dimopoulou, M. Antonini, Image storage in DNA using vector quantization, in: Proceedings of the 28th European Signal Processing Conference, 2021, pp. 516–520, https://doi.org/10.23919/Eusipco47968.2020.9287470.
[32] G. Franzese, Y. Yan, G. Serra, I. D'Onofrio, R. Appuswamy, P. Michiardi, Generative DNA: representation learning for DNA-based approximate image storage, in: Proceedings of the IEEE International Conference on Visual Communications and Image Processing, 2021, pp. 1–5, https://doi.org/10.1109/VCIP53242.2021.9675366.
[33] W. Wu, L. Xiang, Q. Liu, K. Yang, Deep joint source-channel coding for DNA image storage: a novel approach with enhanced error resilience and biological constraint optimization, IEEE Trans. Mol. Biol. Multi-Scale Commun. 9 (4) (2023) 461–471, https://doi.org/10.1109/TMBMC.2023.3331579.
[34] Y. Lu, Z. Zhang, J. Yang, C. Zhang, High fault-tolerant DNA image storage system based on VAE, IEEE Trans. Nanobioscience 24 (3) (2025) 339–347, https://doi.org/10.1109/TNB.2025.3544401.
[35] Y. Zheng, B. Cao, X. Zhang, S. Cui, B. Wang, Q. Zhang, DNA-QLC: an efficient and reliable image encoding scheme for DNA storage, BMC Genomics 25 (2024) 266, https://doi.org/10.1186/s12864-024-10178-5.
[36] Y. Zheng, X. Zhang, L. Sun, B. Cao, B. Wang, Q. Zhang, P. Zheng, INNSE: invertible neural network-based DNA image storage with self-correction encoding, Comput. Struct. Biotechnol. J. 27 (2025) 2492–2502, https://doi.org/10.1016/j.csbj.2025.06.003.
[37] X. Liu, Y. Zheng, X. Li, B. Wang, S. Zhou, B. Cao, P. Zheng, An end-to-end DNA storage coding method based on a low-complexity multiple biological constraints loss and RL-inspired differentiable solver, Expert Syst. Appl. 315 (2026) 131726, https://doi.org/10.1016/j.eswa.2026.131726.
[38] K. Yang, S. Wang, J. Dai, X. Qin, K. Niu, P. Zhang, SwinJSCC: taming Swin Transformer for deep joint source-channel coding, IEEE Trans. Cogn. Commun. Netw. 11 (1) (2025) 90–104, https://doi.org/10.1109/TCCN.2024.3424842.
[39] Q. Ge, M. Ren, T. Qi, C. Han, Y. Yuan, W. Chen, DNA diamond formulates a decomposable composite letter constellation model for DNA data storage, Nat. Commun. 17 (1) (2026) 1704, https://doi.org/10.1038/s41467-026-68861-y.
[40] J.H.D.B. Gervasio, H.C. Oliveira, A.G.C. Martins, J.B. Pesquero, B.M. Verona, N.N. P. Cerize, How close are we to storing data in DNA? Trends Biotechnol. 42 (2) (2024) 156–167, https://doi.org/10.1016/j.tibtech.2023.08.001.
[41] X. Zhao, J. Li, Q. Fan, J. Dai, Y. Long, R. Liu, J. Zhai, Q. Pan, Y. Li, Composite Hedges Nanopores codec system for rapid and portable DNA data readout with high INDEL correction, Nat. Commun. 15 (1) (2024) 9395, https://doi.org/10.1038/s41467-024-53455-3.
[42] F. Liang, X. Liu, A. Liu, C. Fan, Y. Li, Interleaved hybrid simulated annealing and pattern search for time-optimal feedrate scheduling in multi-axis CNC machining, Adv. Eng. Informatics 69 (2026) 104114, https://doi.org/10.1016/j.aei.2025.104114.
[43] B.L. Simmons, N.D. McDonald, N.G. Robinett, Assessment of enzymatically synthesized DNA for gene assembly, Front. Bioeng. Biotechnol. 11 (2023) 1208784, https://doi.org/10.3389/fbioe.2023.1208784.
[44] L. Song, F. Geng, Z.-Y. Gong, X. Chen, J. Tang, C. Gong, L. Zhou, R. Xia, M.-Z. Han, J.-Y. Xu, et al., Robust data storage in DNA by de Bruijn graph-based de novo strand assembly, Nat. Commun. 13 (1) (2022) 5361, https://doi.org/10.1038/s41467-022-33046-w.
[45] W.H. Press, J.A. Hawkins, S.K. Jones Jr., J.M. Schaub, I.J. Finkelstein, HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints, Proc. Natl. Acad. Sci. U.S.A. 117 (31) (2020) 18489–18496, https://doi.org/10.1073/pnas.2004821117.
[46] W. Chen, R. Qin, Q. Guo, J. Guo, Q. Ge, Y. Yuan, Approaching single-molecule assembly-free readout from medium-length encoded DNA, Nat. Commun. 16 (1) (2025) 10059, https://doi.org/10.1038/s41467-025-65004-7.
[47] Z. Yan, G. Qu, X. Chen, G. Zheng, H. Wu, D.N.A. StairLoop, enabling high-fidelity data recovery and robust error correction in DNA-based data storage, Nat. Commun. 16 (1) (2025) 9191, https://doi.org/10.1038/s41467-025-64230-3.
[48] Y. Su, L. Chu, W. Lin, X. Yao, P. Xu, W. Liu, A robust and efficient representation-based DNA storage architecture by deep learning, Small Methods 9 (3) (2025) e2400959, https://doi.org/10.1002/smtd.202400959.
[49] J. Bornholt, R. Lopez, D.M. Carmean, L. Ceze, G. Seelig, K. Strauss, Toward a DNA-based archival storage system, IEEE Micro 37 (3) (2017) 98–104, https://doi.org/10.1109/MM.2017.70.
[50] H.E. Löchel, M. Welzel, G. Hattab, A.-C. Hauschild, D. Heider, Fractal construction of constrained code words for DNA storage systems, Nucleic Acids Res. 50 (5) (2022) e30, https://doi.org/10.1093/nar/gkab1209.
[51] B. Polisky, P. Greene, D.E. Garfin, B.J. McCarthy, H.M. Goodman, H.W. Boyer, Specificity of substrate recognition by the EcoRI restriction endonuclease, Proc. Natl. Acad. Sci. U.S.A. 72 (9) (1975) 3310–3314, https://doi.org/10.1073/pnas.72.9.3310.
[52] Z. Ping, S. Chen, G. Zhou, X. Huang, S.J. Zhu, H. Zhang, H.H. Lee, Z. Lan, J. Cui, T. Chen, et al., Towards practical and robust DNA-based data archiving using the yin–yang codec system, Nat. Comput. Sci. 2 (4) (2022) 234–242, https://doi.org/10.1038/s43588-022-00231-2.
[53] K. Doroschak, K. Zhang, M. Queen, A. Mandyam, K. Strauss, L. Ceze, J. Nivala, Rapid and robust assembly and decoding of molecular tags with DNA-based nanopore signatures, Nat. Commun. 11 (1) (2020) 5454, https://doi.org/10.1038/s41467-020-19151-8.
[54] C. Dong, C.C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: Proceedings of the 14th European Conference on Computer Vision, 2016, pp. 391–407, https://doi.org/10.1007/978-3-319-46475-6_25.
[55] M. Yu, X. Tang, Z. Li, W. Wang, S. Wang, M. Li, Q. Yu, S. Xie, X. Zuo, C. Chen, High-throughput DNA synthesis for data storage, Chem. Soc. Rev. 53 (9) (2024) 4463–4489, https://doi.org/10.1039/d3cs00469d.
[56] F. Jabar, L.-T.-R. Busund, B. Ricciuti, M. Tafavvoghi, T.K. Kilvaer, D.J. Pinato, M. Pøhl, S. Andersen, T. Donnem, D.J. Kwiatkowski, et al., Fully automatic content-aware tiling pipeline for pathology whole slide images, Intelligence-Based Medicine 12 (2025) 100318, https://doi.org/10.1016/j.ibmed.2025.100318.
[57] A.G. Leal, T.Y. Aoyagi, A.G. Costa-Martins, D.T. de Souza, C.M.F. da Silva, E. T. Ueda, M.G.O. Parada, A.E. Feitosa, A. Fujita, Pantheon-DNA: versatile encoding-decoding system with integrated adaptive NGS preprocessing algorithms for DNA data storage, Comput. Struct. Biotechnol. J. 27 (2025) 3941–3951, https://doi.org/10.1016/j.csbj.2025.09.002.

[58] J. Li, C. Mao, S. Wang, X. Li, X. Luo, D. Wang, S. Zheng, J. Shao, R. Wang, C. Fan, et al., A compact cassette tape for DNA-based data storage, Sci. Adv. 11 (37) (2025) eady3406, https://doi.org/10.1126/sciadv.ady3406.
[59] C. Imburgia, L. Organick, K. Zhang, N. Cardozo, J. McBride, C. Bee, D. Wilde, G. Roote, S. Jorgensen, D. Ward, et al., Random access and semantic search in DNA data storage enabled by Cas9 and machine-guided design, Nat. Commun. 16 (2025) 6388, https://doi.org/10.1038/s41467-025-61264-5.
[60] B.W.A. Bögels, B.H. Nguyen, D. Ward, L. Gascoigne, D.P. Schrijver, A.-M.-M. Pistikou, A. Joesaar, S. Yang, I.K. Voets, W.J.M. Mulder, et al., DNA storage in thermoresponsive microcapsules for repeated random multiplexed data access, Nat. Nanotechnol. 18 (2023) 912–921, https://doi.org/10.1038/s41565-023-01377-4.