Next-Frame Decoding for Ultra-Low-Bitrate Image Compression with Video Diffusion Priors

Yunuo Chen[1], Chuqin Zhou[1], Jiangchuan Li[1], Xiaoyue Ling[1], Bing He[1], Jincheng Dai[2], Li Song[1], and Guo Lu[1⋆]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Beijing University of Posts and Telecommunications, Beijing, China

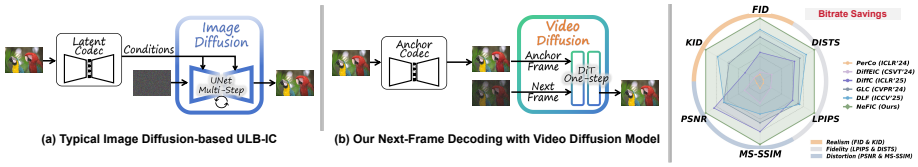

Fig. 1: (a) Some previous diffusion-based ULB-IC methods transmit latents that serve as conditions for image diffusion models during generative decoding. (b) Our method decodes a compact anchor frame and uses a video diffusion model to temporally evolve this anchor into the target image via next-frame prediction. Radar plot compares our model with existing generative codecs on the CLIC2020 test set.

Abstract. We present a novel paradigm for ultra-low-bitrate image compression (ULB-IC) that exploits the "temporal" evolution in generative image compression. Specifically, we define an explicit intermediate state during decoding: a compact anchor frame, which preserves the scene geometry and semantic layout while discarding high-frequency details. We then reinterpret generative decoding as a virtual temporal transition from this anchor to the final reconstructed image. To model this progression, we leverage a pretrained video diffusion model (VDM) as a temporal prior: the anchor frame serves as the initial frame and the original image as the target frame, transforming the decoding process into a next-frame prediction task. In contrast to image diffusion-based ULB-IC models, our decoding proceeds from a visible, semantically faithful anchor, which improves both fidelity and realism for perceptual image compression. Extensive experiments demonstrate that our method achieves superior rate-distortion performance. On the CLIC2020 test set, our method achieves over **50% bitrate savings** across LPIPS, DISTS, FID, and KID compared to DiffC, while also delivering a significant decoding speedup of up to $\times 5$. Code will be released at `https://github.com/UnoC-727/NeFIC`.

Keywords: Image Compression · Generative Compression

1 Introduction

The explosive growth of image data has sharply increased storage and transmission burdens, motivating steady progress in high-performance lossy image

⋆ Corresponding author. E-mail: `luguo2014@sjtu.edu.cn`.

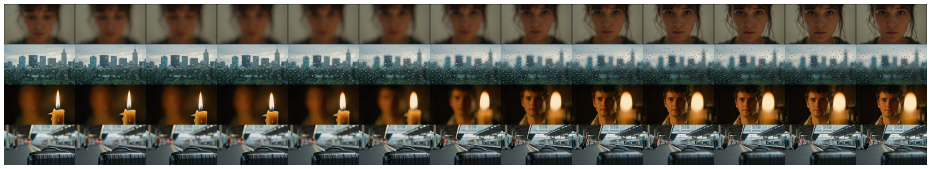

Fig. 2: Visualizations of Generated Refocus Videos. Pretrained VDMs (without finetuning) can generate diverse blur-to-sharp transitions. These foreground-background refocusing effects are common in daily videos and films used for training, enabling VDMs to learn such dynamics to restore details progressively.

compression. Learned image compression (LIC) has advanced rapidly under variational autoencoders (VAEs) and end-to-end optimization [4, 5]. Early models relied on CNN-based analysis-synthesis transforms [28, 65, 88], while later Transformer- and Mamba-based architectures further improved the rate-distortion (RD) performance [15–17, 20, 47, 52, 53, 59, 60, 71, 98, 106, 107]. Concurrently, the rise of Generative Adversarial Networks (GANs) [25] and diffusion models [30,78] has reframed ultra-low-bitrate image compression as a pivotal frontier. By leveraging powerful generative priors, diffusion-based codecs can deliver striking perceptual realism at extremely low bitrates. Most existing approaches [11, 54, 62] control the multi-step denoising process via control networks (Fig.1 (a)). However, the unclear influence of the conditional signals on the diffusion process, and the uncertainty of standard Gaussian initialization, together potentially induce semantic drift and weaken faithfulness to the source [13, 40, 62, 92]. In addition, the iterative denoising process incurs substantially higher computational cost and decoding latency than conventional VAE-style codecs. So far, the central challenge for ULB-IC is to achieve simultaneously: (i) strong perceptual realism, (ii) source-content faithfulness, and (iii) practical inference efficiency.

To achieve these targets, we propose a novel paradigm for ultra-low-bitrate image compression that explores the dynamics in generative image decoding. Our framework reinterprets single-image generative decoding as a virtual temporal process, where a compact anchor evolves to a high-fidelity reconstructed image. As illustrated in Fig. 2, pretrained VDMs can naturally refine blurry frames into sharper frames with richer details. We exploit this blur-to-sharp temporal prior by treating the anchor as a coarse initial frame and the target image as its sharp, detail-resolved final frame. For efficiency, we collapse the original multi-frame evolution into a two-frame refinement process.

Concretely, we define an explicit intermediate state for ULB-IC: a visible anchor that preserves scene geometry, semantic layout, and coarse appearance while discarding high-frequency details. As shown in Fig. 1 (b), this anchor is encoded and transmitted by a VAE-based Anchor Codec. At the decoder, we first decode the anchor from the bitstream and treat it as the initial state. The evolution from anchor to fine-grained reconstruction is then modeled as a video-like progression using a pretrained video diffusion model. We construct a virtual two-frame video, where the anchor is the first frame and the original image is the second. This setup casts generative decoding as a next-frame prediction task conditioned on the anchor, guided by the dynamic priors of the VDM. This yields

a reconstruction consistent with the anchor and perceptually aligned with the source without extra control networks, improving both realism and fidelity.

However, two key challenges arise when applying video diffusion models to image compression. (1) Domain Gap Between Video Generation and Image Decoding. Video diffusion models are primarily trained for gradual temporal evolution across dozens of frames, and emphasize scene transitions and object motions. In contrast, our goal is to achieve high-fidelity detail synthesis through a single-interval transition: a virtual two-frame "video" consisting solely of the compact anchor and the target image. (2) High Computational Cost of Iterative Denoising. State-of-the-art VDMs often require 50 iterative denoising steps and start from standard Gaussian noise, causing high latency and injecting ambiguity that can harm fidelity [13, 62]. Our aim is to turn multi-step video diffusion into a *one-step* generative decoder that preserves both fidelity and realism.

To address these challenges, we propose a two-stage training scheme that adapts video diffusion models for efficient and faithful ultra-low-bitrate image decoding: **Stage I: Next-Frame Decoding Adaptation.** We jointly finetune the pretrained VDM and the anchor codec to synthesize high-fidelity details from compact anchor frames while suppressing undesirable temporal artifacts, such as object motion and scene flicker. This adaptation aligns the generative priors of the VDM with the anchor-conditioned detail synthesis process, enabling accurate and perceptually faithful reconstruction through a single-interval temporal evolution. **Stage II: One-Step Generative Bypass.** To eliminate the latency introduced by iterative denoising, we collapse the multi-step diffusion process to one-step generation by establishing a latent bypass between anchor compression and video generation. Specifically, we condition the encoding of the anchor Compression-VAE on generative latents obtained from the Video-VAE. We then introduce a Bypass Refinement module, which is a learned mapping from the compression latent space to the generative latent space. This bypass regularizes and aligns the latent spaces of the two VAEs, reducing generative uncertainty and enabling faithful one-step decoding.

Our contributions can be summarized as follows.

1. **A novel ULB-IC paradigm:** We reinterpret generative image decoding as a virtual temporal transition from a compact anchor into a faithful and realistic reconstruction.
2. **Efficient generative decoding:** Our two-stage training pipeline adapts VDMs for compression-aware next-frame prediction while enabling one-step generation.
3. **State-of-the-art performance:** Extensive experiments demonstrate that our model achieves superior perceptual realism and fidelity with relatively faster decoding.

2 Related Work

2.1 Generative Image Compression

VAE- and GAN-based Codecs. Previous VAE-based learned image compression models [18, 35, 37, 38, 50, 66, 86, 102, 105] mainly optimize objective metrics,

such as PSNR and MS-SSIM. Agustsson *et al.* [3] and HiFiC [64] pioneered GAN-based methods for low-bitrate image compression and aim to improve realism and fidelity. MR [1] connects generative and non-generative compression. MS-ILLM [68] introduces a local likelihood model to improve realism. EGIC [42] uses discriminators conditioned on segmentation for stronger adversarial training. Recent studies [36,63,94] use vector quantization for extreme semantic compression. TACO [45] proposes a text-guided encoder to improve fidelity.

Diffusion-based Codecs. Recent works [11,12,22–24,26,33,39,43,46,55,57,61,74,79,81,89–93,95,103,104] use generative priors from large pretrained controllable image diffusion models [67,97,99]. Building on DDPM [30] and LDM [75], they show better perceptual realism than conventional GAN-based methods. [62] introduces a privileged decoder to correct the denoising process. DiffC [84] achieves zero-shot compression using reverse-channel coding. ResULIC [40] adds residual coding for semantic alignment. PICD [90] proposes a model for both screen content and natural images. Most models [11,23,24,33] based on standard image diffusion models mainly inject conditions through channel concatenation, which may not be a principled way to exploit context.

2.2 Video Diffusion Model

Text-to-video (T2V) synthesis has progressed rapidly. Early work with large-scale Transformers, such as CogVideo [32] and Phenaki [83], showed that the approach was feasible with great potential. Later advanced diffusion models [9, 27,31,51,77] brought a clear jump in generation quality. With the development in the Diffusion Transformer (DiT) architecture, LinGen [87], CogVideoX [96] and OpenAI's Sora [10] reached a new level for text-to-video generation. Recently, WAN [85], Tencent's HunyuanVideo [41] and OpenAI's Sora-2 [69] together mark the current frontier of modern video generation.

3 Methodology

3.1 Preliminary: In-Context Learning in VDMs.

We build upon CogVideoX-1.5 [96], a text-to-video diffusion model that generates frame sequences by progressively denoising latent noise conditioned on a text prompt. The architecture comprises: (1) a 3D causal Video-VAE for joint spatial-temporal compression whose encoder maps input videos to latents at $1/8$ height and width; (2) a text encoder [73] producing prompt embeddings; and (3) a DiT-based [70] noise-prediction transformer. The transformer operates on the concatenation of text and video tokens with 3D attention and applies 3D rotary position embeddings [80] to queries and keys for spatial-temporal modeling.

As a video foundation model trained on large-scale continuous visual data, CogVideoX exhibits emergent behaviors analogous to LLMs, including unified generation and in-context learning [14, 56]. Given preceding frames $\{c_1, c_2, \ldots\}$ as context, it autoregressively predicts subsequent frames $\{y_1, y_2, \ldots\}$ by capturing long-range spatio-temporal dependencies. This next-token-style formulation supports continued pretraining and task-specific finetuning, enabling coherent future-frame synthesis that remains consistent with the provided visual context.

3.2 Reframing Generative Image Compression

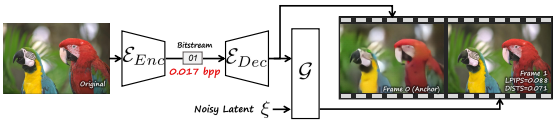

Fig. 3: An illustration of our next-frame prediction paradigm for generative compression.

As discussed in Sec. 1, we focus on modeling the transition from a compact anchor to a high-fidelity reconstruction with video diffusion priors. We adopt a two-frame abstraction: the anchor image serves as the first frame for video generation, and the target image as the second. Accordingly, we reinterpret the anchor as the conditional frame and task the VDM to predict the next frame that restores realistic details. As shown in Fig. 3, the encoding and generative decoding process for image x can be defined as

$$\hat{x} = (\mathcal{G} \circ \mathcal{E}_{Dec})(l; \xi), \quad l = Q\big(\mathcal{E}_{Enc}(x)\big), \tag{1}$$

where $\mathcal{E}_{Enc}$ is the encoder of the Anchor Codec and Q denotes quantization. ξ represents the noisy latent input of the target frame and l refers to the transmitted latent. $\mathcal{E}_{Dec}$ denotes the anchor decoder, and $\mathcal{G}$ is a VDM adapted for single-interval photorealistic next-frame prediction.

In the following sections, we introduce our two-stage training scheme designed to bridge the domain gap between video generation and image decoding, and to mitigate the computational overhead of multi-step diffusion sampling.

3.3 Why is a VDM preferred over an image diffusion model (IDM)?

1. Temporal Prior: VDM's built-in transition capability. Our key insight is that pretrained VDMs can naturally generate **temporal defocus-to-focus transitions**: starting from vague frames, VDMs generate high-frequency details in the subsequent frames. As shown in Fig. 4, details of parrot feathers are temporally refined.

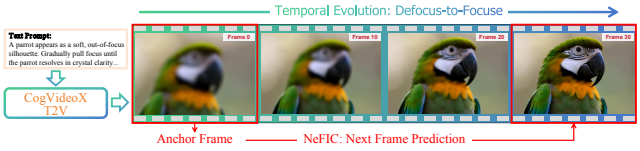

Fig. 4: Defocus-to-Focus Temporal Transition Prior. These generated frames show that VDM generates natural textures where appropriate (e.g., parrot feathers). NeFIC exploits this property and collapses multi-frame transitions into next-frame prediction.

This behavior is not incidental. Trained on large-scale video data, VDMs internalize world knowledge and inter-object relations. Since video corpora contain abundant "defocus-to-focus" temporal patterns, VDMs learn to temporally refine coarse structure into detail-resolved imagery.

We repurpose this generative prior: heavily compressed anchors represent the early blurry frames, while the final reconstructions correspond to the clear frames at the end. This is significantly different from predicting residuals [23,24]. **NeFIC reinterprets the blur-to-clarity temporal transition (rather**

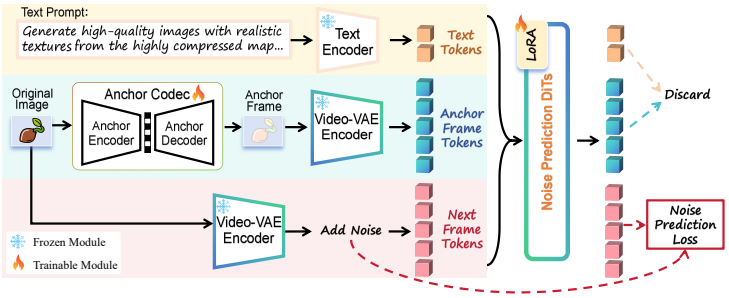

Fig. 5: Overview of Training Stage I. The input to the DiT blocks is a concatenation of three types of tokens: (1) text embeddings, (2) tokens from the anchor frame, and (3) noised tokens from the target frame. During training, only the output corresponding to the noised target tokens is supervised using a noise prediction loss, while the other two segments are discarded.

than motion) in VDM as generative decoding. This pretrained transition prior is not internalized by IDMs.

2. Architectural Advantages: Leveraging VDM's 3D spatiotemporal structure and in-context learning capability.

Inspired by in-context learning in LLMs, we reformulate VDM-based generative decoding as conditional next-frame prediction. We fully exploit the pretrained **3D Attention** and **3D RoPE**: NeFIC performs frame/token-level concatenation to treat the anchor frame as a preceding context.

IDM-based codecs typically inject conditions by concatenating the conditional latent with the noisy latent along the channel dimension [11, 23, 24, 33]. Thus, conditional interactions occur only through implicit cross-channel feature mixing. By contrast, token concatenation coupled with 3D Attention enables explicit and adaptive retrieval: each noisy token can directly attend to and query any anchor token. Meanwhile, 3D RoPE imposes a correspondence prior along the temporal axis, biasing attention toward spatially aligned regions across frames. Thus, NeFIC synthesizes high-frequency details as structurally anchored refinements of the coarse layout. **NeFIC inherits VDM's 3D inductive bias as a more principled condition-injection pathway. Such effective conditional generation and in-context learning capabilities are not supported by standard IDMs.**

3.4 Stage I: Next-Frame Decoding Adaptation

We first cast generative decoding as a single-interval next-frame prediction in the latent space of a frozen Video-VAE (Fig. 5). This stage focuses on synthesizing fine-grained details while suppressing undesirable temporal artifacts, such as object motion and scene flicker. During training, we adapt the DiT [70] blocks to condition on the decoded anchor, and to predict the clean latent of the target frame from its noised latent at diffusion step $t \in \{1, \ldots, T\}$ with schedule (α_t, σ_t) satisfying $\alpha_t^2 + \sigma_t^2 = 1$. For efficiency, we finetune only the Anchor Codec and the attention projections (W_q, W_k, W_v, W_{out}) via LoRA [34], while keeping

the Video-VAE and the text encoder frozen. The video diffusion priors provide photorealistic texture and detail synthesis without auxiliary control networks.

Tokenization and Conditioning. Let E_{text} denote the frozen text encoder, E_{Vid} the frozen Video-VAE encoder, f_θ the DiT blocks, and $\oplus$ token-sequence concatenation. The model consumes three sequences:

1. **Text tokens:** $\mathbf{z}_{\text{text}} = E_{\text{text}}(p)$, where p is a universal prompt ("generate a high-quality image with realistic texture..."). Our framework eliminates the need for detailed scene descriptions, as conditions and semantics are fully transmitted via anchor frames. The text prompt serves only as guidance of the transition behavior.
2. **Anchor tokens:** $\mathbf{z}_{\text{anchor}} = E_{\text{Vid}}\big(\mathcal{E}_{Dec}(l)\big)$, obtained by encoding the decoded anchor $\mathcal{E}_{Dec}(l)$, which serves as the initial-frame condition.
3. **Noisy target tokens:** $\mathbf{z}_t$, formed by encoding the original image x into $\mathbf{z}_0 = E_{\text{Vid}}(x)$ and adding noise:

$$\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

Noise Prediction and Loss. We condition on $\mathbf{c} = \mathbf{z}_{\text{text}} \oplus \mathbf{z}_{\text{anchor}}$ and feed $\mathbf{c} \oplus \mathbf{z}_t$ into f_θ at step t. Using v-prediction parameterization [76], we reconstruct target latent as

$$\mathbf{v}_\theta^{(t)}(\mathbf{z}_t, \mathbf{c}) = f_\theta(\mathbf{z}_t, \mathbf{c}, t), \tag{3}$$

$$\hat{\mathbf{z}}_0 = \alpha_t \mathbf{z}_t - \sigma_t \, \mathbf{v}_\theta^{(t)}(\mathbf{z}_t, \mathbf{c}) \tag{4}$$

The output tokens corresponding to the text and anchor segments are discarded. Only the output token segment aligned with the target tokens is supervised as

$$\mathcal{L}_{noise} \;=\; \|\hat{\mathbf{z}}_0 - \mathbf{z}_0\|_2^2 . \tag{5}$$

Auxiliary Anchor Loss. To keep the decoded anchor $x_{\text{anchor}} = \mathcal{E}_{Dec}(l)$ within the training distribution of the Video-VAE and maintain its conditioning strength, we add an auxiliary reconstruction term on x_{anchor}:

$$\begin{aligned} \mathcal{L}_{\text{aux}} = \lambda_{\text{MSE}} \left\| x - x_{\text{anchor}} \right\|_2^2 \\ + \lambda_{\text{LPIPS}} \, \text{LPIPS}\big(x, \, x_{\text{anchor}}\big), \end{aligned} \tag{6}$$

with $\lambda_{\text{MSE}} = 5$ and $\lambda_{\text{LPIPS}} = 1$. The MSE term enforces pixel-level fidelity so the anchor remains stably encodable by E_{Vid}; the LPIPS term [100] preserves perceptual similarity and high-level semantics, ensuring sufficient and faithful semantic information for conditioning.

Overall Objective. The Stage I objective combines noise prediction, anchor regularization, and a bitrate penalty:

$$\mathcal{L}_{\text{stageI}} = \mathcal{L}_{\text{noise}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}} + \lambda_R \, R, \tag{7}$$

where R is the bitrate loss of compressing the anchor frame and λ_R trades rate against reconstruction quality.

This stage encourages VDM LoRAs to learn the conditional dependence in the compact anchor $\mathbf{z}_{\text{anchor}}$, thereby modeling the next-frame transition and recovering a fine-grained latent $\hat{\mathbf{z}}_0$ under multi-step diffusion.

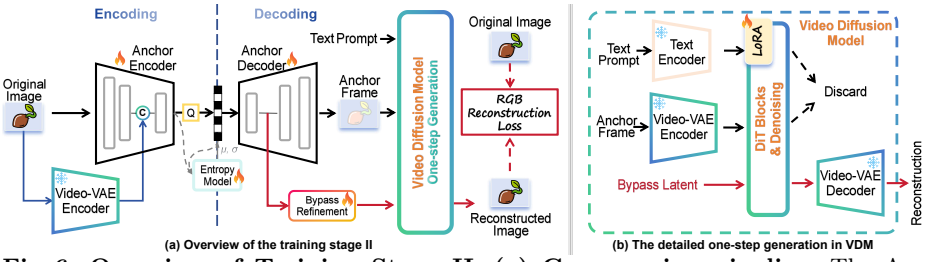

Fig. 6: Overview of Training Stage II. (a) Compression pipeline. The Anchor Encoder is conditioned on Video-VAE latents of the original image and produces entropy-coded anchor latents. The Anchor Decoder reconstructs the anchor frame and a bypass latent. **(b) One-step denoising.** A single-step video diffusion model consumes text embeddings, anchor tokens, and the bypass latent to predict the target latent, which the Video-VAE decoder converts to final reconstruction.

3.5 Stage II: One-Step Generative Bypass

To collapse multi-step denoising into a single step, we initialize the target latent with an informative latent instead of Gaussian noise, which helps to reduce ambiguity and semantic drift [13]. Concretely, we introduce a latent bypass that couples the Anchor Codec with VDM through the two strategies detailed below.

Conditional Anchor Encoding. Due to the latent-space gap between the compression-VAE and generative-VAE [101], raw features from the anchor codec are suboptimal for diffusion. Inspired by conditional coding [48,49,88], we condition and align the anchor encoder with the Video-VAE. The frozen E_{Vid} encodes the original image x to latent $\mathbf{z}_0 = E_{\text{Vid}}(x)$. As shown in Fig. 6, at the same spatial scale the Anchor encoder produces a feature map $\mathbf{h}_{\text{enc}}$. We concatenate them along channels $\mathbf{h}_{\text{cond}} = [\mathbf{h}_{\text{enc}}; \mathbf{z}_0]$, then apply the codec transforms and entropy modeling to $\mathbf{h}_{\text{cond}}$. This biases the Anchor Codec to preserve more scene information and layout semantics consistent with the diffusion prior. As shown in Fig. 7, removing the conditional encoding for the anchor ("w/o Cond-Enc") leads to semantic drift, such as the appearance of extra teeth in the mouth.

Fig. 7: Reconstruction visualizations of three variants.

Bypass Refinement. We decode a bypass latent from the anchor latent to replace the pure noise at inference. Let $\mathbf{h}_{\text{dec}}$ be the decoded feature from $\mathcal{E}_{Dec}$. A Transformer-based Semantic Bypass Refinement T maps $\mathbf{z}_{\text{bypass}} = T(\mathbf{h}_{\text{dec}})$, which serves as the initial noisy latent for one-step generation, carrying semantics aligned to the generation manifold. As shown in Fig. 7, the absence of bypass refinement causes semantic information to become ambiguous and the reconstruction of teeth to appear vague. This occurs because the gap between the compression and generation latent spaces exacerbates the denoising difficulty for DiTs.

One-step generation. As in Stage I, we form $\mathbf{c} = \mathbf{z}_{\text{text}} \oplus \mathbf{z}_{\text{anchor}}$, where $\mathbf{z}_{\text{text}} = E_{\text{text}}(p)$ and $\mathbf{z}_{\text{anchor}} = E_{\text{Vid}}(x_{\text{anchor}})$. To enable one-step generation, we model

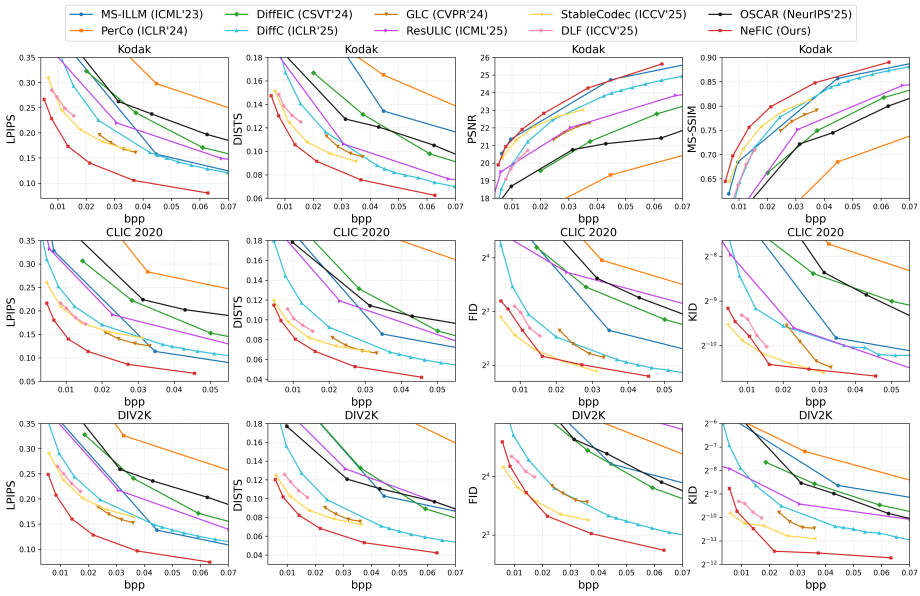

Fig. 8: Rate-distortion and rate-perception curves of recent advanced methods.

the bypass latent as a partially denoised state and set the denoising step $t^\star$ to half of the maximum step value, specifically 500. Using v-prediction, the DiT predicts noise and reconstructs the clean latent in a single step

$$\mathbf{v}_\theta(\mathbf{z}_{\text{bypass}}, \mathbf{c}, t^\star) = f_\theta(\mathbf{z}_{\text{bypass}}, \mathbf{c}, t^\star), \tag{8}$$

$$\hat{\mathbf{z}}_0 = \alpha_{t^\star}\, \mathbf{z}_{\text{bypass}} - \sigma_{t^\star}\, \mathbf{v}_\theta(\mathbf{z}_{\text{bypass}}, \mathbf{c}, t^\star), \tag{9}$$

where $(\alpha_{t^\star}, \sigma_{t^\star})$ is the fixed one-step schedule. The frozen Video-VAE decoder yields reconstructed image $\hat{x} = D_{\text{Vid}}(\hat{\mathbf{z}}_0)$.

CLIP-based GAN loss. To leverage semantic priors from visual foundation models, we adopt a CLIP-based [72] discriminator for adversarial training. Following [44], we train a lightweight classifier head on top while keeping the CLIP backbone frozen. The adversarial loss $\mathcal{L}_{\text{GAN}}$ helps narrow the distribution gap between x and $\hat{x}$ [64, 101].

RGB reconstruction loss. The reconstruction loss combines the GAN term with pixelwise and perceptual criteria:

$$\begin{aligned}\mathcal{L}_{\text{RGB}} = \lambda_{\text{GAN}}\, \mathcal{L}_{GAN} + \lambda_{\text{MSE}}\, \|\hat{x} - x\|_2^2 \\ + \lambda_{\text{LPIPS}}\, \text{LPIPS}\big(\hat{x},\, x\big),\end{aligned} \tag{10}$$

with $\lambda_{\text{GAN}} = 1$, $\lambda_{\text{MSE}} = 2.5$, $\lambda_{\text{LPIPS}} = 0.5$.

Overall objective. We retain the auxiliary anchor loss $\mathcal{L}_{\text{aux}}$ and the bitrate loss R from entropy model as Stage I. The Stage II overall training objective is

$$\mathcal{L}_{\text{stageII}} = \mathcal{L}_{\text{RGB}} + \lambda_{\text{aux}}\, \mathcal{L}_{\text{aux}} + \lambda_R\, R, \tag{11}$$

In Stage II, we update parameters of bypass refinement module T, the Anchor Codec (i.e., $\mathcal{E}_{Enc}, \mathcal{E}_{Dec}$), and the LoRA adapters of f_θ. Other parts remain frozen.

Table 1: Detailed BD-Rate of different ultra-low-bitrate image compression models on Kodak, DIV2K and CLIC2020 test set.

Model	Kodak				DIV2K				CLIC2020 Test			
	LPIPS	DISTS	PSNR	MS-SSIM	LPIPS	DISTS	FID	KID	LPIPS	DISTS	FID	KID
MS-ILLM [68]	10.76	203.21	-32.27	-19.59	35.40	104.69	143.45	173.20	28.71	75.29	94.68	747.20
PerCo [11]	112.61	202.54	303.82	146.77	266.60	409.65	176.61	170.99	326.11	452.83	128.88	54.05
DiffEIC [54]	61.15	84.54	121.28	77.12	106.71	134.88	162.47	141.98	131.51	158.89	73.68	38.45
GLC [36]	-16.86	-6.07	55.25	27.76	-17.79	-15.28	3.25	-21.90	-18.12	-18.72	-32.22	-36.10
DiffC [84]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ResULIC [40]	-9.28	-6.83	7.60	24.38	38.02	68.73	269.67	0.57	42.08	64.07	32.80	-62.06
StableC [101]	-42.19	-41.20	-22.05	-20.82	-27.31	-47.56	-52.06	-70.08	-27.39	-48.35	-70.67	-73.46
DLF [92]	-42.31	-36.97	7.77	-1.25	-20.07	-33.97	-25.71	-34.37	-22.51	-36.44	-60.82	-67.93
OneDC [93]	-44.31	-28.60	12.47	0.37	-48.51	-36.90	-36.12	-	-54.55	-46.49	-43.06	-
OSCAR [26]	35.16	28.74	95.69	51.56	143.25	43.65	171.66	-	196.58	97.54	352.81	-
NeFIC (Ours)	-64.03	-50.84	-34.26	-33.88	-61.71	-56.94	-49.94	-71.20	-64.09	-57.96	-65.76	-75.94

4 Experiments

4.1 Model and Training.

Experimental Settings. We use CogVideoX-1.5 as the video diffusion backbone for our model and pretrain a VAE-based Anchor Codec. Details are in the supplementary material. Our models are trained on Flickr2W [58] with random square crops whose side length is uniformly sampled from 512 to 768 pixels. We set the batch size to 8 and train for 10k steps for stage I and 50k steps for stage II. In stage I, the learning rate is 1e-4. In stage II, we introduce the CLIP-based GAN loss at 35k steps and decay the learning rate to 1e-5 at 45k steps. The entire 2-stage training takes about 3 days. All models are optimized with AdamW. λ_{aux} is fixed at 0.1, λ_{R} is chosen from {5, 3, 1.7, 1.0, 0.5, 0.25}.

Evaluation. We evaluate all the methods on Kodak [21], the CLIC2020 test set [82], and DIV2K [2]. We report objective distortion (PSNR, MS-SSIM), perceptual fidelity metrics (LPIPS [100], DISTS [19]), and perceptual realism metrics (FID [29], KID [6]). Bit-rate efficiency is measured using Bjøntegaard-Delta rate (BD-rate) [7]. We compute PSNR/MS-SSIM/LPIPS/DISTS on full-resolution images. Following HiFiC [64], FID and KID are computed on 256×256 patches for CLIC2020 and DIV2K. We omit FID/KID on Kodak because its small size (24 images) precludes reliable distribution estimation.

Baselines. We take generative image codecs as baselines, including the VAE- and GAN-based generative codec MS-ILLM [68], GLC [36], DLF [92] and recent diffusion-based methods: PerCo [11], DiffEIC [54], DiffC [84], ResULIC [40], OSCAR [26] and StableCodec [101].

4.2 Main Results

Quantitative Evaluation Perceptual Metrics. In Fig. 8, we plot rate-perception curves across various datasets. The corresponding BD-rate results are summarized in Tab. 1 (lower is better; negative indicates bitrate savings), where we report BD-Rate on DIV2K with DiffC as the baseline. Compared with the earlier diffusion-based codec PerCo [11], our model reduces bitrate on

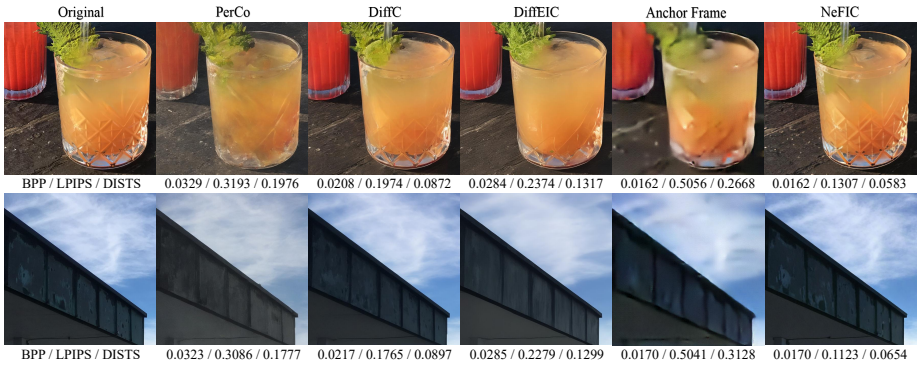

Fig. 9: Qualitative examples on the CLIC2020 dataset. Zoom in for better view. More comparisons are in supplementary.

DIV2K by 89.82%, 91.61%, 87.94%, and 93.25% on LPIPS, DISTS, FID, and KID, respectively. Even against recent SOTA method DLF [92] (ICCV'2025), our method still achieves 51.92%, 48.16%, 31.82%, and 33.47% BD-rate savings on the same four metrics. Across datasets and metrics, our approach consistently attains substantial bitrate reductions at matched perceptual quality, indicating that it preserves both perceptual realism and image fidelity.

Distortion Metrics. Tab. 1 also reports BD-rate comparisons on PSNR and MS-SSIM for Kodak; the corresponding RD curves are shown in Fig. 8. Whereas many generative compression methods sacrifice distortion-oriented objectives for stronger generative ability, our method improves both perceptual and distortion metrics. On Kodak, we outperform recent SOTA model StableCodec [101] (ICCV'2025) by 18.27% BD-rate on PSNR and 22.95% on MS-SSIM. Compared to GAN-based method GLC [36] (CVPR'24), our model also can achieve 57.08% and 46.91% bit savings for PSNR and MS-SSIM. These advantages in objective metrics attest to the faithfulness of our reconstructions. Detailed PSNR and MS-SSIM results for DIV2K and CLIC2020 are provided in the supplementary.

The consistent gains in realism, fidelity, and distortion metrics stem from our evolutionary compression pipeline. The anchor frame preserves coarse yet faithful structure and key semantics, while the video diffusion prior, with its strong temporal consistency, refines them into realistic fine textures. This yields a tight latent-pixel dual-domain coupling: the bypass latent provides a good initialization in latent space, and the anchor frame offers a pixel-domain reference that fully exploits the pretrained video consistency.

Qualitative Evaluation. We present qualitative comparisons of our NeFIC, our anchor frame, and previous diffusion-based ULB-IC models in Fig. 9. At ultra-low bit rates, NeFIC produces details that are more visually consistent with and faithful to the original image. For instance, our model accurately reconstructs the regular ridges on glassware, as well as irregular corrosion patterns on walls. Other models fail to recover those details accurately. Despite using higher bitrates than our approach, these competing models perform worse for LPIPS,

Table 2: Runtime (s) on Kodak and DIV2K, and BD-Rate.

Model	Kodak (s)		DIV2K (s)		BD-Rate DIV2K (LPIPS,%)
	Enc.	Dec.	Enc.	Dec.	
PerCo [11]	0.39	1.92	0.62	38.29	266.60
DiffEIC [54]	0.23	4.82	2.42	58.46	106.71
DiffC [84]	0.6-9.1	2.9-8.4	3.4-47.2	15.8-37.7	0.00
StableCodec [101]	0.11	0.14	18.86	40.63	−27.31
w/o Stage II	0.16	30.61	0.51	296.02	−47.01
NeFIC(Ours)	0.17	0.88	0.58	6.83	−61.71

DISTS, and visual fidelity. As we described, our anchor frame preserves only the scene geometry, semantic layout, and coarse appearance, while providing a powerful temporal initialization for next-frame decoding.

4.3 The Anchor and Bypass

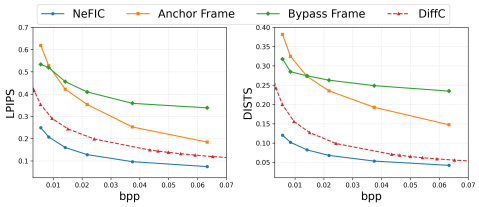

Fig. 10: Comparisons of different frames.

We plot the rate-perceptual curves in Fig. 10 to compare the perceptual quality of the anchor and bypass frames (directly decoding bypass latent by Video-VAE) with the final reconstructed images. As shown, both the anchor and bypass frames exhibit lower perceptual quality, and the significant gap between their curves and the NeFIC curve highlights the effectiveness of our next-frame prediction strategy in achieving superior reconstruction fidelity with VDM. More visualizations are provided in supplementary material.

4.4 Efficiency Evaluation.

We assess runtime efficiency as the average per-image encoding (Enc.) and decoding (Dec.) time on an A100 GPU on the Kodak and DIV2K datasets (Tab. 2). NeFIC achieves 0.17/0.88 s (Enc./Dec.) on Kodak and 0.58/6.83 s on DIV2K, yielding 5.5-8.6× faster decoding and 1.35-4.17× faster encoding than the diffusion-based DiffEIC. On high-resolution DIV2K, NeFIC outperforms StableCodec by 32.5× in encoding and 6.0× in decoding due to avoiding costly tiling. These gains arise from single-interval prediction and one-step generation, which yield smoothly scaling latency and avoid the sharp slowdowns of multi-step diffusion or heavy tiling. Removing the Stage II training for one-step generation increases decoding time to 30.61s for Kodak and 296.02s for DIV2K in a 50-step denoising setting. These delays are unacceptable and underscore the effectiveness and necessity of our generative bypass.

4.5 Ablation Studies.

We ablate the key design components of NeFIC. Unless otherwise noted, all variants are evaluated on DIV2K using LPIPS, DISTS, FID, and KID. We report BD-Rate (%, lower is better) on DIV2K with DiffC as the baseline.

Stage I (Interval Adaptation). Training directly with Stage II is possible but suboptimal. CogVideoX is pretrained on long video sequences (> 25 frames) with diverse scene transitions and motions [96]. If we skip Stage I, the LoRA must

Table 3: Ablation BD-rate results on different components.

Model variants	LPIPS	DISTS	FID	KID
w/o Interval Adaptation	-52.32	-44.71	-40.32	-62.88
w/o One-step Bypass	-47.01	-42.03	-53.72	-77.91
w/o Conditional Encoding	-56.81	-52.34	-42.01	-62.46
w/o Bypass Refinement	-48.01	-47.19	-39.36	-57.91
w/o Aux Anchor Loss	-19.78	-20.46	-31.07	-40.74
w/o CLIP-based GAN	-63.09	-64.02	-30.31	-30.07
NeFIC (Ours)	-61.71	-56.94	-49.94	-71.20

Table 4: Ablation BD-rate results on LoRA Rank and Alpha.

LoRA Rank/Alpha	LPIPS	DISTS	FID	KID
64/64	-44.18	-41.01	-59.91	-82.99
128/128	-54.82	-51.36	-58.34	-80.26
256/256	-61.71	-56.94	-49.94	-71.20

learn compression-aware single-interval prediction and single-step generation at the same time, which makes optimization harder and more unstable. In Tab. 3, removing Stage I ("w/o Interval Adaptation") degrades BD-Rate on all four metrics by 8-10%.

Stage II (One-step Bypass). With Stage I alone, the model supports next-frame decoding for ultra-low bitrate compression but still requires multi-step denoising. Adding Stage II enables end-to-end training and further improves fidelity, at the cost of slightly lower realism. Compared with the Stage I-only setting ("w/o One-step Bypass"), Stage II improves LPIPS and DISTS BD-Rate by about 15%, while FID and KID worsen by roughly 4% and 7%, indicating a mild trade-off between perceptual fidelity and realism.

Conditional Encoding. The semantic latent bypass consists of *Conditional Encoding* and *Bypass Refinement*. Removing Conditional Encoding hampers the alignment of the bypass latent with the VDM generation space and its cooperation with the frozen Video-VAE decoding, yielding BD-rate drops about 4.9% for LPIPS and 7.9% for FID (Tab. 3).

Bypass Refinement. Ablating the bypass refinement causes a larger gap: BD-Rate degrades by ∼10-14% on LPIPS/FID/KID. This suggests the raw VAE-decoded latent still mainly supports anchor-frame RGB reconstruction, while refinement better adapts it to the generation space with stronger semantics.

Auxiliary Anchor Loss. Without the auxiliary anchor loss, the anchor frame diverges from the natural image domain, exhibiting significant semantic and color drift. This misalignment with the training distribution of Video-VAE weakens the conditioning for next-frame generation for single-interval, one-step decoding. As shown in Tab. 3, the absence of this loss leads to a substantial degradation in perceptual compression performance, with BD-Rate degrading to −19.78% for LPIPS and −40.74% for KID.

CLIP-guided GAN Loss. Ablating the CLIP-based adversarial loss slightly improves LPIPS/DISTS but severely harms realism (FID/KID worsen by about 20 and 41 BD-Rate points). Adversarial training is therefore important for preserving generation capability and generative realism.

Table 5: Paradigm and Foundation Model Comparison.

Generation Model	LPIPS	DISTS	FID	KID
(1) Flux-dev-12B [8] (w/o anchor)	-32.01	-33.21	-53.97	-72.87
(2) Flux-dev-12B [8] (channel concatenation)	-35.77	-35.13	-55.92	-72.09
(3) Flux-dev-12B [8] (token concatenation)	-42.18	-42.31	-53.98	-73.19
CogVideoX-5B (Ours)	-61.71	-56.94	-49.94	-71.20

LoRA Settings. Tab. 4 shows that increasing rank/alpha from 64/64 to 256/256 consistently improves LPIPS ($-44.18\% \rightarrow -61.71\%$) and DISTS ($-41.01\% \rightarrow -56.94\%$) while degrading FID and KID by about 10% in BD-rate. We adopt 256/256 as the default setting for LoRA adapters in NeFIC.

Comparison with Image-Diffusion Paradigm. To highlight the benefit of anchor-guided next-frame decoding with a VDM, we compare against a text-to-image (T2I) diffusion codec baseline that replaces our CogVideoX-1.5-5B decoder with Flux-dev-12B [8]. We fine-tune this larger model under three variants: (1) decode the bypass latent in a single step without the anchor branch; (2) concatenate the anchor latent along the channel dimension; (3) concatenate anchor latent as tokens (with 2D attention only, since 3D RoPE and attention is unsupported by IDM). As shown in Tab. 5, the BD-rate results indicate that the anchor condition is not effectively exploited in variants (1)–(3). The T2I baseline benefits from both larger model capacity and the fact that state-of-the-art T2I models often outperform T2V models in visual realism. Our method attains substantially lower BD-rates on LPIPS and DISTS, showing higher perceptual fidelity and better structural consistency with the source image.

We attribute these gains to the explicit anchor and virtual temporal evolution, which together enforce a strong latent–pixel dual-domain coupling. Concretely, the bypass latent provides a strong initialization in latent space, while the decoded anchor frame supplies a pixel-domain reference and leverages the pretrained VDM's learned temporal consistency to mitigate semantic drift. These comparisons suggest that scaling parameters alone cannot achieve such impressive improvements. Additionally, what we propose is a VDM-based **paradigm**, which can benefit from more efficient future VDM backbones.

More discussions about VDM, visualizations, and ablations are provided in the supplementary.

5 Conclusion

We recast ultra-low-bitrate image compression as a virtual temporal evolution from a compact anchor frame to the final, fine-grained reconstruction. We introduce a generative next-frame decoder that leverages pretrained video diffusion priors to synthesize faithful, realistic images. To enable efficient generative decoding, we propose a two-stage training scheme: first, we adapt the VDM's multi-frame generation to a next-frame prediction task tailored for image compression; second, we collapse the multi-step diffusion process into single-step generation. Our ULB-IC model NeFIC achieves superior perceptual compression performance with efficient decoding.

Acknowledgment

This work was supported by the National Key Research and Development Program of China under Grant 2024YFF0509700, the National Natural Science Foundation of China (62471290,62431015,62331014) and the Fundamental Research Funds for the Central Universities.

References

1. Agustsson, E., Minnen, D., Toderici, G., Mentzer, F.: Multi-realism image compression with a conditional generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22324–22333 (2023)
2. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
3. Agustsson, E., et al.: Generative adversarial networks for extreme learned image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
4. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=rJxdQ3jeg
5. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rkcQFMZRb
6. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
7. Bjøntegaard, G.: Calculation of average psnr differences between rd-curves. ITU SG16 Doc. VCEG-M33 (2001)
8. Black Forest Labs: Flux. https://github.com/black-forest-labs/flux (2024), accessed 29 June 2026
9. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
10. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/ (Feb 2024), accessed 29 June 2026
11. Careil, M., Muckley, M.J., Verbeek, J., Lathuilière, S.: Towards image compression with perfect realism at ultra-low bitrates. In: The Twelfth International Conference on Learning Representations (2024)
12. Chang, J.: Generative image coding with diffusion prior. arXiv preprint arXiv:2509.13768 (2025)
13. Chen, J., Pan, J., Dong, J.: Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In: CVPR (2025)
14. Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: Unireal: Universal image generation and editing via learning real-world dynamics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12501–12511 (2025)

15. Chen, Y., Li, Q., He, B., Feng, D., Wu, R., Wang, Q., Song, L., Lu, G., Zhang, W.: S2cformer: Reorienting learned image compression from spatial interaction to channel aggregation (2025), https://arxiv.org/abs/2502.00700
16. Chen, Y., Lyu, Z., He, B., Cao, N., Chen, G., Lu, G., Zhang, W.: Knowledge distillation for learned image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4996–5006 (2025)
17. Chen, Y., Lyu, Z., He, B., Hu, H., Wang, Q., Tian, Y., Song, L., Zhang, W., Lu, G.: Content-aware mamba for learned image compression. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=WwDNiisZQm
18. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7939–7948 (2020)
19. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
20. Du, J., Zhou, C., Cao, N., Chen, G., Chen, Y., Cheng, Z., Song, L., Lu, G., Zhang, W.: Large language model for lossless image compression with visual prompts. arXiv preprint arXiv:2502.16163 (2025)
21. Eastman Kodak: Kodak lossless true color image suite (photocd pcd0992). http://r0k.us/graphics/kodak/ (1993), accessed 29 June 2026
22. El-Nouby, A., Muckley, M.J., Ullrich, K., Laptev, I., Verbeek, J., Jégou, H.: Image compression with product quantized masked image modeling. arXiv preprint arXiv:2212.07372 (2022)
23. Ghouse, N.F., Petersen, J., Wiggers, A., Xu, T., Sautiere, G.: A residual diffusion model for high perceptual quality codec augmentation. arXiv preprint arXiv:2301.05489 (2023)
24. Ghouse, N.F.K.M., Petersen, J., Wiggers, A.J., Xu, T., Sautiere, G.: Neural image compression with a diffusion-based decoder (2023)
25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
26. Guo, J., Ji, Y., Chen, Z., Liu, K., Liu, M., Rao, W., Li, W., Guo, Y., Zhang, Y.: Oscar: One-step diffusion codec across multiple bit-rates. arXiv preprint arXiv:2505.16091 (2025)
27. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
28. He, D., Yang, Z., Peng, W., Ma, R., Qin, H., Wang, Y.: Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5718–5727 (2022)
29. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
30. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
31. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)

32. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
33. Hoogeboom, E., Agustsson, E., Mentzer, F., Versari, L., Toderici, G., Theis, L.: High-fidelity image compression with score-based generative models. arXiv preprint arXiv:2305.18231 (2023)
34. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
35. Iwai, S., Miyazaki, T., Sugaya, Y., Omachi, S.: Fidelity-controllable extreme image compression with generative adversarial networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 8235–8242. IEEE (2021)
36. Jia, Z., Li, J., Li, B., Li, H., Lu, Y.: Generative latent coding for ultra-low bitrate image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26088–26098 (2024)
37. Jiang, W., Wang, R.: MLIC++: Linear complexity multi-reference entropy modeling for learned image compression. In: ICML 2023 Workshop Neural Compression: From Information Theory to Applications (2023), https://openreview.net/forum?id=hxIpcSoz2t
38. Jiang, W., Yang, J., Zhai, Y., Ning, P., Gao, F., Wang, R.: Mlic: Multi-reference entropy model for learned image compression. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7618–7627 (2023)
39. Jin, L., Watanabe, H.: Perceptual image compression via stable diffusion at low bitrate. In: 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 626–629. IEEE (2024)
40. Ke, A., Zhang, X., Chen, T., Lu, M., Zhou, C., Gu, J., Ma, Z.: Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. arXiv preprint arXiv:2505.08281 (2025)
41. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
42. Körber, N., Kromer, E., Siebert, A., Hauke, S., Mueller-Gritschneder, D., Schuller, B.: Egic: enhanced low-bit-rate generative image compression guided by semantic segmentation. In: European Conference on Computer Vision. pp. 202–220. Springer (2024)
43. Körber, N., Kromer, E., Siebert, A., Hauke, S., Mueller-Gritschneder, D., Schuller, B.: Perco (sd): Open perceptual compression. arXiv preprint arXiv:2409.20255 (2024)
44. Kumari, N., Zhang, R., Shechtman, E., Zhu, J.Y.: Ensembling off-the-shelf models for gan training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10651–10662 (2022)
45. Lee, H., Kim, M., Kim, J.H., Kim, S., Oh, D., Lee, J.: Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity. arXiv preprint arXiv:2403.02944 (2024)
46. Lei, E., Uslu, Y.B., Hassani, H., Bidokhti, S.S.: Text+ sketch: Image compression at ultra low rates. arXiv preprint arXiv:2307.01944 (2023)
47. Li, H., Li, S., Dai, W., Li, C., Zou, J., Xiong, H.: Frequency-aware transformer for learned image compression. In: The Twelfth International Conference on Learning Representations (2024)

48. Li, J., Li, B., Lu, Y.: Hybrid spatial-temporal entropy modelling for neural video compression. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1503–1511 (2022)
49. Li, J., Li, B., Lu, Y.: Neural video compression with diverse contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22616–22626 (2023)
50. Li, J., Zhou, C., Chen, Y., Lu, G.: Differentiable vmaf: A trainable metric for optimizing video compression codec. In: 2025 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–5. IEEE (2025)
51. Li, M., Lin, J., Meng, C., Ermon, S., Han, S., Zhu, J.Y.: Efficient spatially sparse inference for conditional gans and diffusion models. Advances in neural information processing systems **35**, 28858–28873 (2022)
52. Li, Y., Zhang, H., Li, L., Liu, D.: Learned image compression with hierarchical progressive context modeling. arXiv preprint arXiv:2507.19125 (2025)
53. Li, Y., Zhang, H., Li, L., Liu, D., Wu, F.: Scaling learned image compression models up to 1 billion. arXiv preprint arXiv:2508.09075 (2025)
54. Li, Z., Zhou, Y., Wei, H., Ge, C., Jiang, J.: Towards extreme image compression with latent feature guidance and diffusion prior. IEEE Transactions on Circuits and Systems for Video Technology (2024)
55. Li, Z., Zhou, Y., Wei, H., Ge, C., Mian, A.S.: Rdeic: Accelerating diffusion-based extreme image compression with relay residual diffusion. arXiv preprint arXiv:2410.02640 (2024)
56. Lin, Y., Huang, M., Zhuang, S., Mao, Z.: Realgeneral: Unifying visual generation via temporal in-context learning with video models. International conference on Computer Vision (ICCV) (2025)
57. Ling, X., Zhou, C., Li, C., Chen, Y., Tian, Y., Lu, G., Zhang, W.: Free-gvc: Towards training-free extreme generative video compression with temporal coherence. arXiv preprint arXiv:2602.09868 (2026)
58. Liu, J., Lu, G., Hu, Z., Xu, D.: A unified end-to-end framework for efficient deep image compression. arXiv preprint arXiv:2002.03370 (2020)
59. Liu, J., Sun, H., Katto, J.: Learned image compression with mixed transformer-cnn architectures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14388–14397 (2023)
60. Lu, J., Zhang, L., Zhou, X., Li, M., Li, W., Gu, S.: Learned image compression with dictionary-based entropy model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12850–12859 (2025)
61. Lu, L., Xie, Y., Jiang, W., Wang, W., Lin, X., Wang, Y.: Hybridflow: Infusing continuity into masked codebook for extreme low-bitrate image compression. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 3010–3018 (2024)
62. Ma, Y., Yang, W., Liu, J.: Correcting diffusion-based perceptual image compression with privileged end-to-end decoder. International conference on machine learning (ICML) (2024)
63. Mao, Q., Yang, T., Zhang, Y., Wang, Z., Wang, M., Wang, S., Jin, L., Ma, S.: Extreme image compression using fine-tuned vqgans. In: 2024 Data Compression Conference (DCC). pp. 203–212. IEEE (2024)
64. Mentzer, F., Toderici, G.D., Tschannen, M., Agustsson, E.: High-fidelity generative image compression. Advances in neural information processing systems **33**, 11913–11924 (2020)

65. Minnen, D., Ballé, J., Toderici, G.D.: Joint autoregressive and hierarchical priors for learned image compression. Advances in neural information processing systems **31** (2018)
66. Minnen, D., Singh, S.: Channel-wise autoregressive entropy models for learned image compression. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 3339–3343. IEEE (2020)
67. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
68. Muckley, M.J., El-Nouby, A., Ullrich, K., Jégou, H., Verbeek, J.: Improving statistical fidelity for neural image compression with implicit local likelihood models. In: International Conference on Machine Learning. pp. 25426–25443. PMLR (2023)
69. OpenAI: Sora 2 is here. `https://openai.com/index/sora-2/` (Sep 2025), accessed 29 June 2026
70. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
71. Qin, S., Wang, J., Zhou, Y., Chen, B., Luo, T., An, B., Dai, T., Xia, S., Wang, Y.: Mambavc: Learned visual compression with selective state spaces. arXiv preprint arXiv:2405.15413 (2024)
72. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
73. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
74. Relic, L., Azevedo, R., Gross, M., Schroers, C.: Lossy image compression with foundation diffusion models. In: European Conference on Computer Vision. pp. 303–319. Springer (2024)
75. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
76. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
77. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
78. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
79. Song, J., Yang, L., Feng, M.: Extremely low-bitrate image compression semantically disentangled by lmms from a human perception perspective. arXiv preprint arXiv:2503.00399 (2025)
80. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
81. Theis, L., Salimans, T., Hoffman, M.D., Mentzer, F.: Lossy compression with gaussian diffusion. arXiv preprint arXiv:2206.08889 (2022)
82. Toderici, G., Shi, W., Timofte, R., Theis, L., Ballé, J., Agustsson, E., Johnston, N., Mentzer, F.: Workshop and challenge on learned image compression (clic2020). In: Proc. CVPR Workshops (2020)

83. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
84. Vonderfecht, J., Liu, F.: Lossy compression with pretrained diffusion models. International Conference on Learning Representations (2025)
85. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
86. Wang, G.H., Li, J., Li, B., Lu, Y.: Evc: Towards real-time neural image compression with mask decay. arXiv preprint arXiv:2302.05071 (2023)
87. Wang, H., Ma, C.Y., Liu, Y.C., Hou, J., Xu, T., Wang, J., Juefei-Xu, F., Luo, Y., Zhang, P., Hou, T., et al.: Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2578–2588 (2025)
88. Wu, S., Chen, Y., Liu, D., He, Z.: Conditional latent coding with learnable synthesized reference for deep image compression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 12863–12871 (2025)
89. Xia, Y., Zhou, Y., Wang, J., An, B., Wang, H., Wang, Y., Chen, B.: Diffpc: Diffusion-based high perceptual fidelity image compression with semantic refinement. In: The Thirteenth International Conference on Learning Representations (2025)
90. Xu, T., Li, J., Li, B., Wang, Y., Zhang, Y.Q., Lu, Y.: Picd: Versatile perceptual image compression with diffusion rendering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28436–28445 (2025)
91. Xu, T., Zhu, Z., He, D., Li, Y., Guo, L., Wang, Y., Wang, Z., Qin, H., Wang, Y., Liu, J., Zhang, Y.Q.: Idempotence and perceptual image compression. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=Cy5v64DqEF
92. Xue, N., Jia, Z., Li, J., Li, B., Zhang, Y., Lu, Y.: Dlf: Extreme image compression with dual-generative latent fusion. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (Oct 2025)
93. Xue, N., Jia, Z., Li, J., Li, B., Zhang, Y., Lu, Y.: One-step diffusion-based image compression with semantic distillation. Advances in Neural Information Processing Systems (2025)
94. Xue, N., Mao, Q., Wang, Z., Zhang, Y., Ma, S.: Unifying generation and compression: Ultra-low bitrate image coding via multi-stage transformer. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
95. Yang, R., Mandt, S.: Lossy image compression with conditional diffusion models. Advances in Neural Information Processing Systems **36**, 64971–64995 (2023)
96. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
97. Zavadski, D., Feiden, J.F., Rother, C.: Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems. In: European Conference on Computer Vision. pp. 343–362. Springer (2024)
98. Zeng, F., Tang, H., Shao, Y., Chen, S., Shao, L., Wang, Y.: Mambaic: State space models for high-performance learned image compression. arXiv preprint arXiv:2503.12461 (2025)
99. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)

100. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
101. Zhang, T., Luo, X., Li, L., Liu, D.: Stablecodec: Taming one-step diffusion for extreme image compression. arXiv preprint arXiv:2506.21977 (2025)
102. Zhang, Y., Lu, G., Chen, Y., Wang, S., Shi, Y., Wang, J., Song, L.: Neural rate control for learned video compression. In: The Twelfth International Conference on Learning Representations (2024)
103. Zhang, Z., Xia, Y., Chen, B., Zhang, T., Wang, H., Qiu, H.: Autoregressive-based progressive coding for ultra-low bitrate image compression. In: The Fourteenth International Conference on Learning Representations (2026)
104. Zhou, C., Ling, X., Chen, Y., Dai, J., Lu, G., Zhang, W.: Dual-representation image compression at ultra-low bitrates via explicit semantics and implicit textures. arXiv preprint arXiv:2602.05213 (2026)
105. Zhou, M., Kong, M.: Glic: General format learned image compression. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10815–10824 (2025)
106. Zhu, Y., Yang, Y., Cohen, T.: Transformer-based transform coding. In: International Conference on Learning Representations (2022)
107. Zou, R., Song, C., Zhang, Z.: The devil is in the details: Window-based attention for image compression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17492–17501 (2022)