ProGVC: Progressive-based Generative Video Compression via Auto-Regressive Context Modeling

Daowen Li[1†], Ruixiao Dong[1†], Kai Li[1*], Ying Chen[1*], Ding Ding[1], and Li Li[2]

[1] Alibaba Group, China
{lidaowen.ldw,dongruixiao.drx,kaishi.lk,YingChen,liangjie.dd}@taobao.com
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
lil1@ustc.edu.cn

Abstract. Perceptual video compression leverages generative priors to reconstruct realistic textures and motions at low bitrates. However, existing perceptual codecs often lack native support for variable bitrate and progressive delivery, and their generative modules are weakly coupled with entropy coding, limiting bitrate reduction. Inspired by the next-scale prediction in the Visual Auto-Regressive (VAR) models, we propose **ProGVC**, a **Pro**gressive-based **G**enerative **V**ideo **C**ompression framework that unifies progressive transmission, efficient entropy coding, and detail synthesis within a single codec. ProGVC encodes videos into hierarchical multi-scale residual token maps, enabling flexible rate adaptation by transmitting a coarse-to-fine subset of scales in a progressive manner. A Transformer-based multi-scale autoregressive context model estimates token probabilities, utilized both for efficient entropy coding of the transmitted tokens and for predicting truncated fine-scale tokens at the decoder to restore perceptual details. Extensive experiments demonstrate that as a new coding paradigm, ProGVC delivers promising perceptual compression performance at low bitrates while offering practical scalability at the same time.

Keywords: Perceptual video compression · Progressive coding · Visual Auto-Regressive model

1 Introduction

Driven by the explosive growth of multimedia consumption, advanced video compression technologies have become increasingly essential to reduce the bitrate cost of storing and transmitting video data.

Depending on the target application, video compression can be designed for human viewing or for machine-oriented visual analysis [7, 33, 44], with this work focusing on the former. Traditional video compression standards, such as

[†] Equal contribution.
[*] Corresponding author.

H.264/AVC [40], H.265/HEVC [31] and H.266/VVC [3], have achieved remarkable success by employing carefully designed modules optimized towards pixel-wise distortion metrics (PSNR, SSIM [39], MS-SSIM [38], *etc.*). In recent years, end-to-end Neural Video Compression (NVC) models, including [2,12,13,18,30] have also demonstrated competitive performance against traditional standards. While these codecs are effective at preserving signal fidelity, they often produce visually unpleasing reconstruction results at low bitrates. These annoying visual artifacts include blurring, blocking, and ringing, indicating that human perception diverges from such distortion measures.

To improve perceptual quality, researchers have explored generative video compression, leveraging powerful generative models to synthesize realistic details. GAN-based methods [14, 43, 46] adopted generative adversarial network (GAN) [9], which yielded sharp-looking results but often suffered from unstable adversarial training and low-fidelity reconstructions. Recently, the advent of Vision Foundation Models, particularly diffusion models [25, 26, 29], has inspired diffusion-based generative codecs [19–21] that exploit powerful priors to generate realistic textures and motions.

Despite these advances, existing perceptual codecs still face several limitations. First, most methods lack support for scalability and variable bitrate adaptation, which are crucial under fluctuating network bandwidth. Second, diffusion-based codecs typically use diffusion model as decoding-time denoiser for the transmitted pixels or latents. Since the diffusion process is largely decoupled from entropy coding, spatio-temporal priors of diffusion model are less utilized to remove signal redundancy. Moreover, diffusion model requires iterative denoising, which increases decoding latency and hinders real-time applications.

In this paper, we aim to address these issues by introducing Visual Auto-Regressive (VAR) [10,15,32] models for generative video compression. The next-scale prediction mechanism in VAR utilizes a multi-scale residual quantizer to encode a video sequence into hierarchical visual token maps (*i.e.*, scales). These scales are then predicted progressively utilizing a Transformer-based Auto-Regressive (AR) model. This coarse-to-fine paradigm naturally supports scalability and bitrate adaptability: the encoder can transmit the first few coarse scales and progressively refine quality by sending additional high-frequency details. Moreover, the autoregressive model provides accurate conditional probabilities for each token that can directly serve as an entropy context model. Furthermore, unlike diffusion models that rely on iterative denoising loops, VAR enables efficient sampling in much fewer sampling steps, leading to substantially lower decoding latency.

Motivated by these insights, we propose *ProGVC* (Progressive-based Generative Video Compression via Auto-Regressive Context Modeling), a perceptual video coding paradigm that unifies progressive scalability, efficient entropy modeling, and detail synthesis in a single autoregressive pipeline. Specifically, ProGVC employs a causal video VAE to extract continuous intra frame features for the first intra frame and inter frame features for subsequent frames. At encoder, a multi-scale residual quantization scheme encodes these features

into K hierarchical multi-scale discrete token maps. To achieve scalability and variable bitrate control, the encoder transmits all intra frame scales, as they are crucial for leveraging temporal priors and incur only a small bitrate overhead, and the first k inter scales ($k \leq K$), where the different choices of k correspond to different bitrates. Moreover, to maximize the compression efficiency for the transmitted scales, we introduce the multi-scale autoregressive context model that acts as the entropy context model. For intra frame scales, the model captures cross-scale spatial dependencies; for inter frame scales, it explicitly conditions on both previously decoded intra and inter frame scales to exploit global spatio-temporal redundancy. The obtained discrete distributions for each token then drive arithmetic coding to minimize bitrate. At the decoder, ProGVC decodes the transmitted scales losslessly and utilizes the autoregressive model to generate the discarded $K - k$ high-frequency inter scales, enabling high-quality perceptual reconstruction with minimal additional rate.

To the best of our knowledge, this is the *first* study to systematically investigate generative video coding built upon an visual autoregressive generative model. We hope this work can shed some light on future research in this field. In summary, our core contributions are as follows:

- We introduce ProGVC, the first progressive generative video coding paradigm built upon the visual autoregressive model, enabling native scalability and variable bitrate by transmitting different subsets of hierarchical scales.
- We design a task-specific multi-scale autoregressive context model for the intra and inter token coding, unifying accurate probability estimation for entropy coding of the transmitted tokens and generative sampling for the un-transmitted details within a single framework, thereby reducing bitrate while preserving reconstruction quality.
- Extensive experimental results demonstrate that ProGVC exhibits superior perceptual quality over fidelity-oriented approaches and achieves competitive compression performance against existing perceptual video codecs.

2 Related works

2.1 Perceptual Video Compression

To improve perceptual quality of the reconstructed videos, generative models have been utilized in perceptual video compression methods. Among them, GAN-based codecs adopt adversarial learning to produce more realistic reconstructions. Mentzer *et al*. [22] first introduced GAN into video compression. Yang *et al*. [43] designed recurrent conditional GAN and adversarial loss functions. Qi *et al*. [27] presented a latent generative coding scheme, performing transform coding with a vector-quantized variational auto-encoder.

Recently, diffusion-prior-based coding paradigms have emerged, where the diffusion foundation model is used as the denoiser in latent or pixel space to synthesize missing details and thus enhance the perceptual quality of the reconstructions. Li *et al*. [11] proposed a hybrid approach that combined image

compression and a diffusion model for video compression. Ma and Chen [19] integrated the Stable Diffusion model [29] as a latent denoiser into the conditional video codec. Mao *et al.* [21] leveraged a video diffusion transformer model to jointly enhance intra and inter frame latents through sequence-level denoising.

Despite their encouraging visual results, prior perceptual video codecs have limited practicality, mainly due to the lack of native support for variable bitrate and scalable bitstream, resulting in the fact that they mostly require training separate models for different rates. In addition, diffusion models are typically leveraged as the decoder-side refiners to suppress quantization noise, rather than being tightly coupled with entropy modeling. Consequently, their powerful spatio-temporal priors are not directly translated into reducing transmitted bits. Furthermore, diffusion-based decoding typically relies on multiple iterative denoising steps, which increases computational cost and limits their applicability in low-delay or real-time streaming scenarios.

2.2 Visual Auto-Regressive Generative Models

Unlike diffusion models which rely on iterative denoising, Auto-Regressive (AR) models generate images and videos by sequentially predicting visual tokens conditioned on prefix context. Early raster-scan AR methods operate directly in pixel domain [24, 34]. To improve efficiency and generation quality, Taming Transformers [8] and VideoGPT [42] proposed a two-stage framework: first leverages a discrete tokenizer mapping visual data to latent tokens via a learned codebook, which is conceptually analogous to quantization in visual compression, and then employs the Transformer that models token distributions autoregressively, resembling entropy modeling in compression. Nevertheless, traditional AR models suffer from structural degradation caused by the fixed raster-scan order, which limits their generative fidelity and practical applicability.

Visual Auto-Regressive (VAR) model [32] alleviates this issue with a hierarchical coarse-to-fine strategy based on a multi-scale tokenizer, progressively refining visual details via next-scale prediction. Its successors, Infinity [10] and InfinityStar [15], extend this paradigm to text-to-image and text-to-video generation, achieving performance comparable to or surpassing leading diffusion-based approaches. Recently, VAR has been adapted to downstream tasks such as controllable generation [6,16], image super-resolution [28] and image compression [47]. Despite these advances, the potential of visual autoregressive models for video compression remains unexplored. Notably, the hierarchical next-scale prediction mechanism of VAR naturally lends itself to a progressive and scalable perceptual video compression framework. Motivated by this, we adopt the visual autoregressive model as the backbone of our perceptual video compression framework.

3 Method

This section first describes the overall framework (Sec. 3.1), highlighting on how scalable compression and variable bitrate adaptation are implemented within

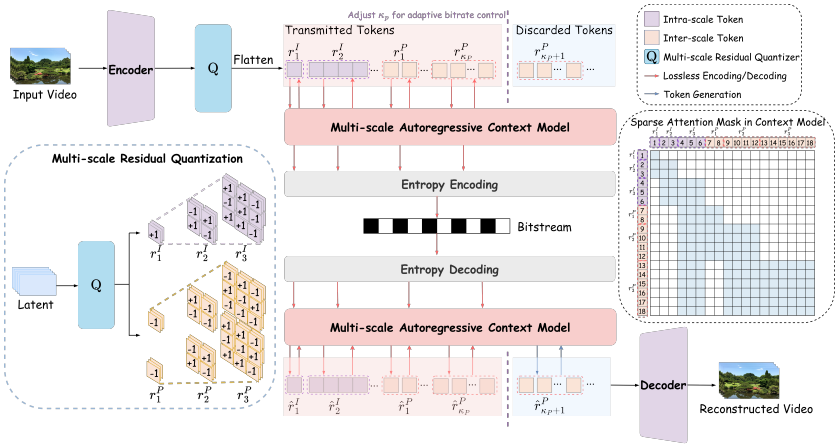

Fig. 1: Overview of the proposed ProGVC video compression framework. We design a progressive generative video compression framework based on the visual autoregressive model. The left panel illustrates the output pyramid structure of the multi-scale residual quantization. Besides, the sparse attention mask used in the context model is shown on the right.

the visual autoregressive model. We then elaborate on the Multi-scale Residual Quantization module (Sec. 3.2) for multi-scale token maps generation, followed by the Multi-scale Auto-Regressive Context Model (Sec. 3.3) for efficient token transmission.

3.1 Overall Framework

We design ProGVC, a progressive perceptual video compression framework based on the visual autoregressive model, as illustrated in Fig. 1. Specifically, a causal video VAE encoder first maps the input video clip into a continuous latent. A multi-scale residual quantizer then discretizes this latent into multi-scale token maps. During encoding, the intra scales and the first k low-frequency inter token maps are losslessly encoded considering their importance in temporal consistency and global layout construction, while higher-frequency inter scales are discarded and later generated at the decoder. We can adjust k for adaptive bitrate control. A multi-scale autoregressive context model estimates the distribution of each token, which is used both for entropy coding of the transmitted tokens and for generation of the discarded scales. Finally, the VAE decoder reconstructs the video from the recovered multi-scale token maps.

Continuous Feature Extraction. For an input video clip $V \in \mathbb{R}^{(1+T)\times W\times H\times 3}$, we regard the first frame as the intra frame, and the remaining frames are inter frames. A causal VAE encoder $\mathcal{E}$ encodes V into continuous spatio-temporal latent features by $f = \mathcal{E}(V) = \{f_t\}_{t=1}^{1+\frac{T}{4}}$, with $f_t \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 64}$. We denote the

intra latent features corresponding to the intra frame as $f^I = f_1$, and the inter latent features for inter frames as $f^P = \{f_t\}_{t=2}^{1+\frac{T}{4}}$.

Multi-scale Residual Quantization. The employed multi-scale residual quantization converts the continuous latent features into multi-scale residual token maps:

$$R^{(\cdot)} = \mathrm{Q}(f^{(\cdot)}) = \{r_k^{(\cdot)}\}_{k=1}^{K}, \quad (\cdot) \in \{I, P\}. \tag{1}$$

Here, Q denotes the multi-scale residual quantization, and R^I and R^P are the resulting multi-scale intra and inter token maps (*i.e.*, scales) respectively. The spatial resolution of the k-th scale $r_k^{(\cdot)}$ increases with k, with higher scales progressively capturing more high-frequency details. Following Binary Spherical Quantization (BSQ) [48], each token in $r_k^{(\cdot)}$ is a binary code of length L_k.

As k increases, the upsampling and element-wise addition result of $\{r_i^{(\cdot)}\}_{i=1}^{k}$ progressively approximates the original continuous features $f^{(\cdot)}$. The detailed procedure to obtain multi-scale residual token maps is presented in Sec. 3.2.

Scalability and Adaptive Bitrate Control. By transmitting multi-scale token maps from coarse to fine, ProGVC first reconstructs global structure and layout from the base scales, and then progressively adds localized, high-frequency details from the later scales. This progressive structure provides explicit scalability: bitstreams containing more scales yield higher reconstruction quality, while truncating the tail scales produces a valid, lower-rate reconstruction. Selecting different subsets of scales for transmission also enables adaptive bitrate control without retraining or modifying the compression model. In practice, we primarily adjust the bitrate by truncating inter scale tokens, since they constitute the majority of transmitted tokens and considering that the intra scales are vital for temporal alignment in modeling inter scales distribution. The discarded higher-frequency inter scales are generated conditioned on the already reconstructed scales at the decoder side. Then the multi-scale summation over all scales is applied to reconstruct the continuous features for final video decoding.

Entropy Coding based on Context Modeling. For the transmitted scales, ProGVC further improves compression efficiency through *lossless* entropy coding. Since each token is represented by binary codes, the resulting bit sequence naturally supports arithmetic coding, which reduces the bitrate given accurate probability estimation. Here we design a multi-scale autoregressive context model that exploits the spatio-temporal priors among scales to estimate the distribution of each scale conditioned on previous scales. Modeling the distributions of intra and inter scales is discussed separately as follows.

Intra context modeling. Following the autoregressive factorization across scales in VAR, the joint distribution of intra scales is factorized as

$$p(r_1^I, \ldots, r_K^I) = \prod_{k=1}^{K} p\big(r_k^I \mid \{r_{k'}^I\}_{k'<k}\big), \tag{2}$$

where r_k^I denotes the k-th intra scale. Given the conditional probability $p(r_k^I \mid \{r_{k'}^I\}_{k'<k})$, the bitstream of the k-th intra scale, denoted as $\mathcal{B}_k^I$, is generated via entropy coding.

Inter context modeling. For inter scales, only the first κ_P scales ($\kappa_P \leq K$) are transmitted. Let r_k^P denotes the k-th inter scale, and $R^I = \{r_k^I\}_{k=1}^K$ represents the intra scales. The joint distribution of inter scales is factorized as

$$p(r_1^P, \ldots, r_{\kappa_P}^P) = \prod_{k=1}^{\kappa_P} p\big(r_k^P \mid \{r_{k'}^P\}_{k'<k},\, R^I\big)\,. \tag{3}$$

That is, each inter scale is conditioned on all previously reconstructed inter scales as well as the intra scales. Analogously, the bitstream of the k-th inter scale tokens, $\mathcal{B}_k^P$, is generated according to the corresponding conditional distribution.

The detailed architecture of the multi-scale autoregressive context model is described in Sec. 3.3.

Entropy Decoding and Token Generation for Video Decompression. At the decoder, the transmitted intra and inter scale tokens are first losslessly recovered from the bitstreams via entropy decoding, denoted as $\{\hat{r}_k^I\}_{k=1}^K$ and $\{\hat{r}_k^P\}_{k=1}^{\kappa_P}$, using the same autoregressive probability factorization as in encoding.

The discarded higher-frequency inter scales are then synthesized through token generation. Specifically, the discarded inter scales ($k > \kappa_P$) are generated conditioned on the already reconstructed scales with the same conditional distribution modeling mechanism, while selecting the most probable entries as

$$\hat{r}_k^P = \arg\max_{\hat{r}_k^P}\; p\Big(\hat{r}_k^P \mid \{\hat{r}_{k'}^P\}_{k'<k},\, \hat{R}^I\Big)\,, \quad k = \kappa_P{+}1, \ldots, K. \tag{4}$$

After obtaining the complete set of hierarchical scales, the intra and inter latent features $\hat{f}^I$ and $\hat{f}^P$ are reconstructed via multi-scale summation. Finally, the reconstructed video $\hat{V}$ is obtained by feeding these latent features to the VAE decoder $\mathcal{D}$ as $\hat{V} = \mathcal{D}(\hat{f}^I, \hat{f}^P)$.

3.2 Multi-scale Residual Quantization

To discretize the continuous latent features $f^{(\cdot)}$, we adopt a multi-scale residual quantization scheme that produces K hierarchical scales $\{r_k^{(\cdot)}\}_{k=1}^K$.

Specifically, at scale k, the residual feature $e_k^{(\cdot)}$, which is initialized with the continous feature $f^{(\cdot)}$ at $k = 1$, is first downsampled into token vectors $\tilde{e}_k^{(\cdot)}$:

$$\tilde{e}_k^{(\cdot)} = D_k\Big(e_k^{(\cdot)}\Big)\,, \tilde{e}_k^{(\cdot)} \in \mathbb{R}^{w_k \times h_k \times T' \times L_k}, e_k^{(\cdot)} \in \mathbb{R}^{W/4 \times H/4 \times T' \times L_k}, \tag{5}$$

where D_k denotes the downsampling operator at scale k, (w_k, h_k) represents the spatial resolution of k-th token map, T' and L_k denote the temporal and channel dimensions, respectively. Each token vector $v \in \mathbb{R}^{L_k}$ from $\tilde{e}_k^{(\cdot)}$ is discretized using Binary Spherical Quantization (BSQ) [48], which binarizes each element v_s of the token according to its sign and normalizes it by the token dimension:

$$\mathrm{BSQ}(v_s) = \frac{\mathrm{sign}(v_s)}{\sqrt{L_k}}, \qquad \mathrm{sign}(v_s) \in \{-1, +1\}. \tag{6}$$

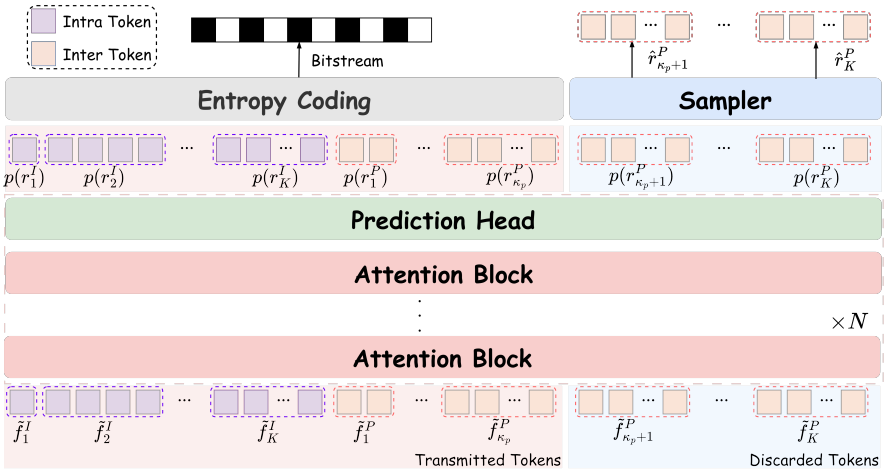

Fig. 2: Overview of the multi-scale autoregressive context model. The transformer network predicts both intra scale and inter scale tokens in an autoregressive manner. The resulting probability distributions are subsequently utilized for entropy coding and for sampling discarded tokens during reconstruction.

By applying BSQ to all k-th scale tokens, the module obtains discrete residual token map $r_k^{(\cdot)}$. $r_k^{(\cdot)}$ is then upsampled back to the latent resolution using the scale-specific upsampling operator U_k, and further subtracted from the current residual feature to generate the residual feature $e_{k+1}^{(\cdot)}$ for the next scale:

$$e_{k+1}^{(\cdot)} = e_k^{(\cdot)} - U_k\left(r_k^{(\cdot)}\right), \qquad k = 1, \ldots, K. \tag{7}$$

Through this hierarchical residual quantization process, the latent representation is progressively refined from coarse to fine scales, natively and explicitly supporting a scalable progressive generative video coding paradigm.

3.3 Multi-scale Auto-Regressive Context Modeling

In this stage, our transformer-based autoregressive context model estimates the distribution of each scale in an autoregressive manner. Fig. 2 exhibits the workflow. Specifically, to estimate the probability of the k-th scale $r_k^{(\cdot)}$, we first construct an aggregated multi-scale representation $\tilde{f}_k^{(\cdot)}$ that serves as input to the autoregressive model for the current scale,

$$\tilde{f}_k^{(\cdot)} = D_k\left(\sum_{i=1}^{k} U_i(r_i^{(\cdot)})\right). \tag{8}$$

The distributions of current scale are then predicted by injecting information from the prefix scales through masked self-attention. In particular, for each attention block, the temporal-spatial priors are incorporated as:

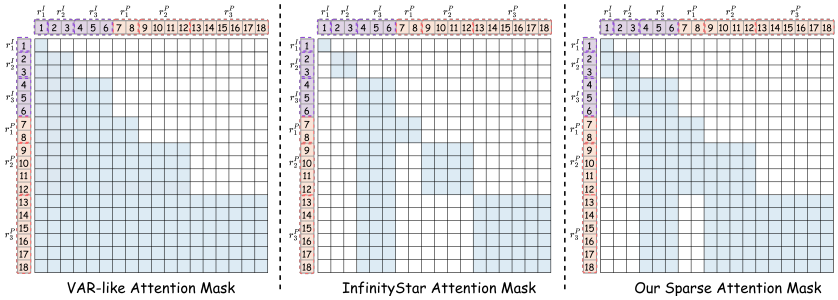

Fig. 3: Illustration of three attention mask designs: VAR-like attention (left), InfinityStar attention (middle), and our sparse attention (right). For visualization simplicity, we illustrate the masks using three spatial scales (1, 2, 3) and a temporal length of $T = 2$ for inter scale tokens.

$$\begin{aligned} \mathcal{Q} = Z_k^{(\cdot)} W^q,\ \mathcal{K} = \text{concat}\left(Z_k^{(\cdot)}, C_k\right) W^k,\ \mathcal{V} = \text{concat}\left(Z_k^{(\cdot)}, C_k\right) W^v, \\ Z_k^{(\cdot)'} = \text{Attention}\left(\mathcal{Q}, \mathcal{K}, \mathcal{V}, \text{Mask}\right) = \text{Softmax}\left(\text{Mask}\left(\mathcal{Q}\mathcal{K}^\top / \sqrt{d}\right)\right)\mathcal{V}, \end{aligned} \tag{9}$$

where W^q, W^k, W^v denote the query, key, and value projection matrices; $Z_k^{(\cdot)}$ and $Z_k^{(\cdot)'}$ are the input and output of this attention block corresponding to k-th scale input $\tilde{f}_k^{(\cdot)}$, respectively. The context set C_k depends on the token type. For intra scales, C_k consists of previous intra scale features $\{Z_{k'}^I\}_{k'<k}$; and for inter scales, the context further includes prefix inter scale features $\{Z_{k'}^P\}_{k'<k}$ together with intra scale features $\{Z_{k'}^I\}_{k'<K}$. The attention mask Mask determines which tokens in $Z_k^{(\cdot)}$ are visible during attention, restricting each token to attend only to its effective prefix context. Multiple masked attention blocks are stacked, followed by an MLP-based prediction head to estimate the discrete token distribution. For transmitted scales, predicted probabilities are used for lossless entropy coding, whereas truncated scales are generated by selecting the most probable discrete entries according to the estimated distribution.

Sparse attention. Theoretically, incorporating richer conditional information can reduce conditional entropy [1], thereby lowering the expected bitrate. However, naively expanding the conditioning context significantly increases the computational cost of autoregressive video modeling, as the context grows rapidly along both spatial and temporal dimensions. Moreover, the aggregated input $\tilde{f}_k^{(\cdot)}$ at scale k already summarizes information from coarser scales, making full attention over all historical tokens partially redundant [35].

As shown in Fig. 3, Infinity [10] employs a block-wise causal mask that allows each scale to attend to all historical tokens, which becomes computationally expensive for long sequences. InfinityStar [15] reduces complexity by adopting a more restrictive attention mask (middle of Fig. 3); however, the limited receptive field weakens context utilization and may increase the bitrate.

To balance compression efficiency and computational complexity, we design a sparse attention mask used in both the encoding and decoding process, as shown in the right of Fig. 3. First, this attention mask allows each scale to attend to itself with its immediately preceding scale. By stacking multiple attention layers, long-range dependencies are progressively captured without explicitly attending to the full history. Furthermore, to leverage temporal priors and maintain temporal consistency in reconstructions, inter scale tokens are allowed to attend to the largest intra scale, which retains rich fine-grained details and exhibits stronger temporal coherence with inter scales. As shown in Sec. 4.3, this sparse attention design achieves a favorable trade-off between compression efficiency and computational overhead.

4 Experiments

4.1 Setup

Dataset. We train ProGVC on the Pexels dataset. To construct a high-quality training corpus, we compute the aesthetic and clarity scores for each video and retain only samples with scores above 4.5 and 0.65, respectively. Videos shorter than 10 seconds are discarded, and only those with aspect ratios close to standard 720p resolution are kept, resulting in a large-scale, high-quality corpus of approximately 480K videos. Evaluation is conducted on widely used benchmarks, including Xiph [41], HEVC Class B [31], and MCL-JCV [36]. Since ProGVC is trained primarily on 720p content following InfinityStar, the HEVC Class B and MCL-JCV datasets are downsampled to 720p.

Training Details. The base configurations of the causal VAE and the multi-scale residual quantization are inherited from InfinityStar. We finetune the VAE jointly with the quantization module, largely following prior work [15, 37, 48]. Specifically, training is conducted on $256 \times 256 \times 81$ video clips for 10K iterations with a batch size of 2. The training objective combines entropy loss, commitment loss, l_2 reconstruction loss, GAN loss, and LPIPS perceptual loss, weighted by 0.1, 0.25, 1, 0.01 and 4, respectively. The learning rate is set to 5×10^{-5}.

We train the multi-scale autoregressive context model for 60k iterations, utilizing the AdamW optimizer [17] with batch size of 8. The learning rate is set to 5×10^{-5}, with momentum coefficients as $(\beta_1, \beta_2) = (0.9, 0.97)$. Following the InfinityStar training protocol, we randomly drop the later inter scales to improve training efficiency. All videos are resized and randomly cropped to a resolution of 720p. Each training sample consists of 81 frames.

Metrics. To evaluate perceptual quality, full-reference quality metrics are adopted, including DISTS [5] and LPIPS [45], along with no-reference quality metric of NIQE [23]. PSNR is used to evaluate signal fidelity. And t-LPIPS [4] is adopted to measure temporal consistency. We use kilobits per second (kbps) to measure the bitrate cost for compressing a video sequence within one second.

Baselines. We compared ProGVC against three categories of video compression methods: a) traditional video codec based on the latest VVC standard, *i.e.*,

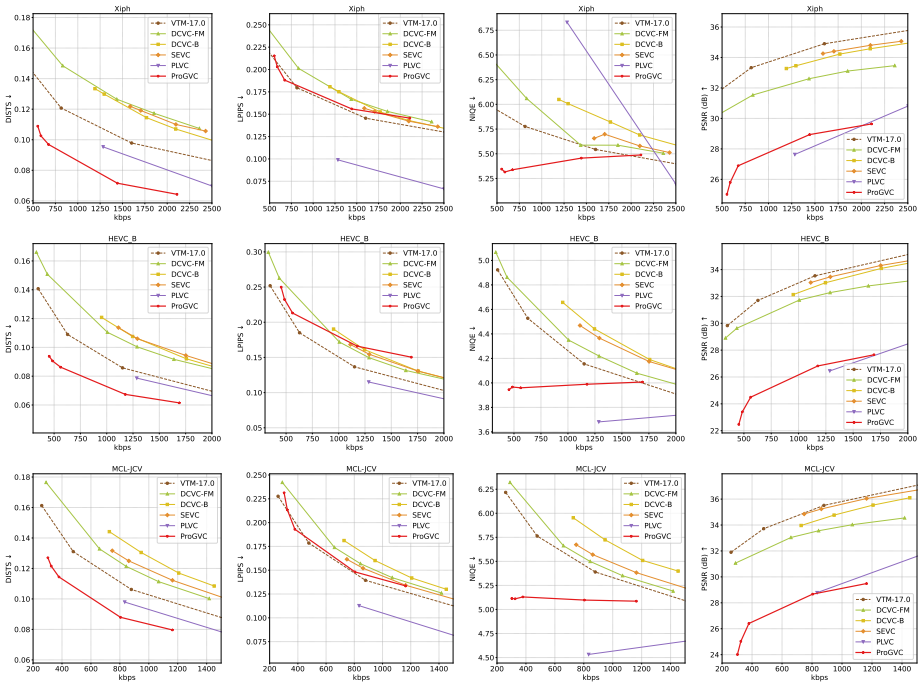

Fig. 4: Rate and perception/fidelity curves on Xiph, HEVC B and MCL-JCV datasets.

VTM-17.0 [3]; b) leading fidelity-oriented neural video codecs, including DCVC-FM [13] , DCVC-B [30] and SEVC [2]; c) perceptual video codecs, represented by PLVC [43], which serves as the strong reproducible generative baseline. Among these baselines, the fidelity-oriented codecs support variable-rate control but not scalable bitstreams, whereas PLVC supports neither scalability nor adaptive rate control. We adopt the random access configuration with GOP size of 32 and one intra frame for VTM-17.0, while for the other baseline methods we use their default configurations. For each test sequence, we evaluate the first 81 frames. All methods are tested in the RGB domain with the same RGB mode to ensure a fair comparison. The evaluated bitrate range is set to 300-2000 kbps. Additional implementation details are provided in the supplementary material.

4.2 Comparison Results

Quantitative Results. As demonstrated in Fig. 4, ProGVC achieves the best performance on DISTS among all competing methods. It also outperforms traditional and fidelity-oriented codecs under NIQE, while delivering a comparable perceptual efficiency on LPIPS. Compared with the perceptual codec PLVC [43], it further improves NIQE-based compression efficiency on Xiph. Tab. 1 reports

Table 1: BD-rate↓ (%) / BD-metric↑ on the Xiph, HEVC B, and MCL-JCV datasets. **Red** and **Blue** indicate the best and the second-best performance in terms of BD-metric, respectively. "N/A" indicates that BD-rate cannot be calculated due to the lack of quality overlap.

Dataset	Method	DISTS	LPIPS	NIQE	PSNR(dB)
Xiph	VTM-17.0 [3]	0.0 / 0.0000	**0.0 / 0.0000**	**0.0 / 0.0000**	**0.0 / 0.0000**
	DCVC-FM [13]	92.0 / -0.0257	40.4 / -0.0228	71.3 / -0.2525	111.1 / -1.8623
	SEVC [2]	87.7 / -0.0179	19.4 / -0.0062	41.8 / -0.1331	51.5 / -0.9035
	DCVC-B [30]	63.2 / -0.0148	17.3 / -0.0065	94.7 / -0.2306	56.8 / -1.0048
	PLVC [43]	**-40.3 / 0.0109**	**-83.7 / 0.0565**	169.9 / -0.0902	825.9 / -5.5261
	ProGVC	**-61.5 / 0.0300**	3.5 / -0.0025	**-62.2 / 0.2670**	891.2 / -5.8437
HEVC-B	VTM-17.0 [3]	0.0 / 0.0000	**0.0 / 0.0000**	0.0 / 0.0000	**0.0 / 0.0000**
	DCVC-FM [13]	62.8 / -0.0231	41.0 / -0.0356	23.1 / -0.1653	50.6 / -1.2807
	SEVC [2]	88.1 / -0.0193	39.2 / -0.0189	60.1 / -0.2066	20.5 / -0.5422
	DCVC-B [30]	78.6 / -0.0184	41.5 / -0.0209	63.9 / -0.2300	28.7 / -0.7316
	PLVC [43]	**-18.1 / 0.0046**	**-30.5 / 0.0150**	**-56.0 / 0.1048**	705.5 / -6.1212
	ProGVC	**-47.7 / 0.0217**	30.3 / -0.0259	**-52.5 / 0.3509**	879.6 / -6.8492
MCL-JCV	VTM-17.0 [3]	0.0 / 0.0000	**0.0 / 0.0000**	0.0 / 0.0000	**0.0 / 0.0000**
	DCVC-FM [13]	49.2 / -0.0215	32.9 / -0.0241	28.6 / -0.2241	68.8 / -1.4598
	SEVC [2]	52.0 / -0.0164	20.1 / -0.0102	30.6 / -0.1601	11.8 / -0.2967
	DCVC-B [30]	77.5 / -0.0247	44.2 / -0.0224	68.4 / -0.3514	37.2 / -0.9451
	PLVC [43]	**-28.8 / 0.0108**	**-45.1 / 0.0322**	**N/A / 0.5959**	571.2 / -5.7383
	ProGVC	**-46.2 / 0.0238**	3.3 / -0.0016	**-65.5 / 0.5287**	893.8 / -6.6569

the BD-rate and BD-metric results using VTM-17.0 [3] as the anchor, confirming that ProGVC achieves superior or comparable perceptual quality to other strong baselines while supporting scalability and variable-rate control. Tab. 2 further reports t-LPIPS which explicitly evaluating temporal consistency, and ProGVC consistently outperforms all baselines. This is mainly attributable to our hierarchical next-scale autoregressive generation and temporal-spatial attention design, where each token attends to tokens from other frames at current and previous scales, effectively capturing cross-frame temporal dependencies.

We note that ProGVC performs more favorably under DISTS and NIQE than under LPIPS. This is likely because our autoregressive context model is trained to capture the natural video distribution, favoring structural coherence and natural textures over strict pixel- or feature-level correspondence. As a result, slight local deviations, as shown in Fig. 5, may be penalized more heavily by LPIPS. Taken together, the multi-metric and qualitative results suggest that ProGVC achieves competitive or superior perceptual quality.

Qualitative Results. Fig. 5 illustrates the visual results of different video compression methods. ProGVC produces more visually pleasant reconstructions at low bitrates. Compared to fidelity-oriented baselines, including VTM-17.0 [3] and SEVC [2], ProGVC produces finer textures and higher structural integrity, mitigating blurring artifacts and fragmented edges. Compared to the perceptual baseline PLVC [43], ProGVC produces more natural textures.

Complexity. We compare the encoding and decoding complexity of ProGVC with several neural video codecs, including the fidelity-oriented codecs DCVC-

Table 2: Temporal consistency comparison on Xiph. BD-rate↓ (%) / BD-metric↑ are reported with VTM-17.0 as the anchor.

Method	t-LPIPS
DCVC-FM [13]	46.14 / -0.0019
SEVC [2]	-15.89 / 0.0004
DCVC-B [30]	58.61 / -0.0011
PLVC [43]	709.87 / -0.0070
ProGVC	**-20.76 / 0.0007**

Table 3: Complexity comparison with other neural codecs.

Methods	Enc. T (s) ↓	Dec. T (s) ↓
DCVC-B [30]	1.28	1.09
SEVC [2]	1.04	**0.88**
DiffVC [19]	0.82	4.54
ProGVC	**0.56**	1.53

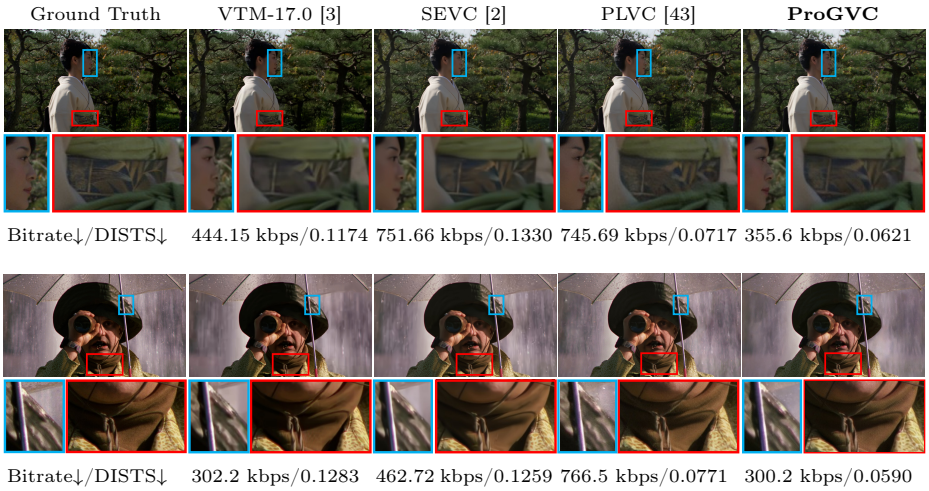

Fig. 5: Visual comparisons with baselines on the *Kimono* sequence in the Xiph dataset and the *videoSRC07* sequence in the MCL-JCV dataset.

B [30] and SEVC [2], as well as the diffusion-based perceptual codec DiffVC [19]. PLVC [43] is excluded due to its incompatible implementation with the unified evaluation platform. Since DiffVC is not open-sourced, we implement its architecture following the original paper without training. All experiments are conducted on the same hardware platform, with an NVIDIA H20 GPU and an Intel Xeon Platinum 8469C CPU (192 cores). For ProGVC, we report the average runtime over the lowest and highest bitrate settings. As summarized in Tab. 3, ProGVC achieves the lowest encoding complexity among all baselines, while its decoding time is comparable to DCVC-B [30] and SEVC [2], and clearly lower than that of DiffVC.

4.3 Ablation Studies

We conduct a series of ablation studies on the key components of our framework. For a fair comparison, we report BD-rate and BD-metric results on the Xiph dataset, evaluated across three metrics: DISTS, LPIPS and PSNR.

Table 4: Effect of multi-scale autoregressive context modeling. "uniform prob." denotes utilizing uniform probability distribution.

Model Variants	BD-rate↓(%) / BD-metric ↑		
	DISTS	LPIPS	PSNR
Base	**0.0 / 0.0000**	**0.0 / 0.0000**	**0.0 / 0.0000**
w/o context model (uniform prob.)	122.4 / -0.0259	119.6 / -0.0360	119.4 / -2.3437
w/o token generation	N/A / -0.1183	280.9 / -0.1273	189.9 / -3.8102

Table 5: Effect of three different temporal-aligned attention masks. The per-frame encoding and decoding times (Enc. T and Dec. T) are averaged over the lowest and highest bitrates.

Attn. Variants	BD-rate↓(%) / BD-metric↑			Enc. T (s) ↓	Dec. T (s) ↓
	DISTS	LPIPS	PSNR		
Ours	0.0 / 0.0000	0.0 / 0.0000	0.0 / 0.0000	0.56	1.53
Self-only	6.8 / -0.0021	12.8 / -0.0062	10.9 / -0.3482	0.50	1.23
Full causal	-6.5 / 0.0021	-3.8 / 0.0019	0.5 / -0.0184	0.64	1.96

Multi-scale Auto-Regressive Context Modeling. We evaluate the proposed context modeling procedure for both entropy coding and discarded token generation. To quantify its contribution to lossless entropy coding, we replace the proposed context model with a uniform probability distribution. As reported in Tab. 4, the proposed context modeling reduces the bitrate by more than 50% compared to the uniform baseline. Moreover, the proposed context modeling further improves both perceptual and fidelity metrics via token generation.

Attention mask design. We conduct ablation studies to evaluate the effectiveness of the proposed sparse attention mask. In particular, we investigate the trade-off between the bitrate reduction by exploiting sufficient contextual information from previous scales and maintaining computational efficiency.

We begin by analyzing temporally aligned intra–intra and inter–inter attention masks, where tokens attend to context from the same frame(s) at earlier scales. Specifically, we compare three variants: (1) *InfinityStar-like self-only attention*, where each scale attends only to itself; (2) *VAR-like full causal attention*, where each scale attends to all prefix scales; and (3) *our proposed design*, in which each scale attends only to itself together with its immediately preceding scale. As reported in Tab. 5, the proposed design achieves a favorable BD-rate–complexity trade-off. Compared with self-only attention, it consistently improves BD-rate across all three metrics. Moreover, it matches the compression performance of full causal attention while substantially reducing computational overhead.

To preserve temporal consistency, inter scales should incorporate information from intra scales. As shown in Tab. 6, we compare the proposed largest-intra-scale conditioning strategy with three alternatives: no intra scales conditioning, conditioning on the smallest intra scale, and conditioning on intra scales at the same spatial resolution. The results indicate that attending to the largest intra scale consistently improves the compression performance of inter frames. The largest intra scale, operating at the finest spatial resolution, provides

Table 6: Effect of different intra scales to the context modeling of inter scales. The per-frame encoding and decoding times (denoted as Enc. T and Dec. T) are averaged across the lowest and highest bitrate settings.

Attn. Variants	BD-rate↓(%) / BD-metric ↑			Enc. T (s) ↓	Dec. T (s) ↓
	DISTS	LPIPS	PSNR		
Ours (largest scale)	**0.0 / 0.0000**	**0.0 / 0.0000**	**0.0 / 0.0000**	0.56	1.53
No intra reference	37.3 / -0.0171	41.7 / -0.0254	46.6 / -1.6698	0.53	1.46
Smallest scale	42.7 / -0.0191	40.4 / -0.0252	46.8 / -1.6803	0.53	1.46
Same resolution scale	11.7 / -0.0042	12.2 / -0.0064	14.4 / -0.5131	0.54	1.49

high-frequency details that better match inter scales than low-frequency scales, thereby offering a stronger predictive prior. Meanwhile, the computational cost of the proposed strategy remains comparable across all variants. Based on these observations, the largest intra scale is adopted as one of the reference context for inter-scale token modeling.

5 Conclusion and Future Work

In this paper, we introduce ProGVC (Progressive-based Generative Video Compression via Auto-Regressive Context Modeling), the first perceptual video coding paradigm based on the visual autoregressive model that unifies progressive scalability, efficient entropy coding, and detail synthesis within a single framework. By integrating a multi-scale residual quantizer with a task-specific multi-scale autoregressive context model, ProGVC achieves lossless compression of low-frequency global structures while generating high-frequency details. Experimental results demonstrate that our model exhibits consistent improvements in perceptual quality over fidelity-oriented codecs and competitive performance compared to perceptual baselines, while maintaining native scalability, adaptive rate control, and low encoding-decoding latency.

Despite these advantages, ProGVC still leaves room for further optimization. First, the supported video resolution and duration are currently constrained by the training configuration of the autoregressive model. In particular, extending AR-based video generation beyond 720p remains challenging due to the substantially increased computational cost of training at higher resolutions, which mainly arises from the current maturity of the underlying AR video generative backbone. Second, the current adaptive rate control strategy is still coarse-grained, as bitrate is adjusted by transmitting or discarding entire scales, resulting in discrete rather than continuous rate points. Future work will scale ProGVC to higher resolutions as the underlying AR generative backbones continue to advance, and explore more general resolution- and length-agnostic architectures. In addition, we will investigate finer-grained rate control mechanisms, such as token-level dropping, to enable smoother bitrate adaptation in practical streaming scenarios.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(Feb), 1137–1155 (2003)
2. Bian, Y., Tang, C., Li, L., Liu, D.: Augmented deep contexts for spatially embedded video coding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2094–2104 (2025)
3. Bross, B., Wang, Y.K., Ye, Y., Liu, S., Chen, J., Sullivan, G.J., Ohm, J.R.: Overview of the versatile video coding (vvc) standard and its applications. TCSVT **31**(10), 3736–3764 (2021)
4. Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thuerey, N.: Learning temporal coherence via self-supervision for gan-based video generation. ACM Transactions on Graphics (TOG) **39**(4), 75–1 (2020)
5. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
6. Dong, R., Wang, Z., Liu, K., Li, L., Chen, Y., Li, K., Li, D., Li, H.: Echogen: Generating visual echoes in any scene via feed-forward subject-driven auto-regressive model. In: The Fourteenth International Conference on Learning Representations (2026), `https://openreview.net/forum?id=ctmyCjo18u`
7. Duan, L., Liu, J., Yang, W., Huang, T., Gao, W.: Video coding for machines: A paradigm of collaborative compression and intelligent analytics. IEEE Transactions on Image Processing **29**, 8680–8695 (2020)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
10. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15733–15744 (2025)
11. Li, B., Liu, Y., Niu, X., Bai, B., Deng, L., Gündüz, D.: Extreme video compression with pre-trained diffusion models. arXiv preprint arXiv:2402.08934 (2024)
12. Li, J., Li, B., Lu, Y.: Neural video compression with diverse contexts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22616–22626 (2023)
13. Li, J., Li, B., Lu, Y.: Neural video compression with feature modulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26099–26108 (2024)
14. Li, M., Shi, Y., Wang, J., Huang, Y.: High visual-fidelity learned video compression. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8057–8066 (2023)
15. Liu, J., Han, J., Yan, B., Zhu, F., Wang, X., Jiang, Y., PENG, B., Yuan, Z., et al.: Infinitystar: Unified spacetime autoregressive modeling for visual generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
16. Liu, K., Wang, Z., Zhou, W., Xu, S., Dong, R., Li, H.: Scaleweaver: Weaving efficient controllable t2i generation with multi-scale reference attention. arXiv preprint arXiv:2510.14882 (2025)

17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
18. Lu, G., Ouyang, W., Xu, D., Zhang, X., Cai, C., Gao, Z.: Dvc: An end-to-end deep video compression framework. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11006–11015 (2019)
19. Ma, W., Chen, Z.: Diffusion-based perceptual neural video compression with temporal diffusion information reuse. ACM Transactions on Multimedia Computing, Communications and Applications **21**(12), 1–22 (2025)
20. Ma, W., Chen, Z.: Diffvc-osd: One-step diffusion-based perceptual neural video compression framework. arXiv preprint arXiv:2508.07682 (2025)
21. Mao, Q., Cheng, H., Yang, T., Jin, L., Ma, S.: Generative neural video compression via video diffusion prior. arXiv preprint arXiv:2512.05016 (2025)
22. Mentzer, F., Agustsson, E., Ballé, J., Minnen, D., Johnston, N., Toderici, G.: Neural video compression using gans for detail synthesis and propagation. In: European Conference on Computer Vision. pp. 562–578. Springer (2022)
23. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
24. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. Advances in neural information processing systems **29** (2016)
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
27. Qi, L., Jia, Z., Li, J., Li, B., Li, H., Lu, Y.: Generative latent coding for ultra-low bitrate image and video compression. IEEE Transactions on Circuits and Systems for Video Technology (2025)
28. Qu, Y., Yuan, K., Hao, J., Zhao, K., Xie, Q., Sun, M., Zhou, C.: Visual autoregressive modeling for image super-resolution. In: Forty-second International Conference on Machine Learning
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Sheng, X., Li, L., Liu, D., Wang, S.: Bi-directional deep contextual video compression. IEEE Transactions on Multimedia (2025)
31. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on circuits and systems for video technology **22**(12), 1649–1668 (2012)
32. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
33. Tian, Y., Ling, X., Geng, C., Hu, Q., Lu, G., Zhai, G.: Smc++: Masked learning of unsupervised video semantic compression. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(2), 1992–2011 (2026)
34. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. pp. 1747–1756. PMLR (2016)
35. Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., Baranchuk, D.: Switti: Designing scale-wise transformers for text-to-image synthesis. arXiv preprint arXiv:2412.01819 (2024)

36. Wang, H., Gan, W., Hu, S., Lin, J.Y., Jin, L., Song, L., Wang, P., Katsavounidis, I., Aaron, A., Kuo, C.C.J.: Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In: 2016 IEEE international conference on image processing (ICIP). pp. 1509–1513. IEEE (2016)
37. Wang, J., Jiang, Y., Yuan, Z., Peng, B., Wu, Z., Jiang, Y.G.: Omnitokenizer: A joint image-video tokenizer for visual generation. Advances in Neural Information Processing Systems **37**, 28281–28295 (2024)
38. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402 Vol.2 (2003). https://doi.org/10.1109/ACSSC.2003.1292216
39. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
40. Wiegand, T., Sullivan, G.J., Bjøntegaard, G., Luthra, A.: Overview of the h. 264/avc video coding standard. IEEE Transactions on circuits and systems for video technology **13**(7), 560–576 (2003)
41. Xiph.org: Video test media. https://media.xiph.org/video/derf/, accessed: 2026
42. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
43. Yang, R., Timofte, R., Van Gool, L.: Perceptual learned video compression with recurrent conditional gan. In: IJCAI. pp. 1537–1544 (2022)
44. Yuan Tian, Guo Lu, G.Z.: Free-vsc: Free semantics from visual foundation models for unsupervised video semantic compression. In: European Conference on Computer Vision. pp. 163–183. Springer (2024)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
46. Zhang, S., Mrak, M., Herranz, L., Blanch, M.G., Wan, S., Yang, F.: Dvc-p: Deep video compression with perceptual optimizations. In: 2021 International Conference on Visual Communications and Image Processing (VCIP). pp. 1–5. IEEE (2021)
47. Zhang, Z., Xia, Y., Chen, B., Zhang, T., Wang, H., Qiu, H.: Autoregressive-based progressive coding for ultra-low bitrate image compression. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=FXu4G5T5QZ
48. Zhao, Y., Xiong, Y., Krähenbühl, P.: Image and video tokenization with binary spherical quantization. arXiv preprint arXiv:2406.07548 (2024)