APSIPA Transactions on Signal and Information Processing, 2025, 14, e33

Overview Paper

Generative Coding: Promise and Challenges

Siwei Ma[1*], Shenpeng Song[1], Bolin Chen[2], Qi Mao[3], Xiaohan Fang[2], Chuanmin Jia[4] and Shiqi Wang[2]

[1] *State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University, Beijing, China*
[2] *Department of Computer Science, City University of Hong Kong, Hong Kong, China*
[3] *Communication University of China, Beijing, China*
[4] *Wangxuan Institute of Computer Technology, Peking University, Beijing, China*

ABSTRACT

Traditional image and video compression techniques, based on handcrafted transforms and distortion metrics, have proven effective in earlier applications. However, their inherent limitations in coding efficiency and perceptual quality become increasingly evident when faced with the demands of diverse and semantically complex visual content. With advances in deep generative models, generative coding has emerged as a promising alternative, offering improved efficiency, perceptual quality, and flexibility. However, it also poses challenges in complexity, interpretability, and deployment. This survey provides a comprehensive overview of generative coding. We formalize the problem and highlight its theoretical links to generation and compression. Representative methods are categorized by model type and technical evolution. Finally, we further present comparative experiments and discuss key challenges and future directions to guide ongoing research.

*Corresponding author: Siwei Ma, swma@pku.edu.cn

Received 28 June 2025; revised 31 August 2025; accepted 14 September 2025
ISSN 2048-7703; DOI 10.1561/116.20250056

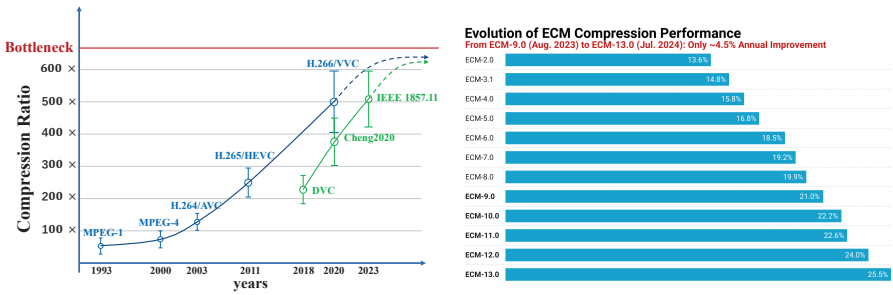
Keywords: Image and video compression, generative compression

1 Introduction

With the widespread adoption of generative artificial intelligence (AI) technologies and high-performance computing devices, digital content creation has undergone a profound transformation, ranging from Professional Generated Content (PGC) [77], to User Generated Content (UGC) [84], and more recently to Artificial Intelligence Generated Content (AIGC) [168, 106]. The rapid advancement of AI has fundamentally shifted the paradigm of content creation from manually designed processes to algorithm-driven workflows. This shift has substantially fueled the exponential growth of visual data, thereby introducing new challenges for compression technologies in terms of storage, transmission, and computational efficiency.

Over the past two decades, traditional image and video compression standards such as H.264/Advanced Video Coding (AVC) [167], H.265/High Efficiency Video Coding (HEVC) [149], and H.266/Versatile Video Coding (VVC) [14] have been designed to preserve essential visual information while minimizing bitrate, thereby achieving high-efficiency data compression. However, these manually crafted compression frameworks primarily rely on the statistical properties of visual data to improve compression efficiency, following Shannon's rate-distortion theory [172], and are gradually approaching a performance ceiling. On one hand, the marginal gains from traditional coding techniques have significantly diminished. The report [23] indicates that only one-quarter of the new tools introduced in the latest coding standard, VVC, yield performance improvements exceeding 1%. Moreover, hardware encoders for the current AVS3/VVC generation require approximately three times the hardware area compared to AVS2/HEVC to achieve a 25% increase in compression efficiency. In addition, manually designed codecs are typically tailored to specific data modalities and are not optimized for general or heterogeneous content, posing significant limitations in addressing the growing demand for multimodal compression.

In response to the limitations of hand-crafted coding paradigms, learning-based compression methods have been introduced as a data-driven extension, leveraging neural networks to enhance or replace traditional modules [98, 192, 190, 163, 115, 71, 72, 103, 120], and to construct end-to-end trainable image and video coding frameworks [7, 9, 123, 187, 113, 150, 112, 64, 175, 189, 105, 102, 96, 161, 97]. Leveraging large datasets and the robust nonlinear transformation representation capabilities of neural networks, the performance of these new compression algorithms has been significantly enhanced. However, despite their initial success, current end-to-end neural compression frameworks are also exhibiting signs of saturation. As shown in Figure 1, both



(a) Compression ratio trends: traditional standards vs. learned end-to-end codecs.

(b) All-intra performance evolution of the Enhanced Compression Model (ECM).

Figure 1: Both traditional and learning-based methods are approaching their limits.

traditional and learning-based methods are approaching performance bottlenecks under existing architectural and optimization constraints, highlighting the need for the exploration of new paradigms. Given the trend of data volume doubling every two years, one must question: where can we find the potential for achieving compression ratios thousands of times higher?

In light of these challenges, deep generative models—such as Variational Auto-encoders (VAEs) [79], Generative Adversarial Networks (GANs) [42], and Denoising Diffusion Probabilistic Models (DDPMs) [58]—have emerged as a promising direction for advancing image and video compression. These models have demonstrated impressive performance across a range of tasks, including text-to-image generation [131, 92], image super-resolution [94, 139], style transfer [76, 186], image animation [162, 142], and video generation [114, 57]. Their strong generative capability stems from the ability to model complex data distributions and produce high-fidelity samples that closely resemble real data. This generative capacity underpins many of the principles shared with compression. Fundamentally, both generation and compression require accurate modeling of the underlying data distribution: generative models aim to synthesize realistic samples from learned distributions [144, 147, 104], while compression algorithms exploit these distributions to encode data more efficiently [140, 7]. From this shared foundation, a natural convergence has emerged, wherein generative models are increasingly integrated into learned compression frameworks. By capturing semantic priors of texture, structure, and motion, generative models enable more compact representations. For instance, in extremely low-bitrate scenarios, high-quality and semantically coherent images can be reconstructed at the receiver side from only a few bytes of transmitted latent variables or text prompts. In this context, generative compression leverages deep generative networks to learn informative priors of visual content, encoding both low-level appearance and high-level semantics.

In summary, with the rapid advancement of GPU computing and artificial intelligence, particularly in generative technologies, the focus of improving coding efficiency has shifted away from traditional hand-crafted codecs. To fulfill the potential of image and video coding as foundational infrastructure for future digital media, it is essential to achieve breakthroughs offering 5×, 10×, or even 100× improvements over current standards. In this context, generative coding is poised to redefine the coding paradigm, delivering not only substantial gains in efficiency but also enhanced perceptual quality [154, 170]. Moreover, with the development of multimodal large models [1, 32] and AIGC-based applications [15, 36], generative coding exhibits inherent compatibility with multimodal representations. By modeling shared latent spaces across visual, textual, and semantic domains [75, 183], generative approaches facilitate unified and scalable compression strategies, enabling more immersive and intelligent media experiences. This synergy paves the way for a new era of intelligent and creative digital media consumption, where high-quality, cross-modal information can be efficiently encoded, transmitted, and reconstructed.

This paper presents a comprehensive review of generative coding, focusing on its potential, capabilities, and challenges. We hope this review elucidates three key aspects of generative coding: its potential to replace traditional methods, its advanced capabilities and efficiency, and the current bottlenecks and challenges. The main contributions are listed as follows.

- We present a principled formulation of generative coding, including preliminary definitions and a general paradigm, and provide a comprehensive survey of representative methods throughout their technical evolution.
- We report experimental results on our proposed approach along with representative baselines to demonstrate its practical effectiveness in generative compression.
- We identify key challenges and outline future research directions to advance the frontier of generative coding.

2 Generative Coding: Formulation and Paradigm

In this section, we present the formulation, which encompasses detailed explanations of both generation and compression, while thoroughly examining their interrelations. Grounded in a robust mathematical foundation, we delineate the definitions and paradigms of generative coding.

2.1 Mathematical Formulation

Generative models aim to learn a mapping from a simple prior distribution (e.g., Gaussian) to a complex target distribution. Formally, given a model $G : Y \to X$ mapping a conditional domain Y to an image domain X, the generation process is defined as:

$$x' = G(y), \quad x' \sim p(x' \mid y), \tag{1}$$

where $p(x' \mid y)$ is the learned conditional distribution. Supposing $p(x \mid y)$ is the true conditional distribution, the quality of generated samples is reflected in how well $p(x' \mid y)$ approximates $p(x \mid y)$. A closer match indicates higher fidelity and typically better perceptual quality.

Coding algorithms aim to represent visual data with minimal bits while maintaining high perceptual quality. Formally, given a T-frame sequence $m_{1:T} = \{m_1, m_2, \ldots, m_T\}$, where each $m_t \in \mathbb{R}^{H \times W \times C}$ denotes a frame with height H, width W, and C color channels, the goal is to encode and decode it via a compact bitstream ξ, such that the reconstructed sequence $\hat{m}_{1:T}$ approximates the original with minimal degradation under a given bitrate constraint. This task is formalized as a rate-distortion optimization problem:

$$\mathcal{J}^* = \min_{\xi} \left[\mathcal{D}(m_{1:T}, \hat{m}_{1:T}) + \lambda \, \mathcal{R}(\xi) \right], \tag{2}$$

where $\mathcal{D}(\cdot)$ denotes the distortion between the original and reconstructed sequences, $\mathcal{R}(\cdot)$ is the bits consumption, and λ is a Lagrange multiplier balancing rate and distortion.

2.2 Connection Between Generation and Compression

Visual compression and visual generation share fundamental methodological connections, as both aim to model the target data distribution and construct precise mapping relationships between probability spaces. In image generation, models learn to approximate the target data distribution by mapping variables sampled from a simple prior (e.g., Gaussian or uniform noise) to realistic outputs. On the other hand, in image compression, especially deep learning-based approaches, the goal is to minimize the bitrates needed to represent an image while controlling distortion within acceptable limits. Information-theoretic principles reveal that the optimal coding strategy is fundamentally determined by the negative log-likelihood of the data distribution [140, 7], highlighting that precise distribution modeling is essential for achieving rate-distortion optimality.

In generative modeling, let x denote a sample from the data space and z a corresponding latent variable. To approximate the true data distribution $p(x)$, a model distribution $q_\theta(x)$ is defined as:

$$q_\theta(x) = \int q_\theta(x, z)\, dz = \int q_\theta(x \mid z) q(z)\, dz, \tag{3}$$

where θ represents the parameters of the generative model.

However, directly optimizing the divergence between the true and model distributions, $\mathrm{KL}(p(x) \parallel q_\theta(x))$, is typically intractable due to the marginalization over latent variables. To address this, variational inference introduces an auxiliary latent posterior distribution, enabling the transformation of the marginal KL divergence into a tractable upper bound on the joint KL divergence:

$$\begin{aligned} \mathrm{KL}(p(x, z) \parallel q(x, z)) &= \mathrm{KL}(p(x) \parallel q(x)) \\ &\quad + \int p(x)\, \mathrm{KL}(p(z \mid x) \parallel q(z \mid x))\, dx \\ &\geq \mathrm{KL}(p(x) \parallel q(x)). \end{aligned} \tag{4}$$

This joint KL divergence thus serves as a variational upper bound for optimizing the generative model. By further deriving the variational objective and estimating the integral via Monte Carlo sampling with the reparameterization trick, the joint KL divergence can be rewritten as:

$$\mathcal{L}_G = \mathbb{E}_{x \sim p(x)} \left[-\log q(x \mid z) + \mathrm{KL}(p(z \mid x) \parallel q(z)) \right], \tag{5}$$

where $\mathbb{E}_{x \sim p(x)} \left[\mathrm{KL}(p(z \mid x) \parallel q(z)) \right]$, encourages the learned posterior $p(z \mid x)$ to align with the prior $q(z)$, thereby enabling sampling from the prior for unsupervised generation.

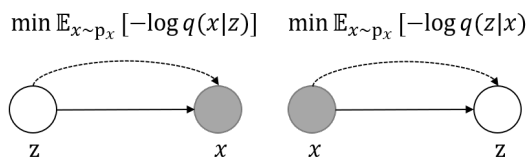
For end-to-end learning-based compression, let x, z, and x' denote the input, the compressed latent representation, and the reconstructed output, respectively. The rate-distortion optimization objective, consistent with Equation 2, can be reformulated as:

$$\mathcal{L}_E = \mathbb{E}_{x \sim p(x)} [-\log q(z \mid x)] + \lambda\, \mathbb{E}[d(x, x')], \tag{6}$$

where the first term, $\mathbb{E}_{x \sim p(x)} [-\log q(z \mid x)]$, represents the expected negative log-likelihood of the latent variable z given the input x, which corresponds to the estimated bitrates.

Interestingly, under extreme assumptions, both objectives reduce to conditional maximum likelihood estimation. Specifically, for Equation 5, omitting the KL term yields an objective that maximizes the conditional likelihood $q(x \mid z)$ with $z \sim p(z \mid x)$, emphasizing reconstruction fidelity and similar to

a deterministic autoencoder, disregarding the generative quality of samples drawn from the prior. Conversely, for Equation 6, when the distortion term is omitted (i.e., $\lambda = 0$), the objective reduces to maximizing the conditional likelihood $q(z \mid x)$, which corresponds to learning an encoder that efficiently models the latent representation, prioritizing bit rate minimization without considering reconstruction accuracy. In this scenario, as illustrated in Figure 2, the two optimization processes differ primarily in the direction of inference between the input and output domains.



(a) Generation objective. (b) Compression objective.

Figure 2: Illustration of the generation and compression processes under extreme assumptions where both objectives reduce to conditional maximum likelihood estimation with reversed inference directions between the input and output domains.

In summary, although visual generation and visual compression serve distinct practical purposes, they share a common theoretical foundation based on effective data distribution modeling. This alignment suggests significant complementarity and potential for integration, providing a strong basis for cross-disciplinary approaches and the development of new paradigms like generative compression.

2.3 Paradigm

Generative compression algorithms aim to leverage deep generative models to learn compact latent representations of data, and generatively reconstruct target data from these representations, thereby achieving high-quality inference for compression tasks and promising rate-distortion performance. Building upon the preceding derivations, we now conclude the framework for generative coding, as illustrated in Figure 3. Within this framework, the generative model, conditioned on the bitstream, functions as the decoder and constitutes the core of the entire system. To formalize this concept, we define *generative coding for images/videos* as follows: ***Generative coding for images/videos*** *refers to deep learning-based approaches that generate reconstructed visual data conditioned on compressed representations encoded from the input, optimizing efficiency, fidelity, and perceptual quality.*

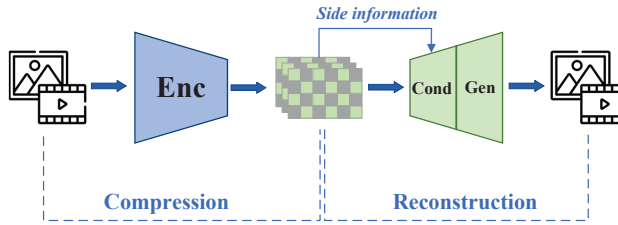


Figure 3: Illustration of generative coding framework.The input visual content (image or video, shown by the icon on the left) is first encoded by the encoder (Enc) into a compact latent representation, which serves as *side information*. During reconstruction, the conditional generator (Gen) acts as the decoder. The side information is utilized as a conditioning signal (Cond) to guide the generator in reconstructing the visual content.

3 Generative Coding: Progress Survey

Herein, we propose potential solutions for designing a generative coding framework that achieves high encoding efficiency and perceptually superior reconstruction, informed by a review from the perspectives of different technical approaches and focal points. Specifically, we categorize existing generative coding studies primarily based on the type of core generative model employed, such as variational autoencoders (VAEs), diffusion models, and generative adversarial networks (GANs). Although some studies incorporate hybrid architectures (e.g., combining a VAE encoder with a diffusion-based generator), we classify them based on the dominant generative mechanism, with such cases grouped under diffusion models. To provide historical context, Figure 4 illustrates the chronological evolution of generative models and their impact on coding research. In the following subsections, we provide a detailed survey of existing works.

3.1 Explicit and Direct Probability Modeling: Flow and Autoregressive Models

Flow-based models, also known as *normalizing flows* [30], constitute a class of generative models built upon a sequence of bijective and differentiable transformations. Their core principle is to map a complex data distribution $p_X(x)$ to a simple and tractable latent distribution $p_Z(z)$ (typically an isotropic Gaussian) through invertible functions f_θ. Thanks to their reversibility, flow models enable efficient and lossless bidirectional mapping: new samples can be generated by sampling from the latent space and applying the inverse transformation f_θ^{-1} to reconstruct data in the original space.

Autoregressive models [159], by contrast, model the data generation process as a sequence of conditional probability estimations. Leveraging the chain

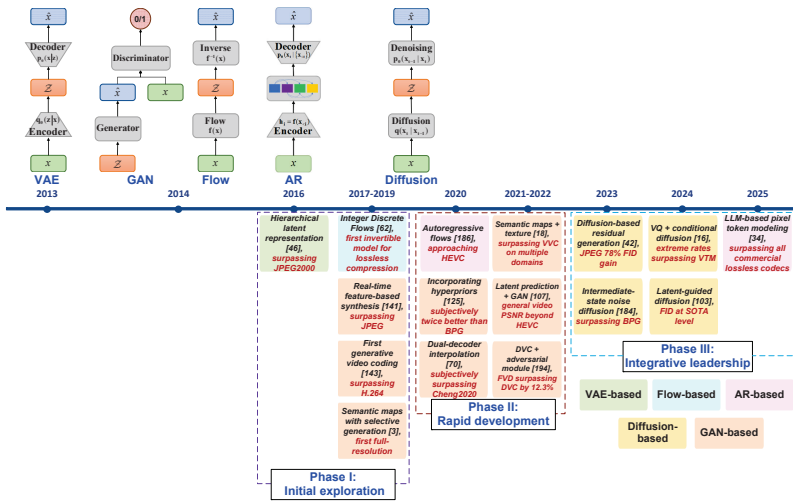


Figure 4: Development timeline of generative models and their applications in coding. The upper part illustrates the evolution of core generative models, while the lower part highlights representative breakthroughs. The progression is divided into three phases, demonstrating how generative paradigms progressively enhance coding performance.

rule of probability, they decompose the joint distribution into an ordered product of conditionals:

$$p(x) = \prod_{i=1}^{n} p(x_i \mid x_1, x_2, \ldots, x_{i-1}) = \prod_{i=1}^{n} p(x_i \mid x_{<i}), \tag{7}$$

where $x = (x_1, x_2, \ldots, x_n)$ denotes the input sequence, and $x_{<i}$ refers to all previously generated elements. Each term $p(x_i \mid x_{<i})$ is a conditionally sampleable distribution, enabling sequential data generation in a flexible and autoregressive manner. This formulation supports precise modeling of complex dependencies in data sequences.

Despite architectural differences, flow-based and autoregressive models share a common optimization strategy: maximizing the data log-likelihood, $\log p_\theta(\mathbf{x})$, for explicit probabilistic modeling. In the following, we review representative works that leverage these modeling approaches in the context of image compression.

3.1.1 Flow-based Modeling

et al. Flow-based models were first applied to lossless image compression by Hoogeboom *et al.* with the Integer Discrete Flows (IDF) method [61]. IDF leverages invertible transformations to enable lossless, bidirectional mapping

between data and latent spaces, achieving state-of-the-art performance at the time. Its hierarchical design, combining squeeze operations [31], integer flow layers, and factor-out modules, supports progressive decoding by modeling the prior in a conditional and multi-scale manner.

In 2021, Helminger *et al.* [55] introduced normalized flows for lossy image compression via a three-level hierarchy. Despite suboptimal rate-distortion performance, their invertible design enabled iterative and multi-pass decoding without error accumulation (i.e., idempotent reconstruction). To further improve expressiveness, ANFIC [59] introduced a hybrid architecture combining multi-layer VAEs with invertible flows. Stacking VAE modules hierarchically achieves quasi-invertibility, retaining flow-based reversibility while enhancing flexibility. ANFIC outperformed prior flow models [116] and several VAE-based methods [9, 86, 63] in BD-rate.

However, the aforementioned flow-based approaches still suffer from relatively high bitrates and fail to fully exploit the potential of generative models in achieving high perceptual quality at extremely low bitrates. To address this limitation, an innovative method was proposed by [37], which employs flow models as pre- and post-processing modules around a conventional compression backbone. This framework leverages invertible transformations to map the original image into an intermediate, compression-friendly representation. This design reduces coding redundancy while preserving perceptual quality. Compared to representative end-to-end methods for extreme low-bitrate compression [74, 188], this approach achieves higher-fidelity reconstruction at bitrates below 0.05 bpp, demonstrating the promising potential of flow-based models in this challenging regime.

In summary, flow-based models provide a principled framework for image compression, featuring exact likelihood estimation, invertibility, and idempotent reconstruction, where the decoding process avoids error accumulation across iterative passes, even though lossy quantization still introduces reconstruction errors. Their unique properties support stable, multi-pass, and lossless reconstruction. Nonetheless, flow-based approaches exhibit limitations: they are less flexible in architectural design, face challenges in modeling high-dimensional, complex data distributions, and incur substantial computational costs due to invertibility constraints and Jacobian determinant evaluations. Consequently, while flow-based compression provides theoretical elegance and certain unique advantages, overcoming these limitations is essential for improving both ratedistortion performance and scalability in practical applications.

3.1.2 Autoregressive Modeling

In recent years, autoregressive modeling has been widely adopted in image compression [123, 193, 107]. As discussed earlier, autoregressive mechanisms

are frequently employed in vector quantization (VQ)-based schemes as well. Broadly speaking, research in this direction can be categorized into two main approaches.

The first involves classical autoregressive image modeling, which typically estimates the conditional probability distribution of pixels using masking strategies during generation, often enhanced by context models to improve entropy coding efficiency. These methods are generally integrated into the entropy model to reduce bitrates [124, 86, 130, 127]. The second approach extends the paradigm of large language models (LLMs) to image compression, giving rise to a novel language-driven perspective on image compression [52, 20, 33, 93, 146]. This emerging direction highlights the potential of integrating generative modeling with multimodal perception and compression, and is expected to play a pivotal role in the future convergence of generative compression and multimodal generation.

In the domain of autoregressive image compression, the overall framework is illustrated in Figure 5, where the decoder is typically trained using generative techniques. El-Nouby *et al.* [127] proposed replacing conventional vector quantization with product quantization (PQ), an intermediate form between vector and scalar quantization. This design expands the trade-off space between rate and reconstruction quality. Combined with a masked image modeling-based conditional entropy model, their method enables robust reconstruction even when only partial tokens are available. For video compression, Yang *et al.* [178] unified generative modeling, variational inference, and autoregressive flows. By introducing structured priors between motion and residual latent variables, they improved entropy modeling and achieved rate-distortion performance comparable to HEVC.

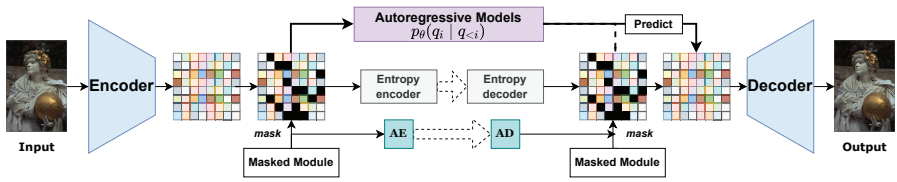


Figure 5: Overview of an autoregressive generative image compression framework. AE and AD denote arithmetic encoder and decoder, respectively. The autoregressive models estimate $p_\theta(q_i \mid q_{<i})$ for both entropy coding and sequential token prediction during decoding. The masked module facilitates context modeling by limiting access to future tokens during training and inference.

Recent studies have extended autoregressive modeling to large language models (LLMs) for image and video compression. Han *et al.* [52] directly modeled bitstreams from traditional codecs (e.g., JPEG, AVC/H.264) using 7B-parameter LLaMA-2 [157], introducing JPEG-LM and AVC-LM for generative decoding. Building on this, Chen *et al.* [20] proposed a lossless approach

by encoding per-pixel RGB values with prompt templates and fine-tuning LLaMA-3 [43], achieving competitive bit-per-subpixel (bpsp) on CLIC.mobile [155] (2.08) and Kodak [25] (2.83), albeit with an average decoding time of 273.11 seconds on Kodak. Similarly, Du *et al.* [33] used LLMs as entropy models to predict residual distributions conditioned on lossy reconstructions, attaining comparable bpsp but with even longer decoding times (495.6 seconds). These results highlight the promise of LLM-based compression while emphasizing the need for more efficient decoding.

Another line of work incorporates LLMs at the encoding stage for semantic-aware compression without requiring additional training of the large models themselves [93, 146]. These methods leverage vision-language models such as GPT-4 Vision [1], Grounding DINO [109], and SAM [80] to extract object-level semantic hierarchies from input images, including global captions, individual object descriptions, and corresponding binary masks. A low-bitrate end-to-end codec is then used to compress a reference image, forming a multimodal bitstream that includes textual descriptions, masks, and image tokens. At the decoder side, ControlNet [182] guides the reconstruction by sequentially conditioning on the semantic information, ultimately achieving high-fidelity generation.

Despite recent progress, LLM-driven image compression remains far from practical deployment. The primary bottleneck lies in its excessive decoding complexity and inference latency. Striking a balance between generative flexibility and computational efficiency will be crucial for advancing this direction and realizing its potential in real-world applications.

In summary, autoregressive methods distinguish themselves by factorizing data distributions into conditionals, enabling precise dependency modeling and high-fidelity generation. This makes them particularly suited for integrating semantic information and extending toward multimodal compression with visionlanguage models. However, these advantages come at the cost of sequential decoding, which severely limits efficiency, and of high computational and memory demands, particularly in LLM-based frameworks. These characteristics suggest that while autoregressive compression is theoretically powerful and semantically rich, its future progress will critically depend on innovations such as token pruning and compression [12, 91, 166] to reduce decoding complexity and improve scalability.

3.2 Approximate Inference: VAEs and Diffusion Models

Variational Autoencoders (VAEs) [79] constitute one of the earliest generative frameworks grounded in variational inference and probabilistic graphical modeling. They adopt an encoderdecoder architecture, where the encoder maps input data to a probability distribution in the latent space, and the decoder reconstructs the original data from sampled latent variables.

Denoising diffusion models [147, 58] represent another class of generative approaches that achieve high-quality sample generation by inverting a noise corruption process. The core idea involves constructing a forward diffusion process (typically modeled as a Markov chain) that gradually transforms data from the target distribution into standard Gaussian noise [144]. A neural network is then trained to approximate the reverse denoising process, enabling the reconstruction of data from pure noise.

Unlike flow-based and autoregressive models that support exact likelihood estimation, variational autoencoders (VAEs) and diffusion models rely on approximate inference to model data distributions. In the following, we review representative works based on these two paradigms.

3.2.1 VAEs Modeling

As one of the earliest generative models, the VAEs has been widely adopted in image and video coding tasks, serving as a fundamental building block in the construction of learned codec frameworks.

Gregor *et al.* [45] first applied generative modeling to lossy image compression by introducing a hierarchical latent variable structure based on VAEs. This design enables multi-scale representation, capturing both high-level semantics and low-level details. By hierarchically organizing latent variables, the model retains only top-layer latent during inference and reconstructs the rest during decoding, effectively balancing compression rate and reconstruction quality. This conceptual compression framework has inspired a range of learned image codecs [160, 153, 9].

Building on this foundation, Jia *et al.* [73] extended the generative latent space to discrete representations via vector quantization [44], advancing the VQ-VAE framework [160]. Their method employs a three-stage training scheme: (i) learning a latent space via encoder-decoder training, (ii) applying transform coding for entropy compression of latent, and (iii) joint fine-tuning. This approach significantly reduces hyperprior redundancy and outperforms pixel-space methods at extremely low bitrates, while also exhibiting richer semantic expressiveness and better alignment with human perception.

Similar to image coding, variational autoencoders (VAEs) form a foundational framework for generative video compression. Early work by Han *et al.* [111] integrated deep generative models with quantization and entropy coding, employing a sequential VAE to encode video frames individually. Building on this, Habibian *et al.* [49] introduced a 3D autoencoder with a deterministic encoder and an autoregressive prior to jointly model short video clips (e.g., 8 frames). These efforts mark a transition from frame-wise modeling to temporally coherent sequence-level modeling in generative video compression.

Overall, VAEs are inherently limited in generative fidelity and struggle to scale toward high-resolution synthesis. Even so, VAEs made two lasting contributions to visual compression: they established the paradigm of probabilistic latent-space modeling with a systematic variational optimization scheme, and they enabled hierarchical or quantized latent representations that integrate naturally with entropy coding. These advances not only shaped early codec design but also laid the foundation for more expressive generative approaches.

3.2.2 Diffusion Modeling

Since 2021, diffusion-based generative image coding has rapidly emerged as a prominent research direction in the field of image compression. As summarized in Table 1, existing approaches can be broadly categorized into two groups: Pure diffusion-based coding frameworks and hybrid diffusion frameworks. The key distinction lies in the source of the diffusion models input and its functional role within the overall compression pipeline.

Table 1: Summary of diffusion modeling approaches for generative coding.

		Publication	Publication Venue	Remarks
Diffusion	Pure Diffusion Codec (LDM with Feature or Text Guidance)	Lucas Theis 2022 [152]	Arxiv	Progressive Encoding Scheme
		Tom Bordin 2023 [13]	IEEE MMSP	Downscaled Segmentation Maps for Guidance
		Ruihan Yang 2023 [177]	NIPS	Image latent-guided diffusion
		Eric Lei 2023 [89]	ICML workshop	Text + Sketch Guidance
		Marlene Careil 2024 [16]	ICLR	Dual Latent Space with VQ-text Conditioning
		Tom Bachard 2024 [6]	PCS	Hybrid CLIP Latent & Color Map Encoding
		Lucas Relic 2024 [134]	ECCV	VAE-diffusion Hybrid Framework
		Meiqin Liu 2024 [108]	Arxiv	Unified Diffusion-based Video Codec
		Wenzhuo Ma 2025 [117]	Arxiv	Diffusion with Temporal Reuse
	Hybrid Codec (Base Codec with Diffusion Post-processing)	Junlong Gao 2023 [39]	DCC	Cheng2020 + Text Enhancement
		Noor Fathima Ghouse 2023 [41]	Arxiv	Diffusion-based Residual Generation
		Junlong Gao 2024 [38]	TCSVT	Cross-modal Encoding
		Haowei Kuang 2024 [85]	ACM MM	Postprocessing with Syntax-guided Diffusion
		Emiel Hoogeboom 2024 [60]	Arxiv	Diffusion-autoencoder with rectified flows
		Yiyang Ma 2024 [118]	ICML Workshop	Diffusion-decoder with privileged correction
		Zhiyuan Li 2024 [101]	TCSVT	Postprocessing VAE Latent Representations

In pure diffusion-based frameworks, the diffusion model operates directly on a latent representation extracted from the input image via an encoder. It serves as the central component responsible for both compression and reconstruction. Representative examples include Latent Diffusion Models (LDMs) guided by semantic priors. By contrast, hybrid diffusion frameworks adopt a modular design that combines a traditional or learned-based codec with a diffusion model. In this setting, the base codec performs the primary encoding and decoding, while the diffusion model is applied to intermediate reconstructions, either in the image or latent domain, produced by the base codec. The following sections provide a detailed review of representative works in each category.

Pure Diffusion Codec

Within the framework of fully diffusion-based image compression, the modeling paradigm has evolved from unconditional generation to conditional guidance. Theis *et al.* [152] introduced a pioneering unconditional diffusion-based compression framework, marking a departure from traditional compression pipelines that rely on transform coding and quantization. In contrast to mainstream end-to-end learned compression systems, which are typically composed of an encoder transform, entropy model, and decoder reconstruction module [8, 179], this method applies Gaussian perturbations directly in the pixel space and reconstructs the image via a reverse diffusion process. The formulation enables inherently progressive decoding and represents a paradigm shift beyond standard compression architectures.

Subsequently, researchers introduced conditional diffusion modeling to enhance the modeling capacity. The Conditional Diffusion Compression (CDC) model proposed by Yang *et al.* [177] draws on the latent variable modeling paradigm of VAEs, using the latent variable z extracted by the encoder as a conditioning input to guide the reverse diffusion process in reconstructing image content. Meanwhile, the remaining high-frequency texture details are modeled through the generative capacity of the diffusion process. As the first diffusion-based compression framework reported to surpass GAN-based methods, CDC has established a representative baseline for this direction and demonstrated strong performance across various datasets and evaluation protocols.

Extending this line of work, subsequent research broadened the scope of conditional information from single latent variables to multi-level structural priors, as is shown in Figure 6. By incorporating structured guidance signals, such as semantic segmentation maps and color maps, these approaches introduced stronger constraints into the diffusion process to facilitate more faithful reconstruction. Bordin *et al.* [13] conditioned diffusion models on semantic segmentation maps, further guided by color maps to enhance semantic fidelity and color consistency. Their method maintains high reconstruction quality even with heavily downsampled segmentation inputs, demonstrating strong generalization beyond input priors. Building on the idea of leveraging alternative structural cues, Lei *et al.* [89] achieved extreme compression (< 0.01 bpp) through text-sketch decomposition and ControlNet [182]-based reconstruction, though at the cost of severe distortions in object details. Pushing further, Bachard *et al.* proposed CoCliCo [6], which incorporates global semantic embeddings from CLIP [132] alongside color maps to guide latent-space diffusion. CoCliCo achieves visually rich outputs at extremely low bitrates (e.g., <0.002 bpp), showcasing the potential of semantic guidance for high-fidelity compression under extreme constraints. In parallel, Careil *et al.* [16] combined vector-quantized latent representations with text-guided condi-

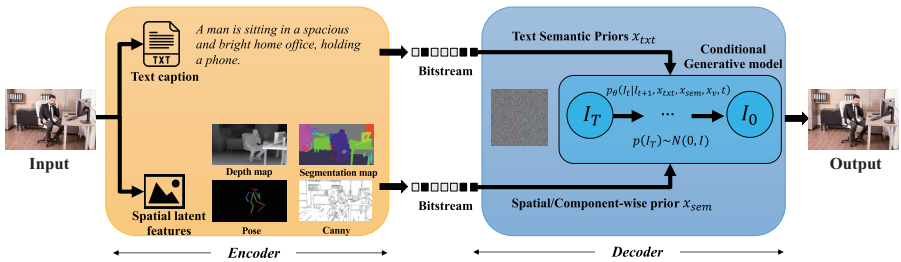


Figure 6: Illustration of generative coding with multi-level priors. The input image is encoded into multiple priors including structural priors (e.g., depth, segmentation, pose, edges) and semantic priors (e.g., text captions). These priors are compressed into bitstreams and transmitted to the decoder. A conditional generative model, exemplified by diffusion, reconstructs the original image by progressively denoising from noise, guided by semantic and spatial priors.

tional decoding, achieving photorealistic reconstructions at ultra-low bitrates ($0.1 \sim 0.003$ bpp). Their approach surpasses traditional codecs by an order of magnitude, highlighting the strength of pre-trained diffusion models as image priors.

The aforementioned works have not explicitly addressed the complexity of diffusion models, which often suffer from high decoding latency. To tackle the computational overhead during diffusion inference, Relic *et al.* [134] model the quantization error as a denoising task and introduce a parameter estimation module to jointly learn adaptive quantization parameters and the optimal number of denoising steps. Unlike conventional diffusion models that initiate sampling from pure Gaussian noise, their method starts directly from quantized latents, enabling effective elimination of redundant diffusion steps through joint prediction. Under low-bitrate conditions, the approach achieves high perceptual quality using less than 10% of the sampling steps, significantly improving decoding efficiency.

Extending to temporal dynamics, Liu *et al.* [108] proposed a fully generative video compression framework that unifies intra- and inter-frame modeling within a single conditional diffusion process. Unlike traditional hybrid codecs relying on explicit motion estimation, their method employs diffusion inversion to implicitly align inter-frame information. Reference frame features serve as priors to guide a selective DDIM [145] denoising on motion-sensitive regions, effectively capturing temporal dependencies and enhancing frame consistency. Building on this, Ma *et al.* [117] extended the DCVC-DC architecture [96] by introducing a conditional diffusion model that incorporates both temporal context and latent reconstructions of the current frame. This design improves visual quality while preserving temporal coherence. To address the computational cost of diffusion inference, they further proposed a temporal informa-

tion reuse strategy that recycles diffusion trajectories from previous P-frames, achieving significant acceleration with negligible quality degradation.

Hybrid Diffusion-based Compression

These methods typically enhance the initial reconstruction produced by a base codec using diffusion models, forming a hybrid compression framework.

A representative line of work is proposed by Gao *et al.* [39, 38], which builds upon the Cheng2020 [22] model and applies a text-guided diffusion model as a post-processing enhancement. This framework integrates the enhancement process into a unified rate-distortion optimization objective and demonstrates substantial advantages under extremely low bitrate conditions. Another line of work builds on the U-Net architecture by injecting guidance at various stages of the diffusion process. Kuang *et al.* [85] proposed a consistency-guided diffusion model, which extracts a syntax vector from the initial reconstruction and incorporates a Syntax-driven Feature Fusion (SFF) module to guide the multi-scale diffusion process within the U-Net backbone. This design jointly improves perceptual quality and fidelity. Li *et al.* [101] introduced a control module that conditions a frozen pre-trained diffusion model on latent representations of the reconstructed image. By training only the controller, the method achieves lightweight and efficient adaptation.

Other approaches draw inspiration from residual coding by modeling the residual between the original and reconstructed images using diffusion models [41]. This enables a plug-and-play post-processing module that significantly enhances perceptual quality without altering conventional metrics such as PSNR.

Within hybrid frameworks, recent advances address the high inference complexity through model simplification and encoder-side guidance. Hoogeboom *et al.* [60] introduced Rectified Flows [110] to design a lightweight post-processing module that reduces sampling steps while improving efficiency. Despite fewer steps, it outperforms standard diffusion and scales well with more iterations. Ma *et al.* [118] proposed a correction-guided mechanism that estimates the score function [148] or its linear approximation at the encoder side based on the original and initially reconstructed images. This estimated guidance is transmitted as auxiliary information to the decoder, steering the diffusion process initialized from the latent representation. The method improves reconstruction at low bitrates and significantly reduces decoding cost.

To conclude, diffusion models have redefined generative compression by turning data synthesis into a noise-to-signal mapping with clear interpretability. Their stepwise denoising paradigm stabilizes training and yields reconstructions of exceptional perceptual quality, while flexible noise scheduling enables broad extensions to conditional and multimodal settings. Yet these benefits are tempered by slow sampling and high computational overhead, leaving efficiency the key obstacle to practical deployment.

3.3 Implicit Modeling: Generative Adversarial Networks

Generative Adversarial Networks (GANs), one of the most influential generative paradigms in the deep learning era, were pioneered by Goodfellow *et al.* [42] in 2014. Rooted in game theory and the principle of adversarial training, GANs employ a dual-network architecture consisting of a generator G and a discriminator D to model complex data distributions. The generator $G(z) : \mathcal{Z} \to \mathcal{X}$ maps latent $z \sim p_z$ to the data space in order to synthesize samples that aim to deceive the discriminator, while the discriminator $D(x) : \mathcal{X} \to [0, 1]$ is trained to distinguish real data from generated samples. Formally, this adversarial mechanism is expressed as a minimax optimization problem:

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))], \tag{8}$$

which implicitly minimizes the JensenShannon (JS) divergence between the model distribution p_G and the data distribution p_{data}.

Unlike explicit density estimation and approximate inference frameworks, GANs characterize data distributions through adversarial training rather than tractable likelihoods. Build on this foundation, this section provides a systematic review of GAN-based generative image coding approaches, as summarized in Table 2. We classify these approaches into five key categories: (1) End-to-end Coding Frameworks with Adversarial Training; (2) Variable Bitrate and Rate Control Mechanisms; (3) Image Interpolation Structures Based on Dual Decoders; (4) Multi-modal Hierarchical Prior Modeling Frameworks; and (5) Discretized Representation Learning Based on VQ-GAN. These categories reflect a technological progression from "*Perceptual Optimization* ⇒ *Practicality Extension* ⇒ *Structural Innovation* ⇒ *Semantic Enhancement* ⇒ *Paradigm Shift in Discrete Representation*," showcasing the evolution of GAN-based generative image coding in performance, flexibility, and expressive capacity.

GAN-exclusive Decoder Training

Early studies in the domain did not alter the encoding-decoding architecture but instead focused on optimizing the decoder directly through adversarial training. Rippel *et al.* [135] first introduced GANs into lossy image compression, employing a pyramid-based autoencoder for feature extraction and a generator for high-detail image synthesis. Their method demonstrated the potential of GANs to surpass traditional codecs like JPEG in rate-distortion performance, while enabling real-time encoding and decoding via a lightweight framework. Santurkar *et al.* [137] formally introduced the concept of "Generative Compression" in the literature, merging the efficient encoding capabilities of VAEs with the high-quality reconstruction advantages of GANs. They further extended this idea to video, making one of the earliest attempts at generative video compression via frame-by-frame processing.

Table 2: Summary of GANs-based approaches for generative coding.

		Publication	Publication Venue	Remarks
GAN	**GAN-exclusive Decoder Training (End-to-end GAN Integrated Framework)**	Oren Rippel 2017 [135]	ICML	First GAN-based Image Compression Framework
		Shibani Santurkar 2018 [137]	PCS	First Attempt at Generative Video Compression
		Eirikur Agustsson 2019 [3]	ICCV	First Full-resolution Generative Image Compression Generative Compression with Partial Reconstruction
		Younhee Kim 2020 [78]	CVPR workshop	Hybrid VVC-GAN Architecture with Adversarial Loss
		Fabian Mentzer 2020 [122]	NIPS	Landmark GAN-based High-fidelity Compression
		Shubham Dash 2020 [133]	WACV	Stacked Autoencoders with Patch GAN
		Jooyoung Lee 2020 [88]	CVPR workshop	GAN-based Decoder and Enhancement Fine-tuning
		Zhengxue Cheng 2021 [21]	CVPR workshop	Enhanced Cheng2020 Framework with GAN Optimization
		Zeyu Yan 2021 [174]	ICML	Theoretical Foundations of Adversarial Training for Compression
		Bowen Liu 2021 [105]	CVPR	GAN + ConvLSTM latent prediction
		Saiping Zhang 2021 [185]	VCIP	Enhancing DVC via GAN-refined motion and residuals
		Dailan He 2022 [54]	CVPR	Enhanced ELIC Framework with GAN Optimization
	Variable Bitrate and Rate Control Design (Adaptive Rate Allocation Strategy)	Chaoyi Han 2020 [50]	TCSVT	Quantization on Channel Variance for Latent Compression
		Lirong Wu 2020 [169]	WACV	Importance Map-driven Bit Allocation Mechanism
		Yixin Gao 2021 [40]	CVPR	Adaptive spatial bit allocation with GAN enhancement
		Rushil Gupta 2022 [48]	CVPR workshop	Input: Importance Map for Bit Allocation
		Eirikur Agustsson 2023 [2]	CVPR	Conditional Decoder with Realism-aware Factor
		Shoma Iwai 2024 [68]	WACV	Rate Control via Factor-level Guided
	Dual-decoder Interpolation (Distortion-perceptual Loss Hybrid)	Michael Tschannen 2018 [158]	NIPS	WGAN-WVAE: Zero-to-perfect bitrate interpolation
		Hiroaki Akutsu 2020 [4]	CVPR workshop	Dual-loss Decoder (Adversarial + MS-SSIM Optimization)
		Shoma Iwai 2020 [69]	ICPR	Dual-loss Decoder (Adversarial + MSE Optimization)
		Zeyu Yan 2022 [173]	ICML	Theoretical Proof: DP-curve Coverage via Decoder Interpolation
		Nikolai Körber 2024 [81]	ECCV	Semantic-aware Discriminator with Segmentation Guidance
	Hierarchical Multimodal GAN Architecture (Cross-domain Representation Learning)	Danlan Huang 2022 [65]	JSAC	Semantic Segmentation-driven Rate Control
		Chen Zhu 2022 [191]	TCSVT	Perceptual frame enhancement via edge guidance
		Ren Yang 2022 [176]	IJCAI	Recurrent GAN exploiting motion and hidden states
		Fabian Mentzer 2022 [121]	ECCV	Dual-path GAN for inter-frame synthesis
		Jianhui Chang 2023 [17]	IJCV	Hierarchical Encoding: Structure + Texture
		Xuhao Jiang 2023 [75]	AAAI	Text Feature-guided Generation
	Vector-quantized GAN (Discrete Representation)	Matthew Muckley 2023 [126]	ICML	Non-binary Discriminator for VQ-VAE Representation
		Naifu Xue 2023 [171]	ICME	VQ Tokenization with Checkerboard Transformer
		Qi Mao 2024 [119]	DCC	VQ Tokenization with Autoregressive Modeling
		Anqi Li 2025 [90]	ICLR	Patch-based Encoding with Adaptive Rate Control

While early generative compression methods demonstrated feasibility, they were constrained to low resolutions (typically $\leq 64 \times 64$). To overcome this, Agustsson *et al.* [3] proposed a multi-scale framework with an enhanced discriminator, enabling the first successful compression of full-resolution Kodak [25] and Cityscapes [26] 2K images. They also introduced a selective compression strategy guided by semantic maps, preserving salient regions while aggressively compressing backgrounds, paving the way for perception- and semantics-aware coding. Building upon these, Mentzer *et al.* [122] made a landmark contribution with a refined conditional GAN framework, integrating hyperprior modules [9] while strategically simplifying the architecture. This approach demonstrated 50% bitrate reduction over MSE-optimized and conventional codecs at equivalent perceptual quality, establishing itself as the definitive baseline for subsequent research.

Subsequent research integrated GAN training into learned compression frameworks by incorporating adversarial modules as perceptual optimization components. Building on existing architectures, several studies advanced end-

to-end image coding through the integration of adversarial components: Kim *et al.* [78] extended VVC [14], Lee *et al.* [88] enhanced JointIQ-Net [87], Cheng *et al.* [21] improved upon Cheng2020 [22], and He *et al.* [54] advanced ELIC [53]. Parallel innovations include Dash *et al.* [133], who fused a Stacked Autoencoder with PatchGAN [67] via adversarial training, and Yan *et al.* [174], whose hybrid framework pairing an MSE-optimized encoder with an adversarial decoder outperformed conventional frameworks using distortion plus adversarial loss. The latter study further established seminal theoretical contributions: perfect perceptual quality necessitates doubling the minimal MSE distortion, classical rate-distortion encoders retain optimality in perceptual tasks, and distortion loss proves redundant for training a perceptual decoder. These insights provide a crucial theoretical basis for designing more efficient perceptual compression algorithms.

For video compression, Liu *et al.* [105] enhanced latent modeling by incorporating a GAN-augmented autoencoder with ConvLSTM-based prediction, achieving competitive compression performance while enabling downstream tasks such as anomaly detection. Zhang *et al.* [185] integrated perceptual enhancement into the end-to-end deep video codec (DVC) framework [113], using GANs to refine motion and residual reconstructions. These works represent early efforts to extend generative techniques from image to video domains, demonstrating the potential of adversarial training in capturing temporal consistency and enhancing perceptual quality.

Variable Bitrate and Rate Control Design

The above implementations still necessitate distinct training for each predefined rate configuration, leaving the fundamental rate-distortion trade-off underexplored. Notably, generative compression architectures exhibit inherent compatibility with content-aware compression strategies. Intuitively, they enable low-bitrate generative synthesis for perceptually less significant regions while allocating higher bitrates to semantically critical areas to ensure high-fidelity reconstruction. Among these advancements, Han *et al.* proposed channel-wise correlation variance masking to optimize redundancy in quantized latent spaces [50], while Wu *et al.* [169] developed perceptual significance mapping via importance matrices for quantization-aware bit allocation. Gao *et al.* [40] further advanced spatial scale hyperpriors in context modules for content-adaptive bitrate control, and adopted the gain unit and weighted quantization from G-VAE [27] to enable continuous variable bitrate compression. Gupta *et al.* [48] developed a GAN-based framework that integrates user-defined bitrate constraints and importance maps to guide bitrate allocation, significantly enhancing perceptual quality in semantically critical regions.

However, these approaches still rely on existing end-to-end variable-rate frameworks and have not achieved fundamental breakthroughs under the generative paradigm. Agustsson *et al.* [2] proposed a novel generative compression

framework that exploits the flexibility of conditional GANs. By introducing a tunable realism factor β, the model enables diverse reconstructions, ranging from low-distortion to high-perceptual quality, using a single latent representation. This approach achieves a state-of-the-art balance between distortion and perceptual fidelity. Building on this, Iwai *et al.* [68] further integrated a bitrate control variable *level* alongside β, resulting in a unified compression framework for both adaptive rate control and perceptual optimization.

Dual-decoder Interpolation

After addressing the limitation that existing methods require separately trained models for each predefined bitrate, researchers turned to more flexible architectures to tackle the distortionperception trade-off. A notable innovation is the dual-decoder framework, which employs two decoders independently optimized with distortion and perceptual objectives. As shown in Figure 7, this design enables interpolation between the two outputs, allowing for continuous and controllable balancing between distortion and perceptual quality within a generative compression framework.

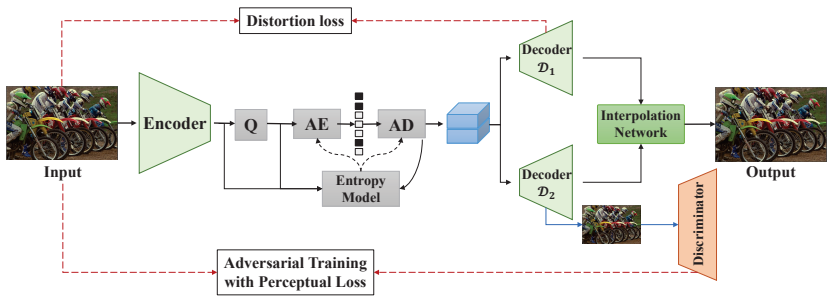


Figure 7: Schematic of the image compression framework based on dual-decoder interpolation. Q, AE, and AD denote the quantizer, arithmetic encoder, and arithmetic decoder, respectively.

Tschannen *et al.* [158] pioneered this approach by integrating Wasserstein Autoencoder [156] and Wasserstein GAN [5] principles, demonstrating seamless transitions between generative data modeling (zero bitrate) and exact reconstruction (high bitrates). Subsequent refinements include Akutsu *et al.*'s [4] GAN-conditioned primary decoder with selective detail decoding via weighted MS-SSIM optimization, augmented by causal attention and super-resolution modules. Iwai *et al.* [69] further advanced this paradigm through two-stage parameter interpolation that suppresses low-bitrate (< 0.1 bpp) artifacts. Notably, Yan *et al.* [173] provided theoretical validation, demonstrating that any point along the distortion-perception (D-P) tradeoff bound can be attained via a simple linear interpolation between the outputs of a min-

imum MSE decoder and a specially designed perfect perceptual decoder. The state-of-the-art is exemplified by Körber *et al.*'s [81] Output Residual Prediction (ORP) module, which dynamically adjusts the residual R between MSE- and GAN-optimized decoder outputs during reconstruction, achieving distortion performance comparable to VTM 20.0 while simultaneously advancing perceptual quality.

Hierarchical Multimodal GAN Architecture

Beyond decoder design, some studies focus on prior modeling and semantic guidance in encoding. Integrating multimodal information as auxiliary guidance enhances generative reconstruction quality, while hierarchical feature extraction improves compression efficiency, particularly at ultra-low bitrates. Several works have explored combining these strategies within generative compression frameworks.

Huang *et al.* [65] introduced a semantic-layered framework that encodes object categories, features, and spatial relations. A reinforcement learning-based bit allocation and a hierarchical adversarial decoder with cross-scale attention enable perceptually coherent reconstruction at low bitrates. Chang *et al.* [17] demonstrated superior performance via dual-layer semantic-texture disentanglement and GAN-driven synthesis, outperforming hybrid codecs in both perceptual quality and downstream vision task compatibility. Jiang *et al.* [75] enhanced text-guided compression with an Image-text Attention module for cross-modal feature fusion and a decoder-side module that adaptively integrates text priors. A multimodal semantic-consistency loss ensures alignment among reconstructed images, original inputs, and associated texts, leading to improved perceptual quality over HiFiC [122].

Extending these advances to video compression, Zhu *et al.* [191] guided texture generation via edge maps and motion estimation, improving visual quality. Mentzer *et al.* [121] proposed a GAN-based inter-frame synthesis framework combining UFlow and generative I-frame compression, with a dual-path discriminator supervising both structure and texture. Building on this, Yang *et al.* [176] designed a recurrent discriminator and generator architecture that jointly exploits motion, latent features, and ConvLSTM-based history. Their method significantly outperformed HEVC and several learned codecs in perceptual quality at low bitrates.

Vector-quantized GAN

Increasing attention has been paid to the discretization of the latent space in generative compression, aiming to more effectively capture semantic priors embedded in generative models. Compared to continuous latent variables, the discrete tokens introduced by VQ-GAN [35] significantly improve training stability and representation capacity.

Muckley *et al.* [126] pioneered a non-binary adversarial discriminator based on VQ-VAE autoencoders, integrating with the Mean-scale [123] framework to form the MS-ILLM architecture. This approach enhances statistical fidelity through locally conditioned latent modeling, achieving state-of-the-art rate-distortion-perception trade-offs. Xue and Mao *et al.* [171, 119] proposed a hierarchical tokenization scheme with transformer-based modeling, enabling ultra-low bitrate compression ($\leq$ 0.03 bpp) while preserving visual details. Notably, they made a pioneering attempt to unify entropy probability estimation with the probabilistic formulation of generative models, addressing a core limitation in prior methods. More recently, Li *et al.* [90] developed a variable-length VQ-GAN supporting fine-grained bitrate control via spatially adaptive VQ indices and non-parametric entropy coding. A probabilistic conditional decoder further improves reconstruction realism through hierarchical aggregation of multi-granularity representations, achieving optimal efficiency, perceptual quality, and adaptability.

Overall, GAN-based approaches have achieved striking progress in perceptual compression, enabling highly realistic reconstructions at high resolutions and substantially advancing generative quality. However, they also face persistent challenges, including unstable training dynamics, mode collapse, and inherent difficulty in probabilistic density modeling, which makes them less compatible with entropy modeling or ratedistortion optimization. Despite these limitations, the paradigm has significantly expanded the expressive power and application boundaries of generative models, marking GANs as a pivotal milestone in the development of generative compression research.

4 Generative Coding: Key Issues

The preceding survey highlights the advanced capabilities of generative coding in terms of perceptual quality, semantic consistency, and compression efficiency, demonstrating its feasibility across a wide range of scenarios. However, achieving optimal performance and enabling practical deployment of generative coding systems necessitates addressing several critical challenges. These challenges are essential to the successful application and further development of this emerging paradigm. The following subsections elaborate on these key issues.

Modeling Frameworks

One of the central challenges in current generative compression lies in the fact that its framework design remains immature and insufficiently well-defined, as shown in Figure 8. The key design is how to effectively balance perceptual generation with fidelity-oriented reconstruction.

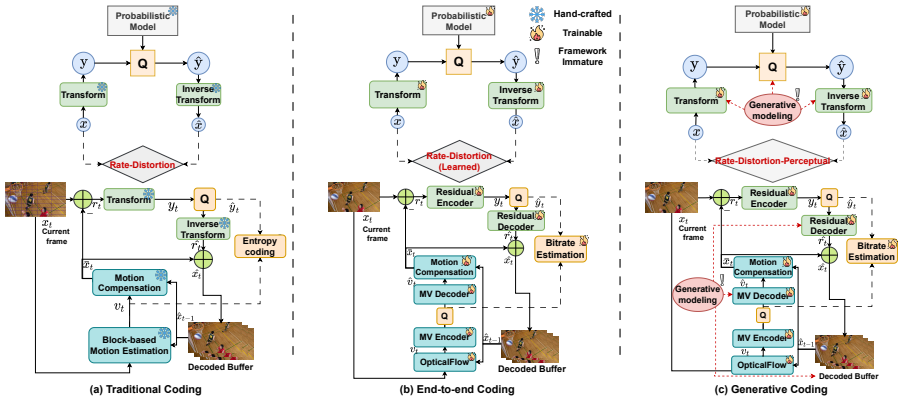


Figure 8: Illustration of image and video compression frameworks. From left to right: (a) traditional coding, (b) end-to-end coding, and (c) generative coding. The first row presents image compression pipelines, while the second row illustrates the corresponding video compression paradigms. Unlike traditional and end-to-end approaches with well-established frameworks, generative coding remains non-standardized and is an area of ongoing research.

Generative coding requires not only the generation of plausible content but also the preservation of the essential semantic and structural information from the input. In particular, clarifying the pathway from lossy compression to lossless representation is essential. This transition not only involves ensuring the interpretability and controllability of model architectures, but also directly impacts the reliability and deployability of compression systems in real-world applications.

In summary, although the previous section reviewed several efforts to explore and utilize prior information, the path toward establishing a unified and mature framework for generative compression is still far from complete. We believe that a promising direction toward addressing this gap is to focus on the design of generative models specifically tailored for coding tasks.

Metrics and Evaluation

The absence of effective evaluation metrics remains a critical barrier to the advancement and standardization of generative compression. Traditional low-level distortion measures, while sensitive to fine-grained structural variations, fail to capture semantic consistency and perceptual relevance, rendering them insufficient for evaluating the quality of generative reconstructions. Conversely, high-level perceptual metrics, although better aligned with human visual similarity, often compromise input fidelity and deviate from the fundamental objectives of compression. As illustrated in Table 3, each category of evaluation metric presents distinct trade-offs across dimensions. There is a pressing need for comprehensive, multi-level evaluation protocols that jointly

Table 3: Comparison of commonly used evaluation metrics in generative compression. Symbols ✓, ✗, and ● denote high, low, and medium performance respectively.

Metric	Semantic Consistency	Perceptual Quality	Fidelity	Interpretability	Widely Adoption
PSNR/SSIM [164]	✗	✗	✓	✓	✓
LPIPS [184]	✓	✓	✗	✗	✓
FID [56]/KID [11]	✓	✓	✗	✗	✓
VIF [141]	✗	●	✓	●	●
DISTS [29]	✗	✓	✓	✗	●
NIQE [125]	✗	●	✗	●	✓
mIoU	✓	✗	✗	✗	✗
VMAF [100]	✗	✓	●	✗	✓
Human Opinion Score	✓	✓	✓	✓	✗

account for semantic alignment, perceptual quality, and reconstruction accuracy.

A fundamental barrier lies in redefining the concept of "distortion" for generative coding, which should be tailored to meet the specific requirements of different applications. Key considerations include determining the degree of generated content that is acceptable to viewers under various bitrates and scene conditions. It is also important to recognize that human perception is not the only targetgenerative compression holds significant promise for machine-oriented applications (e.g., Image/Video coding for machine [34, 180, 151]), which opens up new and exciting directions for future research.

Complexity and Deployment Feasibility

Generative models are typically large in scale, incurring substantial training and inference costs, which hinders their practical deployment. For any compression standard to be viable in real-world scenarios, lightweight and low-cost deployment on edge or terminal devices is essential. While some workssuch as those based on GANshave demonstrated real-time encoding and decoding capabilities [135, 128], the overall deployment landscape remains challenging.

Table 4 provides a comparative analysis of encoding and decoding times for representative traditional, end-to-end, GAN-based, and diffusion-based compression methods on the Kodak [25] dataset. As shown, traditional codecs like VTM [14] suffer from prohibitively high encoding times, whereas GAN-based approaches such as HiFiC [122] and MS-ILLM [126] achieve significantly lower latency. However, diffusion-based methods (e.g., CDC [177] and DiffEIC [101]), though promising in perceptual quality, remain computationally intensive, especially during decoding.

The future viability of generative compression as a practical standard hinges on the development of model designs that balance performance and computational efficiency. This includes advancing key technologies such as model compression [28], quantization [47, 138], and adaptation to heteroge-

Table 4: Model parameters, and encoding/decoding speed (in seconds) on the Kodak [25] dataset. Values are reported as mean ± standard deviation.

Category	Method	Params	BPP	Encoding Time	Decoding Time	Platform	Config
Traditional	VTM [14]	–	0.406	60.871 ± 22.461	0.158 ± 0.014	Intel Xeon Silver 4310	QP = 32
End-to-end	ELIC [53]	34M	0.497	2.878 ± 0.188	2.792 ± 0.173	NVIDIA RTX 3090	–
GAN Based	HiFiC [122]	185M	0.303	0.660 ± 0.164	1.435 ± 0.210	NVIDIA RTX 3090	–
	MS-ILLM [126]	181.5M	0.433	0.171 ± 0.388	0.086 ± 0.024	NVIDIA RTX 3090	–
Diffusion Based	CDC [177]	850M	0.485	–	3.324 ± 0.031	NVIDIA RTX 3090	Denoising steps=50
	DiffEIC [101]	950M	0.123	0.330 ± 0.383	6.854 ± 0.128	NVIDIA RTX 3090	Denoising steps=50

neous hardware platforms [70, 46]. Ultimately, these advances will facilitate a meaningful transition from centralized systems to edge-level deployment.

5 Experiment

In this section, we present a comprehensive evaluation to assess the performance and feasibility of generative compression. To better illustrate the current landscape, we first include two case studies as representative examples of emerging directions. Generative video compression is still in its early stage, and most existing advances are concentrated on specific domains such as talking-face or human-body videos. In this context, our first study demonstrates the effectiveness of generative compression for talking-face videos, highlighting the potential in specialized domains. Furthermore, leveraging large multimodal language models (MLLMs) for video coding remains a bold and largely unexplored attempt. Our second study provides a preliminary exploration in this direction, marking one of the first unified applications of foundational MLLMs and generative models in video coding. Second, we conduct extensive comparisons with representative baselines in the field to benchmark and highlight their strengths and limitations.

5.1 Our Attempts at Generative Compression

5.1.1 Efficient Compression for Talking-face Videos

Talking-face videos exhibit strong spatio-temporal regularities, offering significant opportunities for enhanced compression efficiency in communication systems. Traditional model-based approaches have demonstrated the potential for ultra-low bitrate transmission. More recent advancements employ 3D techniques [24] and deep generative frameworks [42, 142, 162] to further improve rate-distortion performance. However, these methods typically rely on hand-crafted and explicit representations, such as facial landmarks or keypoints, which constrain their ability to capture complex motion dynamics

and limit their integration with established hybrid video coding standards. This underscores the need for a more compact and expressive representation that can effectively model temporal trajectories in talking-face videos. Such a representation should facilitate high-fidelity reconstruction under constrained bandwidth conditions and align with the hybrid video coding paradigm to ensure compatibility and robustness in practical deployment scenarios.

For this purpose, we design a novel talking-face video compression framework based on Compact Temporal Trajectory Representation (CTTR) [19], which replaces explicit, hand-crafted representations with a learned, compact encoding of temporal motion. By capturing the intrinsic dynamics of facial movements through data-driven modeling, CTTR enables more efficient and robust video reconstruction. To further enhance performance, we introduce a dynamic reference refresh mechanism and enforce temporal consistency via end-to-end training. This design not only achieves superior rate-distortion performance at ultra-low bitrates, but also ensures compatibility with traditional hybrid video coding frameworks.

As shown in Figure 9, the encoder side of the proposed framework integrates an intra-coding scheme based on traditional hybrid video coding to compress key-reference frames, a compact feature representation module to model the dynamic trajectory variations of inter frames, and a context-based feature encoding module for efficient feature compression via inter prediction and entropy coding. An identified key-reference frame of a specific sequence is compressed using the VVC encoder, which provides the fundamental texture representation. Based on this reference, the remaining frames are represented through a learned compact feature space. Both the reconstructed key-reference and inter frames are projected into this highly compact latent space. The resulting feature representations are then processed through inter prediction, quantization, and entropy coding. On the decoder side, the key-reference frames are reconstructed via VVC decoding, while inter-frame features are recovered through context-based decoding and feature compensation. A sparse motion field is then constructed from the reconstructed features and refined into a dense motion map and occlusion map. These, together with the decoded key-reference frames, are fed into a neural frame generation module to produce high-quality talking-head video with realistic motion and appearance.

As shown in Table 5, the proposed CTTR framework achieves remarkable bitrate savings compared to the latest VVC codec, with reductions of 78.84% in Rate-DISTS, 73.00% in Rate-LPIPS, and 77.72% in Rate-FID at a resolution of 256×256. Additionally, CTTR outperforms the state-of-the-art generative methods Face_vid2vid [162] and CFTE [18], offering over 20% bitrate savings across all three deep learning-based quality metrics. These results demonstrate the superior compression efficiency of CTTR under deep perceptual quality metrics.

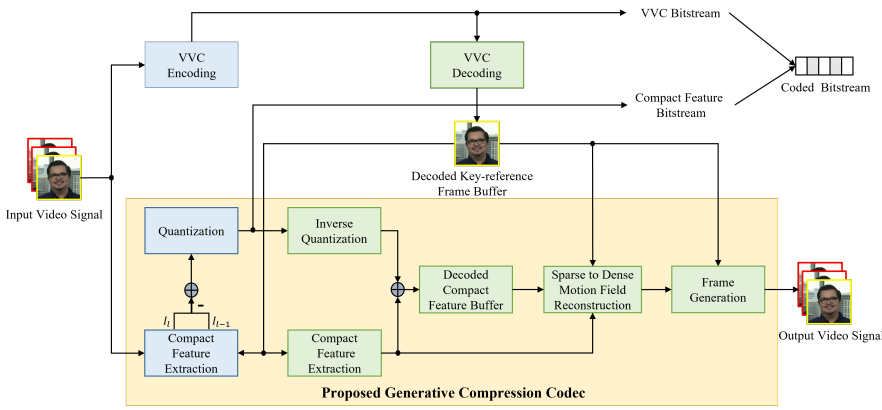


Figure 9: Overview of proposed generative compression framework for talking-face video.

Table 5: Average bit-rate savings of 40 talking-face sequences at the resolution of 256×256. Anchor: CTTR (ours).

Anchors	Rate-DISTS	Rate-LPIPS	Rate-FID
VVC [14]	-78.84%	-73.00%	-77.72%
FOMM [142]	-64.19%	-65.80%	-64.85%
FOMM2.0 [143]	-43.54%	-46.28%	-43.02%
Face_vid2vid [162]	-24.70%	-30.80%	-25.67%
CFTE [18]	-24.32%	-25.80%	-28.13%

Moreover, to demonstrate the coding efficiency, we evaluate trade-offs in model size (Params), computational cost measured in multiply-adds (MAdd), and actual inference time on both the encoder and decoder. Experiments are conducted on a Tesla A100 GPU with 15-core Intel Xeon Platinum 8369B CPUs, using a test sequence of 250 frames at 256Œ256 resolution. Encoder and decoder times are averaged over five quantization parameters (QPs) (32, 37, 42, 47, and 52). As shown in Table 6 , compared with the traditional VVC codec, generative compression schemes substantially reduce encoder-side inference time, although decoding complexity increases. Among these methods, the proposed CTTR approach achieves a favorable balance between Params, MAdd, and encoder-side Time, while decoding efficiency is slightly affected by the inclusion of an additional temporal discriminator.

5.1.2 Cross-modality Video Coding with MLLMs

Traditional approaches of video compression often struggle to maintain semantic integrity and perceptual quality under tight bitrate constraints. Meanwhile,

Table 6: Complexity comparison in terms of model size (Params), computational cost (MAdd), and inference time (Time) at both encoder and decoder sides.

	Measures	VVC [14]	FOMM [142]	FOMM2.0 [143]	Face_vid2vid [162]	CFTE [18]	CTTR(Ours)
Encoding	Params(M)	-	14.215	14.147	72.195	14.131	14.131
	MAdd(G)	-	1.277	1.059	317.324	0.989	0.989
	Time(Sec)	1323.495	21.065	19.648	18.829	16.402	16.095
Decoding	Params(M)	-	45.575	45.575	53.101	43.885	43.910
	MAdd(G)	-	53.643	53.643	178.282	54.822	189.858
	Time(Sec)	0.650	14.274	14.257	20.093	13.232	22.880

Multimodal Large Language Models (MLLMs) [51, 136] have shown a remarkable ability to process sequential inputs and understand temporal patterns in multimodal data. Their potential to encode rich semantic cues from video content opens new possibilities for video compression, particularly for generating compact representations that preserve meaning and structure, even at extremely low bitrates.

In light of this, we introduce a novel Cross-modality Video Coding (CMVC) framework that incorporates MLLMs and video generation techniques to achieve efficient and flexible video coding [183]. CMVC supports two tailored reconstruction modes: Text-text-to-video (TT2V), which uses text-based representations to reconstruct video content with strong semantic alignment at ultra-low bitrate, and Image-text-to-video (IT2V), which combines text with keyframes to enhance visual fidelity and temporal consistency at extremely low bitrate. The IT2V mode further benefits from Low-rank Adaptation (LoRA) [62] based frame interpolation for smoother transitions.

The pipeline of the proposed CMVC framework is illustrated as Figure 10. It consists of an encoder-decoder architecture designed to enable efficient video compression with high semantic and perceptual fidelity. In the encoding stage, we apply a keyframe selection strategy to segment the video into discrete clips, as in part (a) of Figure 10. This facilitates the extraction of spatial features from keyframes and temporal information from the motion between them. CLIP-based similarity metrics are used to ensure coverage of salient temporal changes. Both the selected keyframes and the corresponding motion between them are then transformed into compact textual representations via V2T models, forming the core multimodal inputs for compression, as in part (b) of Figure 10. The decoder supports two distinct reconstruction modes tailored to different bit-rate and quality requirements. As illustrated in part (c) of Figure 10, the TT2V mode reconstructs video content solely from text, emphasizing semantic coherence and enabling ultra-low bitrate transmission. In contrast, as shown in part (d) of Figure 10, the IT2V mode incorporates both decoded keyframe images and motion descriptions to enhance perceptual quality, and further applies a Stable Diffusion-based generative model with LoRA tuning to refine temporal consistency across frames. By leveraging the cross-modal representation capabilities of foundational MLLMs, the proposed

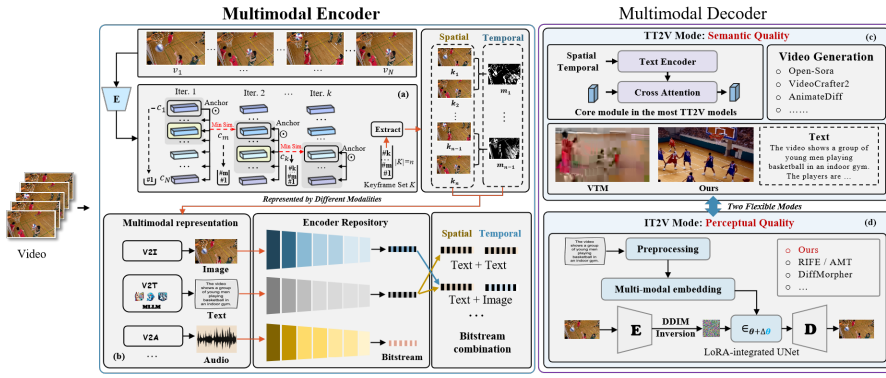


Figure 10: The framework of the proposed CMVC scheme. The input video is divided into clips via keyframe selection, from which spatial and temporal priors are extracted. Multimodal representations are obtained using MLLMs and compressed into bitstreams by the encoder (E). At the decoder side (D), we demonstrate two modes: TT2V for semantic quality and IT2V for perceptual quality, both enabling flexible integration with state-of-the-art generative models.

CMVC framework provides a unified and adaptable solution for video coding under extreme bitrate constraints.

In the IT2V mode, the proposed CMVC framework is compared against traditional codecs (x264 [167], x265 [149], VTM [14]), deep learning-based codecs (DCVC [95], DCVC-DC [96]), and state-of-the-art video generation models (RIFE [66], AMT [99], DiffMorpher [181]). As shown in Figure 11, CMVC consistently achieves superior perceptual quality, particularly under extremely low bitrate constraints, where existing pretrained deep codecs show limitations. Our approach demonstrates stable and high-quality reconstruction across various datasets, with improved bitrate control via keyframe selection and quality adjustment. This work marks the first unified application of foundational MLLMs and generative models in video coding. Nevertheless, CMVC entails substantially higher complexity due to the heavy computational demands of large multimodal language models. In the IT2V mode, the encoding and decoding processes take considerably longer, requiring 722.54 and 211.25 seconds, respectively, when utilizing 4 NVIDIA GeForce RTX 3090 GPUs. We present this study as an initial attempt, underscoring both the potential and challenges of this direction.

5.2 Benchmarking Against State-of-the-art Methods

We conducted a comprehensive evaluation by comparing several representative generative image compression methods, alongside neural compression

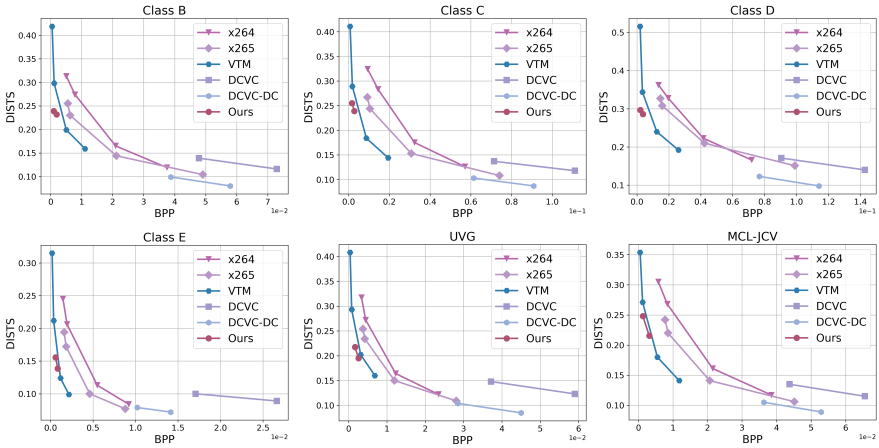


Figure 11: The R-D performance comparison in the IT2V mode. The comparisons are performed on the Class B, Class C, Class D, Class E, UVG, and MCL-JCV, respectively. In the figures, "Ours" refers to the proposed CMVC framework.

techniques and traditional approaches. The generative image compression methods considered include:

1. High Fidelity Compression (HiFiC) [122], trained for three target bitrates $r_t \in \{0.14, 0.30, 0.45\}$.

2. Conditional Diffusion Compression (CDC) [177] consists of two variants: χ-prediction and ϵ-prediction. The χ-prediction model achieves compression performance similar to that of ϵ-prediction, requiring only a few decoding steps (compared to hundreds) and without the need for post-processing. The implementation includes two model configurations controlled by the hyperparameter ρ, where $\rho = 0.9$ and $\rho = 0.0$ represent different perceptual loss weightings. Here, we focus on the χ-prediction variant with a denoising step set to 50.

3. Diffusion-based Extreme Image Compression (DiffEIC) [101], trained for five target bitrates $r_t \in \{0.02, 0.04, 0.06, 0.09, 0.12\}$.

4. Mean-scale implicit local likelihood model (MS-ILLM) [126], which is a VQGAN-based method that improves statistical fidelity using local adversarial discriminators.

5. Multi-realism Image Compression (MRIC) [2], which aims to train a single generator G to adapt to any realism weight $\beta \in [0, 2.56]$, where $\beta = 0$ corresponds to the lowest distortion and $\beta = \beta_{max}$ corresponds to the highest perceptual realism in the reconstruction.

6. Perceptual compression (PerCo(SD)) [82, 16], which is conditioned on both a vector-quantized latent image representation and a textual image description.

7. Rate-distortion Optimization for Cross Modal Compression (RDO-CMC) [39], in which we employed its variant based on a diffusion model for the decoder.

8. Unified Image Generation-compression (UIGC) [171], which is based on a vector-quantized representation and employs a multi-stage transformer architecture for extreme compression.

For comparison, we include the neural image compression method ELIC [53], as well as traditional methods VVC [14], implemented using the reference software VTM-23.0 with intra configuration, and BPG [10]. And the QPs for BPG and VTM is set to $\{30, 35, 38, 42, 50\}$. We assess the aforementioned methods using the Kodak image dataset [25], which comprises 24 images, each with a resolution of 768×512. To ensure a fair and comprehensive evaluation of the compression performance of each model, we employed both PSNR and MS-SSIM [165] to assess distortion, and LPIPS [184] and FID [56] to evaluate perceptual quality. Specifically, we utilized the AlexNet-based [83] LPIPS for perceptual distance. Additionally, considering that variations in low-level preprocessing, such as image resizing and patching can introduce significant differences and unexpected effects on the FID metric, we adopted the clean-FID implementation [129] to mitigate such issues.

As shown in Figure 12, generative models generally exhibit superior perceptual quality at lower bitrates, but tend to underperform in terms of fidelity compared to traditional and end-to-end learned methods. Among the generative approaches, MS-ILLM offers a favorable trade-off between perceptual quality and fidelity. Compared to VTM, it achieves an 86.25% improvement in LPIPS-BD rate, at the expense of a 70.72% increase in PSNR-BD rate.

6 Conclusion

In this paper, we present a comprehensive review of generative coding, an emerging paradigm that leverages generative models for efficient visual data compression. We provide a principled formulation that emphasizes the intrinsic theoretical connections between generation and compression, highlighting their potential for mutual integration and cross-fertilization. Through a systematic survey categorized by the underlying generative model types, we trace the technical evolution of representative methods and distill their core innovations. We also present experimental studies, including both our own ex-

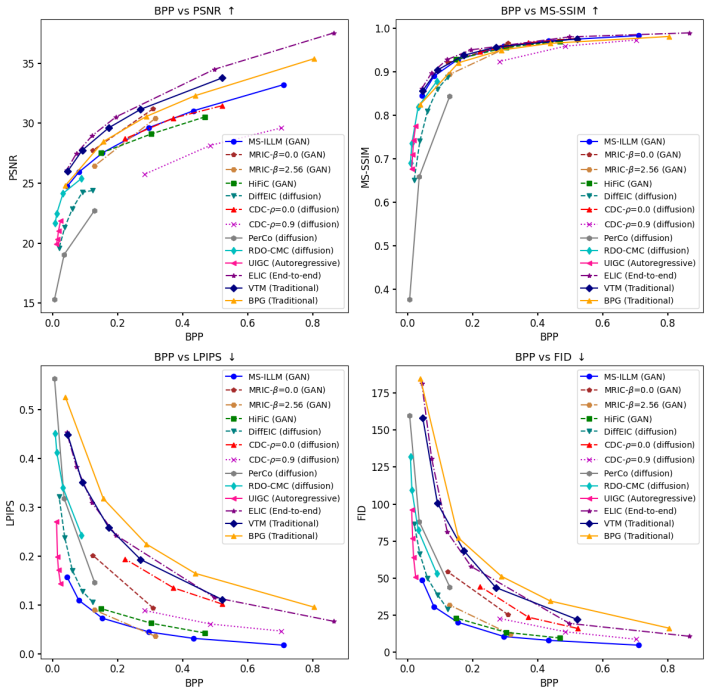


Figure 12: Tradeoffs between bitrate (x-axes, in bpp) and different metrics (y-axes) for various models tested on Kodak. The labels indicate the paradigm of each method.

plorations and comparisons with representative baselines, offering practical insights into the feasibility and limitations of current approaches.

Generative coding is poised to reshape the future of visual communication, particularly in the context of foundation models and machine-centric media understanding. Based on current developments, we believe that the following directions are particularly promising for advancing generative compression research:

- Paradigm Innovation in Generative Video Compression: While generative image compression has witnessed notable advances, its video counterpart remains at a nascent stage, often constrained within the modular replacement of conventional codecs (e.g., motion estimation or reference reconstruction). Most existing approaches follow a frame-by-frame progression built upon motion compensation, which inherently limits their temporal modeling and compression efficiency.

- Leveraging LLMs for Lossy Compression: Large language models have shown strong capabilities in cross-modal understanding and structured generation. While early attempts have applied LLMs to lossless image compression, their use in lossy visual compression remains underexplored. Future work may investigate LLM-driven compression pipelines, where semantic tokenization and reconstruction are guided by language-based priors.

- Deployability and Edge-cloud Adaptation: High computational cost and low interpretability limit the practical use of generative coding. To enable deployment across edge and cloud systems, future work should focus on lightweight, efficient decoders, model compression (e.g., distillation, quantization), and balancing computation with coding performance.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report", *arXiv preprint arXiv:2303.08774*, 2023.

[2] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, "Multi-realism image compression with a conditional generator", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 22324–33.

[3] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression", in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, 221–31.

[4] H. Akutsu, A. Suzuki, Z. Zhong, and K. Aizawa, "Ultra low bitrate learned image compression by selective detail decoding", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, 118–9.

[5] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks", in *International Conference on Machine Learning*, PMLR, 2017, 214–23.

[6] T. Bachard, T. Bordin, and T. Maugey, "CoCliCo: Extremely low bitrate image compression based on CLIP semantic and tiny color map", in *2024 Picture Coding Symposium (PCS)*, IEEE, 2024, 1–5.

[7] J. Ballé, V. Laparra, and E. Simoncelli, "End-to-end optimized image compression", in *International Conference on Learning Representations*, 2016.

[8] J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding", *IEEE Journal of Selected Topics in Signal Processing*, 15(2), 2020, 339–53.

[9] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior", *arXiv preprint arXiv:1802.01436*, 2018.

[10] F. Bellard, "BPG Image format", *URL https://bellard.org/bpg*, 2015.

[11] M. Bikowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans", *arXiv preprint arXiv:1801.01401*, 2018.

[12] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman, "Token merging: Your vit but faster", *arXiv preprint arXiv:2210.09461*, 2022.

[13] T. Bordin and T. Maugey, "Semantic based generative compression of images for extremely low bitrates", in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2023, 1–6.

[14] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications", *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10), 2021, 3736–64.

[15] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt", *arXiv preprint arXiv:2303.04226*, 2023.

[16] M. Careil, M. J. Muckley, J. Verbeek, and S. Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates", in *The Twelfth International Conference on Learning Representations*, 2023.

[17] J. Chang, J. Zhang, J. Li, S. Wang, Q. Mao, C. Jia, S. Ma, and W. Gao, "Semantic-aware visual decomposition for image coding", *International Journal of Computer Vision*, 131(9), 2023, 2333–55.

[18] B. Chen, Z. Wang, B. Li, R. Lin, S. Wang, and Y. Ye, "Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression", in *2022 Data Compression Conference (DCC)*, IEEE, 2022, 13–22.

[19] B. Chen, Z. Wang, B. Li, S. Wang, and Y. Ye, "Compact Temporal Trajectory Representation for Talking Face Video Compression", *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11), 2023, 7009–23.

[20] K. Chen, P. Zhang, H. Liu, J. Liu, Y. Liu, J. Huang, S. Wang, H. Yan, and H. Li, "Large Language Models for Lossless Image Compression: Next-Pixel Prediction in Language Space is All You Need", *arXiv preprint arXiv:2411.12448*, 2024.

[21] Z. Cheng, T. Fu, J. Hu, L. Guo, S. Wang, X. Zhao, D. Zhou, and Y. Song, "Perceptual image compression using relativistic average least squares gans", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 1895–900.

[22] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 7939–48.

[23] W.-J. Chien, J. Boyce, Y.-W. Chen, R. Chernyak, K. Choi, R. Hashimoto, Y.-W. Huang, H. Jang, R.-L. Liao, and S. Liu, "JVET AHG report: Tool reporting procedure and testing (AHG13)", *The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, doc. no. JVET-T0013*, October, October 2020.

[24] M. F. Chowdhury, A. F. Clark, A. C. Downton, E. Morimatsu, and D. E. Pearson, "A switched model-based coder for video signals", *IEEE Transactions on Circuits and Systems for Video Technology*, 4(3), 2002, 216–27.

[25] E. K. Company, "Kodak Lossless True Color Image Suite", http://r0k.us/graphics/kodak, 1999.

[26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 3213–23.

[27] Z. Cui, J. Wang, B. Bai, T. Guo, and Y. Feng, "G-vae: A continuously variable rate deep image compression framework", *arXiv preprint arXiv:2003.02012*, 2(3), 2020, 4.

[28] P. V. Dantas, W. S. da Silva Jr, L. C. Cordeiro, and C. B. Carvalho, "A comprehensive review of model compression techniques in machine learning", *APPLIED INTELLIGENCE*, 54(22), November 2024, 11804–44, ISSN: 0924-669X.

[29] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image Quality Assessment: Unifying Structure and Texture Similarity", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5), 2022, 2567–81.

[30] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation", *arXiv preprint arXiv:1410.8516*, 2014.

[31] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp", *arXiv preprint arXiv:1605.08803*, 2016.

[32] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "PaLM-E: An Embodied Multimodal Language Model", in *International Conference on Machine Learning*, PMLR, 2023, 8469–88.

[33] J. Du, C. Zhou, N. Cao, G. Chen, Y. Chen, Z. Cheng, L. Song, G. Lu, and W. Zhang, "Large Language Model for Lossless Image Compression with Visual Prompts", *arXiv preprint arXiv:2502.16163*, 2025.

[34] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics", *IEEE Transactions on Image Processing*, 29, 2020, 8680–95.

[35] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 12873–83.

[36] L. G. Foo, H. Rahmani, and J. Liu, "Ai-generated content (aigc) for various data modalities: A survey", *ACM Computing Surveys*, 57(9), 2025, 1–66.

[37] F. Gao, X. Deng, J. Jing, X. Zou, and M. Xu, "Extremely low bit-rate image compression via invertible image generation", *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8), 2023, 6993–7004.

[38] J. Gao, C. Jia, Z. Huang, S. Wang, S. Ma, and W. Gao, "Rate-distortion optimized cross modal compression with multiple domains", *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8), 2024, 6978–92.

[39] J. Gao, C. Jia, S. Wang, S. Ma, and W. Gao, "Rate-distortion optimization for cross modal compression", in *2023 Data Compression Conference (DCC)*, IEEE, 2023, 218–27.

[40] Y. Gao, Y. Wu, Z. Guo, Z. Zhang, and Z. Chen, "Perceptual friendly variable rate image compression", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 1916–20.

[41] N. F. Ghouse, J. Petersen, A. Wiggers, T. Xu, and G. Sautière, "A residual diffusion model for high perceptual quality codec augmentation", *arXiv preprint arXiv:2301.05489*, 2023.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", in *Advances in Neural Information Processing Systems*, 2014, 2672–80.

[43] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, "The llama 3 herd of models", *arXiv preprint arXiv:2407.21783*, 2024.

[44] R. Gray, "Vector quantization", *IEEE Assp Magazine*, 1(2), 1984, 4–29.

[45] K. Gregor, F. Besse, D. J. Rezende, I. Danihelka, and D. Wierstra, "Towards conceptual compression", in *Advances in Neural Information Processing Systems*, 2016, 3549–57.

[46] Y. Guo, W. Gao, S. Ma, and G. Li, "Accelerating Transform Algorithm Implementation for Efficient Intra Coding of 8K UHD Videos", 18(4), March 2022, ISSN: 1551-6857.

[47] Z. Guo, Z. Zhang, R. Feng, and Z. Chen, "Soft then Hard: Rethinking the Quantization in Neural Image Compression", in *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, *Proceedings of Machine Learning Research*, PMLR, 18–24 Jul 2021, 3920–9.

[48] R. Gupta, S. BV, N. Kapoor, R. Jaiswal, S. R. Nangi, and K. Kulkarni, "User-guided variable rate learned image compression", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 1753–8.

[49] A. Habibian, T. v. Rozendaal, J. M. Tomczak, and T. S. Cohen, "Video compression with rate-distortion autoencoders", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, 7033–42.

[50] C. Han, Y. Duan, X. Tao, M. Xu, and J. Lu, "Toward variable-rate generative compression by reducing the channel redundancy", *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7), 2020, 1789–802.

[51] L. Han, J. Ren, H.-Y. Lee, F. Barbieri, K. Olszewski, S. Minaee, D. Metaxas, and S. Tulyakov, "Show me what and tell me how: Video synthesis via multimodal conditioning", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 3615–25.

[52] X. Han, M. Ghazvininejad, P. W. Koh, and Y. Tsvetkov, "Jpeg-lm: Llms as image generators with canonical codec representations", *arXiv preprint arXiv:2408.08459*, 2024.

[53] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 5718–27.

[54] D. He, Z. Yang, H. Yu, T. Xu, J. Luo, Y. Chen, C. Gao, X. Shi, H. Qin, and Y. Wang, "Po-elic: Perception-oriented efficient learned image coding", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 1764–9.

[55] L. Helminger, A. Djelouah, M. Gross, and C. Schroers, "Lossy Image Compression with Normalizing Flows", in *Neural Compression Workshop at ICLR 2021*, 2021.

[56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium", *Advances in Neural Information Processing Systems*, 30, 2017.

[57] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen Video: High Definition Video Generation with Diffusion Models", *arXiv preprint arXiv:2210.02303*, abs/2210.02303, 2022.

[58] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", en-US, *Advances in Neural Information Processing Systems*, 33 (January), January 2020, 6840–51.

[59] Y.-H. Ho, C.-C. Chan, W.-H. Peng, H.-M. Hang, and M. Domaski, "Anfic: Image compression using augmented normalizing flows", *IEEE Open Journal of Circuits and Systems*, 2, 2021, 613–26.

[60] E. Hoogeboom, E. Agustsson, F. Mentzer, L. Versari, G. Toderici, and L. Theis, "High-Fidelity Image Compression with Score-based Generative Models", *arXiv preprint arXiv:2305.18231*, 2023.

[61] E. Hoogeboom, J. Peters, R. Van Den Berg, and M. Welling, "Integer discrete flows and lossless compression", *Advances in Neural Information Processing Systems*, 32, 2019.

[62] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.", *ICLR*, 1(2), 2022, 3.

[63] Y. Hu, W. Yang, Z. Ma, and J. Liu, "Learning end-to-end lossy image compression: A benchmark", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[64] Z. Hu, G. Lu, and D. Xu, "FVC: A New Framework towards Deep Video Compression in Feature Space", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 1502–11.

[65] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding", *IEEE Journal on Selected Areas in Communications*, 41(1), 2022, 55–71.

[66] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-Time Intermediate Flow Estimation for Video Frame Interpolation", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[67] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 1125–34.

[68] S. Iwai, T. Miyazaki, and S. Omachi, "Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, 2900–9.

[69] S. Iwai, T. Miyazaki, Y. Sugaya, and S. Omachi, "Fidelity-controllable extreme image compression with generative adversarial networks", in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, IEEE, 2020, 8235–42.

[70] C. Jia, X. Hang, S. Wang, Y. Wu, S. Ma, and W. Gao, "FPX-NIC: An FPGA-Accelerated 4K Ultra-High-Definition Neural Video Coding System", *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9), 2022, 6385–99.

[71] C. Jia, S. Wang, X. Zhang, S. Wang, J. Liu, S. Pu, and S. Ma, "Content-aware convolutional neural network for in-loop filtering in high efficiency video coding", *IEEE Transactions on Image Processing*, 28(7), 2019, 3343–56.

[72] C. Jia, S. Wang, X. Zhang, S. Wang, and S. Ma, "Spatial-temporal residue network based in-loop filter for video coding", in *IEEE Visual Communications and Image Processing*, 2017, 1–4.

[73] Z. Jia, J. Li, B. Li, H. Li, and Y. Lu, "Generative Latent Coding for Ultra-Low Bitrate Image Compression", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, 26088–98.

[74] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao, "An end-to-end compression framework based on convolutional neural networks", *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10), 2017, 3007–18.

[75] X. Jiang, W. Tan, T. Tan, B. Yan, and L. Shen, "Multi-modality deep network for extreme learned image compression", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 1, 2023, 1033–41.

[76] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review", *IEEE transactions on visualization and computer graphics*, 26(11), 2019, 3365–85.

[77] J. Kim, "The institutionalization of YouTube: From user-generated content to professionally generated content", *Media, culture & society*, 34(1), 2012, 53–67.

[78] Y. Kim, S. Cho, J. Lee, S.-Y. Jeong, J. S. Choi, and J. Do, "Towards the perceptual quality enhancement of low bit-rate compressed images", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, 136–7.

[79] D. P. Kingma and M. Welling, "Auto-encoding variational bayes", *arXiv preprint arXiv:1312.6114*, 2013.

[80] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, 4015–26.

[81] N. Körber, E. Kromer, A. Siebert, S. Hauke, D. Mueller-Gritschneder, and B. Schuller, "Egic: enhanced low-bit-rate generative image compression guided by semantic segmentation", in *European Conference on Computer Vision*, Springer, 2024, 202–20.

[82] N. Körber, E. Kromer, A. Siebert, S. Hauke, D. Mueller-Gritschneder, and B. Schuller, "PerCo (SD): Open Perceptual Compression", *arXiv preprint arXiv:2409.20255*, 2024.

[83] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks", *arXiv preprint arXiv:1404.5997*, 2014.

[84] J. Krumm, N. Davies, and C. Narayanaswami, "User-generated content", *IEEE Pervasive Computing*, 7(4), 2008, 10–11.

[85] H. Kuang, Y. Ma, W. Yang, Z. Guo, and J. Liu, "Consistency Guided Diffusion Model with Neural Syntax for Perceptual Image Compression", in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, 1622–31.

[86] J. Lee, S. Cho, and S.-K. Beack, "Context-adaptive entropy model for end-to-end optimized image compression", *arXiv preprint arXiv:1809.10452*, 2018.

[87] J. Lee, S. Cho, and M. Kim, "An end-to-end joint learning scheme of image compression and quality enhancement with improved entropy minimization", *arXiv preprint arXiv:1912.12817*, 2019.

[88] J. Lee, D. Kim, Y. Kim, H. Kwon, J. Kim, and T. Lee, "A training method for image compression networks to improve perceptual quality of reconstructions", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, 144–5.

[89] E. Lei, Y. B. Uslu, H. Hassani, and S. S. Bidokhti, "Text+ sketch: Image compression at ultra low rates", *arXiv preprint arXiv:2307.01944*, 2023.

[90] A. Li, F. Li, Y. Liu, R. Cong, Y. Zhao, and H. Bai, "Once-for-All: Controllable Generative Image Compression with Dynamic Granularity Adaption", *arXiv preprint arXiv:2406.00758*, 2024.

[91] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, *et al.*, "Llava-onevision: Easy visual task transfer", *arXiv preprint arXiv:2408.03326*, 2024.

[92] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation", *Advances in Neural Information Processing Systems*, 32, 2019.

[93] C. Li, G. Lu, D. Feng, H. Wu, Z. Zhang, X. Liu, G. Zhai, W. Lin, and W. Zhang, "Misc: Ultra-low bitrate image semantic compression driven by large multimodal model", *IEEE Transactions on Image Processing*, 2024.

[94] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models", *Neurocomputing*, 479, 2022, 47–59.

[95] J. Li, B. Li, and Y. Lu, "Deep Contextual Video Compression", in *Advances in Neural Information Processing Systems*, ed. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Vol. 34, Curran Associates, Inc., 2021, 18114–25.

[96] J. Li, B. Li, and Y. Lu, "Neural Video Compression With Diverse Contexts", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, 22616–26.

[97] J. Li, B. Li, and Y. Lu, "Neural video compression with feature modulation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 26099–108.

[98] J. Li, B. Li, J. Xu, and R. Xiong, "Efficient multiple-line-based intra prediction for HEVC", *IEEE Transactions on Circuits and Systems for Video Technology*, 28(4), 2016, 947–57.

[99] Z. Li, Z.-L. Zhu, L.-H. Han, Q. Hou, C.-L. Guo, and M.-M. Cheng, "AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, 9801–10.

[100] Z. Li, A. Aaron, and C. Bampis, "Toward a Practical Perceptual Video Quality Metric", Netflix Tech Blog, June 2016.

[101] Z. Li, Y. Zhou, H. Wei, C. Ge, and J. Jiang, "Towards Extreme Image Compression with Latent Feature Guidance and Diffusion Prior", *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[102] H. Lin, B. Chen, Z. Zhang, J. Lin, X. Wang, and T. Zhao, "DeepSVC: Deep Scalable Video Coding for Both Machine and Human Vision", in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, 9205–14.

[103] K. Lin, C. Jia, Z. Zhao, L. Wang, S. Wang, S. Ma, and W. Gao, "Residual in Residual Based Convolutional Neural Network In-loop Filter for AVS3", in *Picture Coding Symposium*, 2019, 1–5.

[104] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling", *arXiv preprint arXiv:2210.02747*, 2022.

[105] B. Liu, Y. Chen, S. Liu, and H.-S. Kim, "Deep Learning in Latent Space for Video Prediction and Compression", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 701–10.

[106] G. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, D. I. Kim, and X. Shen, "Semantic communications for artificial intelligence generated content (AIGC) toward effective content creation", *IEEE Network*, 2024.

[107] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 14388–97.

[108] M. Liu, C. Xu, Y. Gu, C. Yao, and Y. Zhao, "I^{2}VC: A Unified Framework for Intra-& Inter-frame Video Compression", *arXiv preprint arXiv:2405.14336*, 2024.

[109] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection", in *European Conference on Computer Vision*, Springer, 2024, 38–55.

[110] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow", *arXiv preprint arXiv:2209.03003*, 2022.

[111] S. Lombardo, J. Han, C. Schroers, and S. Mandt, "Deep generative video compression", *Advances in Neural Information Processing Systems*, 32, 2019.

[112] G. Lu, C. Cai, X. Zhang, L. Chen, W. Ouyang, D. Xu, and Z. Gao, "Content adaptive and error propagation aware deep video compression", in *European Conference on Computer Vision*, 2020, 456–72.

[113] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 11006–15.

[114] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation", in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 10209–18.

[115] C. Ma, D. Liu, X. Peng, L. Li, and F. Wu, "Convolutional neural network-based arithmetic coding for HEVC intra-predicted residues", *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7), 2019, 1901–16.

[116] H. Ma, D. Liu, N. Yan, H. Li, and F. Wu, "End-to-end optimized versatile image compression with wavelet-like transform", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[117] W. Ma and Z. Chen, "Diffusion-based perceptual neural video compression with temporal diffusion information reuse", *arXiv preprint arXiv:2501.13528*, 2025.

[118] Y. Ma, W. Yang, and J. Liu, "Correcting Diffusion-Based Perceptual Image Compression with Privileged End-to-End Decoder", in *International Conference on Machine Learning*, PMLR, 2024, 34075–93.

[119] Q. Mao, T. Yang, Y. Zhang, Z. Wang, M. Wang, S. Wang, L. Jin, and S. Ma, "Extreme image compression using fine-tuned vqgans", in *2024 Data Compression Conference (DCC)*, IEEE, 2024, 203–12.

[120] X. Meng, C. Jia, X. Zhang, S. Wang, and S. Ma, "Deformable Wiener Filter for Future Video Coding", *IEEE Transactions on Image Processing*, 31, 2022, 7222–36.

[121] F. Mentzer, E. Agustsson, J. Ballé, D. Minnen, N. Johnston, and G. Toderici, "Neural Video Compression using GANs for Detail Synthesis and Propagation", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, 416–31.

[122] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression", *Advances in Neural Information Processing Systems*, 33, 2020, 11913–24.

[123] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression", *Advances in Neural Information Processing Systems*, 31, 2018.

[124] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression", in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, 3339–43.

[125] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a Completely Blind Image Quality Analyzer", *IEEE Signal Processing Letters*, 20(3), 2013, 209–12.

[126] M. J. Muckley, A. El-Nouby, K. Ullrich, H. Jégou, and J. Verbeek, "Improving statistical fidelity for neural image compression with implicit local likelihood models", in *International Conference on Machine Learning*, PMLR, 2023, 25426–43.

[127] A. El-Nouby, M. J. Muckley, K. Ullrich, I. Laptev, J. Verbeek, and H. Jégou, "Image compression with product quantized masked image modeling", *arXiv preprint arXiv:2212.07372*, 2022.

[128] M. Oquab, P. Stock, D. Haziza, T. Xu, P. Zhang, O. Celebi, Y. Hasson, P. Labatut, B. Bose-Kolanu, T. Peyronel, *et al.*, "Low bandwidth video-chat compression using deep generative models", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, 2388–97.

[129] G. Parmar, R. Zhang, and J.-Y. Zhu, "On aliased resizing and surprising subtleties in gan evaluation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 11410–20.

[130] Y. Qian, M. Lin, X. Sun, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression", *arXiv preprint arXiv:2202.05492*, 2022.

[131] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 1505–14.

[132] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision", in *International Conference on Machine Learning*, PMLR, 2021, 8748–63.

[133] S. K. Raman, A. Ramesh, V. Naganoor, S. Dash, G. Kumaravelu, and H. Lee, "CompressNet: Generative Compression at Extremely Low Bitrates", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, 2325–33.

[134] L. Relic, R. Azevedo, M. Gross, and C. Schroers, "Lossy image compression with foundation diffusion models", in *European Conference on Computer Vision*, Springer, 2024, 303–19.

[135] O. Rippel and L. Bourdev, "Real-time adaptive image compression", in *International Conference on Machine Learning*, PMLR, 2017, 2922–30.

[136] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, "Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 10219–28.

[137] S. Santurkar, D. Budden, and N. Shavit, "Generative compression", in *Picture Coding Symposium (PCS)*, IEEE, 2018, 258–62.

[138] H. Schwarz, M. Coban, M. Karczewicz, T.-D. Chuang, F. Bossen, A. Alshin, J. Lainema, C. R. Helmrich, and T. Wiegand, "Quantization and Entropy Coding in the Versatile Video Coding (VVC) Standard", *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10), 2021, 3891–906.

[139] S. Shang, Z. Shan, G. Liu, L. Wang, X. Wang, Z. Zhang, and J. Zhang, "Resdiff: Combining cnn and diffusion model for image super-resolution", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 8, 2024, 8975–83.

[140] C. E. Shannon, "A mathematical theory of communication", *The Bell system technical journal*, 27(3), 1948, 379–423.

[141] H. R. Sheikh and A. C. Bovik, "Image information and visual quality", *IEEE Transactions on image processing*, 15(2), 2006, 430–44.

[142] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation", *Advances in Neural Information Processing Systems*, 32, 2019.

[143] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion Representations for Articulated Animation", in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, 13648–57.

[144] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics", in *International Conference on Machine Learning*, pmlr, 2015, 2256–65.

[145] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models", in *International Conference on Learning Representations*, 2020.

[146] J. Song, L. Yang, and M. Feng, "Taming Large Multimodal Agents for Ultra-low Bitrate Semantically Disentangled Image Compression", *arXiv preprint arXiv:2503.00399*, 2025.

[147] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution", *Advances in Neural Information Processing Systems*, 32, 2019.

[148] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations", *arXiv preprint arXiv:2011.13456*, 2020.

[149] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard", *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12), 2012, 1649–68.

[150] Z. Tang, H. Wang, X. Yi, Y. Zhang, S. Kwong, and C.-C. J. Kuo, "Joint graph attention and asymmetric convolutional neural network for deep image compression", *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1), 2022, 421–33.

[151] Z. Tang, X. Yi, and H. Wang, "Toward Learned Image Compression for Multiple Semantic Analysis Tasks", *IEEE Transactions on Broadcasting*, 2025.

[152] L. Theis, T. Salimans, M. D. Hoffman, and F. Mentzer, "Lossy compression with gaussian diffusion", *arXiv preprint arXiv:2206.08889*, 2022.

[153] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders", en-US, *arXiv preprint arXiv:1703.00395*, March, March 2017, DOI: 10.17863/cam.51995.

[154] T. Tian, H. Wang, L. Zuo, C.-C. J. Kuo, and S. Kwong, "Just noticeable difference level prediction for perceptual image compression", *IEEE Transactions on Broadcasting*, 66(3), 2020, 690–700.

[155] G. Toderici, W. Shi, R. Timofte, L. Theis, J. Ballé, E. Agustsson, N. Johnston, and F. Mentzer, "Workshop and challenge on learned image compression (clic2020)", in *CVPR*, 2020.

[156] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein Auto-Encoders", in *International Conference on Learning Representations*, 2018.

[157] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models", *arXiv preprint arXiv:2307.09288*, 2023.

[158] M. Tschannen, E. Agustsson, and M. Lucic, "Deep generative models for distribution-preserving lossy compression", *Advances in Neural Information Processing Systems*, 31, 2018.

[159] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, *et al.*, "Conditional image generation with pixelcnn decoders", *Advances in Neural Information Processing Systems*, 29, 2016.

[160] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning", *Advances in Neural Information Processing Systems*, 30, 2017.

[161] G.-H. Wang, J. Li, B. Li, and Y. Lu, "EVC: Towards Real-Time Neural Image Compression with Mask Decay", in *International Conference on Learning Representations*, 2023.

[162] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, 10039–49.

[163] Y. Wang, X. Fan, C. Jia, D. Zhao, and W. Gao, "Neural Network Based Inter Prediction for HEVC", in *IEEE International Conference on Multimedia and Expo*, IEEE Computer Society, 2018, 1–6.

[164] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", *IEEE Transactions on Image Processing*, 13(4), 2004, 600–12.

[165] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment", in *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2, Ieee, 2003, 1398–402.

[166] S. Wei, T. Ye, S. Zhang, Y. Tang, and J. Liang, "Joint token pruning and squeezing towards more aggressive compression of vision transformers", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, 2092–101.

[167] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard", *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7), 2003, 560–76.

[168] J. Wu, W. Gan, Z. Chen, S. Wan, and H. Lin, "Ai-generated content (aigc): A survey", *arXiv preprint arXiv:2304.06632*, 2023.

[169] L. Wu, K. Huang, and H. Shen, "A gan-based tunable image compression system", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, 2334–42.

[170] X. Wu, H. Wang, S. Hu, S. Kwong, and C.-C. J. Kuo, "Perceptually weighted mean squared error based rate-distortion optimization for HEVC", *IEEE Transactions on Broadcasting*, 66(4), 2020, 824–34.

[171] N. Xue, Q. Mao, Z. Wang, Y. Zhang, and S. Ma, "Unifying generation and compression: Ultra-low bitrate image coding via multi-stage transformer", in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2024, 1–6.

[172] H. Yamamoto, "Rate-distortion theory for the Shannon cipher system", *IEEE Transactions on Information Theory*, 43(3), 1997, 827–35.

[173] Z. Yan, F. Wen, and P. Liu, "Optimally Controllable Perceptual Lossy Compression", in *International Conference on Machine Learning*, PMLR, 2022, 24911–28.

[174] Z. Yan, F. Wen, R. Ying, C. Ma, and P. Liu, "On perceptual lossy compression: The cost of perceptual reconstruction and an optimal training framework", in *International Conference on Machine Learning*, PMLR, 2021, 11682–92.

[175] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with recurrent auto-encoder and recurrent probability model", *IEEE Journal of Selected Topics in Signal Processing*, 15(2), 2020, 388–401.

[176] R. Yang, R. Timofte, and L. Van Gool, "Perceptual Learned Video Compression with Recurrent Conditional GAN.", in *IJCAI*, 2022, 1537–44.

[177] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models", *Advances in Neural Information Processing Systems*, 36, 2023, 64971–95.

[178] R. Yang, Y. Yang, J. Marino, and S. Mandt, "Hierarchical autoregressive modeling for neural video compression", *arXiv preprint arXiv:2010.10258*, 2020.

[179] Y. Yang, S. Mandt, L. Theis, *et al.*, "An introduction to neural data compression", *Foundations and Trends ő in Computer Graphics and Vision*, 15(2), 2023, 113–200.

[180] X. Yi, H. Wang, S. Kwong, and C.-C. J. Kuo, "Task-driven video compression for humans and machines: Framework design and optimization", *IEEE Transactions on Multimedia*, 25, 2022, 8091–102.

[181] K. Zhang, Y. Zhou, X. Xu, B. Dai, and X. Pan, "DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, 7912–21.

[182] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models", in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, 3836–47.

[183] P. Zhang, J. Li, K. Chen, M. Wang, L. Xu, H. Li, N. Sebe, S. Kwong, and S. Wang, "When video coding meets multimodal large language models: A unified paradigm for video coding", *arXiv preprint arXiv:2408.08093*, 2024.

[184] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 586–95.

[185] S. Zhang, M. Mrak, L. Herranz, M. G. Blanch, S. Wan, and F. Yang, "Dvc-p: Deep video compression with perceptual optimizations", in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, IEEE, 2021, 1–5.

[186] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-Based Style Transfer With Diffusion Models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, 10146–56.

[187] Z. Zhang, B. Chen, H. Lin, J. Lin, X. Wang, and T. Zhao, "ELFIC: A Learning-based Flexible Image Codec with Rate-Distortion-Complexity Optimization", in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, 9252–61.

[188] L. Zhao, H. Bai, A. Wang, and Y. Zhao, "Learning a virtual codec based on deep convolutional neural network to compress image", *Journal of Visual Communication and Image Representation*, 63, 2019, 102589.

[189] T. Zhao, W. Feng, H. Zeng, Y. Xu, Y. Niu, and J. Liu, "Learning-Based Video Coding with Joint Deep Compression and Enhancement", in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[190] Z. Zhao, S. Wang, S. Wang, X. Zhang, S. Ma, and J. Yang, "Enhanced bi-prediction with convolutional neural network for high-efficiency video coding", *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11), 2018, 3291–301.

[191] C. Zhu, J. Xu, D. Feng, R. Xie, and L. Song, "Edge-based video compression texture synthesis using generative adversarial network", *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10), 2022, 7061–76.

[192] L. Zhu, S. Kwong, Y. Zhang, S. Wang, and X. Wang, "Generative adversarial network-based intra prediction for video coding", *IEEE Transactions on Multimedia*, 22(1), 2019, 45–58.

[193] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding", in *International Conference on Learning Representations*, 2022.