

## What Matters in Practical Learned Image Compression

Kedar Tatwawadi

Parisa Rahimzadeh

Zhanghao Sun

Zhiqi Chen

Ziyun Yang

Sanjay Nair

Divija Hasteer

Oren Rippl

Apple

{kedar.tatwawadi, oren.rippel}@apple.com

### Abstract

One of the major differentiators unlocked by learned codecs relative to their hard-coded traditional counterparts is their ability to be optimized directly to appeal to the human visual system. Despite this potential, a perceptual yet practical image codec is yet to be proposed.

In this work, we aim to close this gap. We conduct a comprehensive study of the key modeling choices that govern the design of a practical learned image codec, jointly optimized for perceptual quality and runtime — including within the ablations several novel techniques. We then perform performance-aware neural architecture search over millions of backbone configurations to identify models that achieve the target on-device runtime while maximizing compression performance as captured by perceptual metrics.

We combine the various optimizations to construct a new codec that achieves a significantly improved tradeoff between speed and perceptual quality. Based on rigorous subjective user studies, it provides 2.3-3 $\times$  bitrate savings against AV1, AV2, VVC, ECM and JPEG-AI, and 20-40% bitrate savings against the best learned codec alternatives. At the same time, on an iPhone 17 Pro Max, it encodes 12MP images as fast as 230ms, and decodes them in 150ms — faster than most top ML-based codecs run on a V100 GPU.

### 1. Introduction

Since their emergence [1–3], learned image codecs have shown meaningful compression gains over traditional codecs. In recent years, the field has made significant progress in addressing several challenges that had once hindered practical deployment — improving computational efficiency, achieving fine-grained rate control with minimal overhead, and ensuring reliable cross-platform coding which is not inherent to hyperprior-based codecs [e.g. 4–9]. A major milestone in this evolution is the standardization of JPEG-AI [9], which not only highlights the technical maturity of learned codecs but also their growing industrial traction, signaling a clear transition beyond academic research.

Despite these remarkable advancements in building learned image codecs, a major opportunity remains largely

untapped. The key advantage of learned codecs over traditional hand-engineered approaches lies in their ability to be directly optimized for the task at hand — which is often

Image Codec	Perceptual BD-rate against PICO	12MP coding speed (ms)			Practicality		
		Device	Enc.	Dec.	Rate control	Cross-device	
Traditional	HEIC	227%	iPhone 17	68	14	✓	✓
	AV1	169%	iPhone 17	710*	41	✓	✓
	AV2	173%	CPU	3.6e6*	4,700*	✓	✓
	VVC (VTM)	151%	CPU	2.7e6*	2,600*	✓	✓
	ECM	148%	CPU	3.2e7*	18,000*	✓	✓
Learned	JPEG-AI	129% (HOP)	iPhone 17, HOP NN only	850	560	✓	✓
			iPhone 17, BOP NN only	130	72	✓	✓
	DCVC-RT Intra	172%	A100	140	130	✓	✓
			iPhone 17, NN only	710	980	✓	✓
	TCM	220%	RTX 3090	4,500	4,300	✗	✗
		V100	2,200	490,000	✗	✗	
		A100	2,400	2,600	✗	✗	
Learned & perceptual	C3-WD	40%	V100 encode, CPU decode	3.5e10	2,600*	✓	✓
			V100	1,100	3,100	✗	✗
	HiFiC	40%	iPhone 17, NN only	680	3,200	✗	✗
	MRIC	21%	V100	2,100	2,400	✗	✗
PICO (Ours)	0%	V100, NN only	140	82	✓	✓	
		iPhone 17, NN only	160	96	✓	✓	
		iPhone 17	230	150	✓	✓	

Figure 1. Comparisons of state-of-the-art traditional and learned codecs across different considerations of practicality. The reported perceptual BD-rates are based on human ratings from a large-scale subjective study (Sec. 5). For speed comparisons on iPhone 17 Pro Max, we use the exact architecture implementations found in the repositories of the baselines, and apply the same compiler optimizations as for PICO. Benchmarks marked with \* indicate that the runtime is expected to be faster once accelerated in hardware.

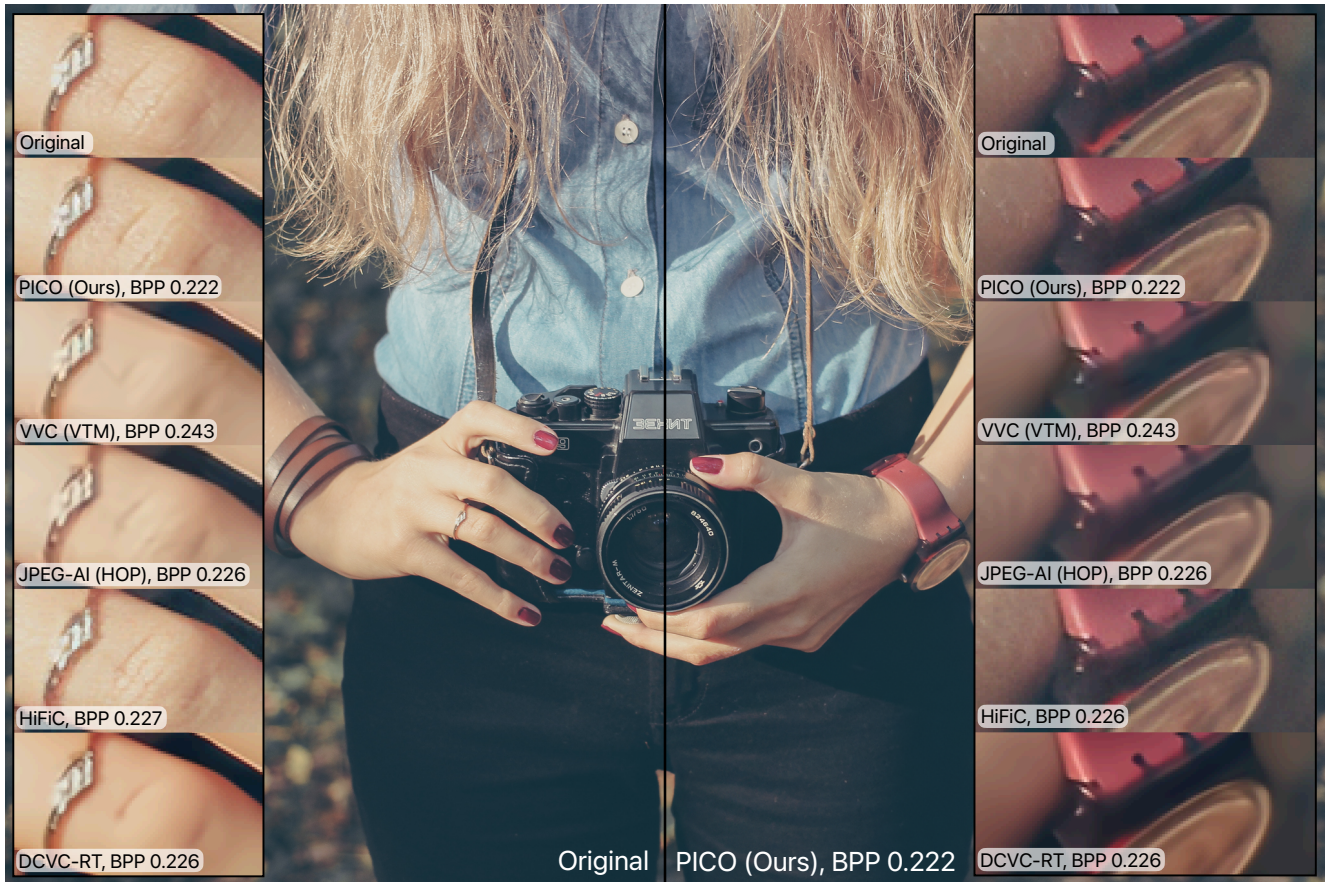


Figure 2. Qualitative comparisons of reconstruction quality for equal filesize/BPP (bits-per-pixel). PICO features significant improvements to fine-grained detail preservation, and even at low bitrates remains indistinguishable from the original.

to appeal to the human visual system. Several studies have explored this direction, establishing the foundations for applying modern generative techniques to image compression [10–13]. Although these works have demonstrated the exciting potential for perceptual optimization, their runtimes are an order of magnitude away from practical deployment. Moreover, most of them lack features necessary for any practical codec, such as cross-platform support or rate control.

In this work, we aim to close this gap. Our key contributions are as follows:

- We present the first work to comprehensively ablate across a broad spectrum of modeling decisions, and millions of model configurations, to explicitly optimize the trade-off between perceptual quality and runtime. The ablations include several novel architectures and algorithmic techniques, aimed at maximizing the codec’s expressivity — crucial for its generative capability — while explicitly avoiding incurring computational overhead.
- We introduce carefully-designed training and loss recipes that enable stable optimization of lightweight codecs towards high perceptual quality. We further propose specialized losses to surgically mitigate text and tiling artifacts.

- Building on these systematic ablations, we introduce **PICO (Perceptual Image Codec)**, a new image codec that integrates all essential components for practical deployment. Through extensive subjective user studies, PICO achieves 2.3–3× bitrate savings over AV1, AV2, VVC, ECM, and JPEG-AI, and 20–40% savings compared to the strongest learned codec baselines (Fig. 1, 2, 6). On an iPhone 17 Pro Max, PICO encodes 12MP images in as little as 230ms and decodes them in 150ms — faster than most state-of-the-art learned codecs run on a V100 GPU.

## 2. Related work

**Traditional image codecs** such as BPG [14], VVC [15], AV1 [16] and next-generation ECM [17] and AV2 [18] are based on hand-crafted pipelines that exploit redundancy by combining transformations with entropy coding. While these codecs have been extensively optimized, their design is fundamentally constrained by heuristically-designed components, leading to several limitations. For example, although they can be slightly tuned towards given metrics, their structure makes it inherently challenging to explicitly optimize them for perceptual quality. They also typically require dedi-

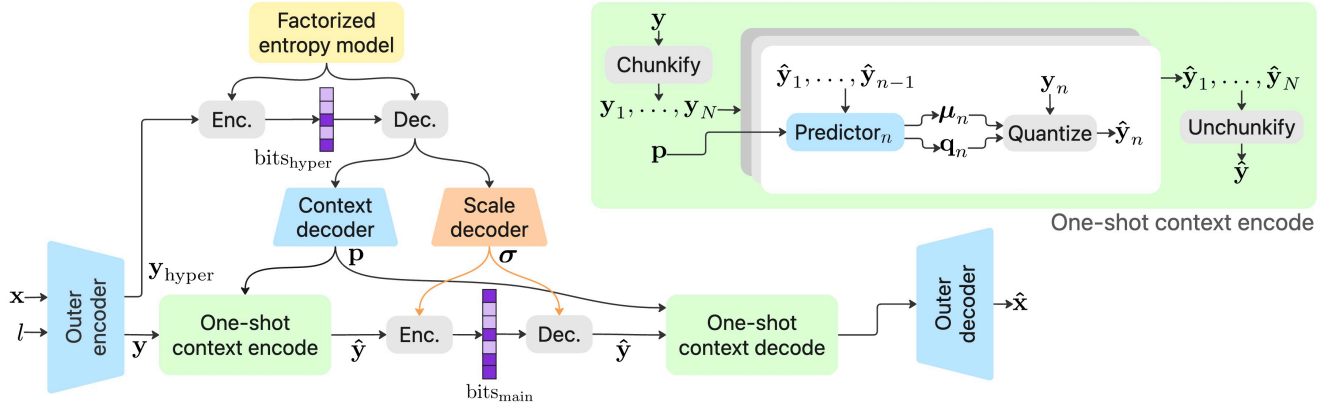


Figure 3. The overall model architecture. Individual components described in Sections 3 and 4.1. The **scale decoder** computation is bit-exact to guarantee entropy decodability.

cated hardware, leading to long adoption and update cycles.

**Learned image codecs** aim to resolve these issues via end-to-end modeling using neural networks [1–3], allowing them to be explicitly optimized to achieve optimal tradeoffs between bitrate, and given differentiable metrics. This unlocked the ability to train codecs directly for perceptual quality [10, 12, 19–21]. Recent research [11, 22] proposes to employ latent diffusion for image compression.

**Practical learned image compression** Despite their promise, learned image codecs have faced several major challenges. First, achieving high perceptual quality requires the model to align with the human visual system. Prior art introduced perceptual training objectives [10, 12, 23], yet still produce noticeable artifacts. Second, practical on-device deployment scenarios demand fast encoding and decoding. Many learned codecs (including all perceptual codecs mentioned) rely on heavyweight neural architectures [11, 21], autoregressive entropy models [24–27], or test-time optimization [12, 28] to enhance compression efficiency — at the cost of computational overhead. Recent research proposes more efficient neural architectures [8, 9] but focuses on metrics such as PSNR or SSIM, which poorly reflect perceptual quality. Third, encoding/decoding across devices with differing hardware/software configurations needs to be supported. Proposed enablements include integer-only coding to avoid inherent non-determinism in floating point operations [4], vector quantization to avoid decoding failures [29], and additional signaling to safeguard against errors [7].

### 3. Codec framework

Before diving into the details of the codec design search space, we first describe the framework at a high level.

#### 3.1. High-level codec framework

The ubiquitous hyperprior architecture described in [24] includes four sub-networks: encoder, decoder, hyper-encoder, and hyper-decoder. The encoder and decoder networks are

responsible for converting the input image  $x$  to a latent tensor  $\hat{y}$  and back to a reconstruction  $\hat{x}$ , while the hyper-encoder and hyper-decoder are used to provide parameters for entropy coding of the latent tensor  $\hat{y}$ . Specifically, the hyper-decoder outputs parameters location  $\mu$  and scale  $\sigma$ , which are used by the entropy coder to map to a discrete distribution for lossless coding of the latent  $\hat{y}$ .

At a high level, our model framework is similar to the hyperprior architecture, albeit with a few key differences. First, we split the hyper-decoder network into two sub-networks: a *scale decoder* and a *context decoder* (Fig. 3). The scale decoder outputs the scale parameter  $\sigma$  used for entropy coding of the latent  $\hat{y}$  and hence must produce the *exact same* output during the encoding and decoding processes given the extreme sensitivity of entropy decoding to parameter mismatch. Separating the scale decoder into a standalone model is crucial in facilitating guaranteed cross-device robustness, as well as unlocking additional speed gains via pipelining (Sec. 3.2). The context decoder can be thought of as a generalization of the location  $\mu$  output from the hyper-decoder model (see Sec. 4). Another key difference is that the hyper-encoder network is absorbed into the encoder network (Fig. 3). This simplification allows for the encoder to be compiled and executed as a single network.

#### 3.2. Extensions for practical deployment

We further extend this model in several ways to allow for practical deployment:

**Guaranteed cross-platform robustness** As was observed in many prior works [e.g. 4, 6–8], the parameters provided to the entropy coder must be *bit-exact*: the slightest discrepancy in computation between the entropy encoder and decoder will result in decoding failure. To guarantee success, we build the scale decoder to provide deterministic output across devices. We first quantize the model to UINT8 so that all the weights and activations within the network are integers. This step is necessary but in fact not sufficient, as there remain some floating point (FP) operations through the

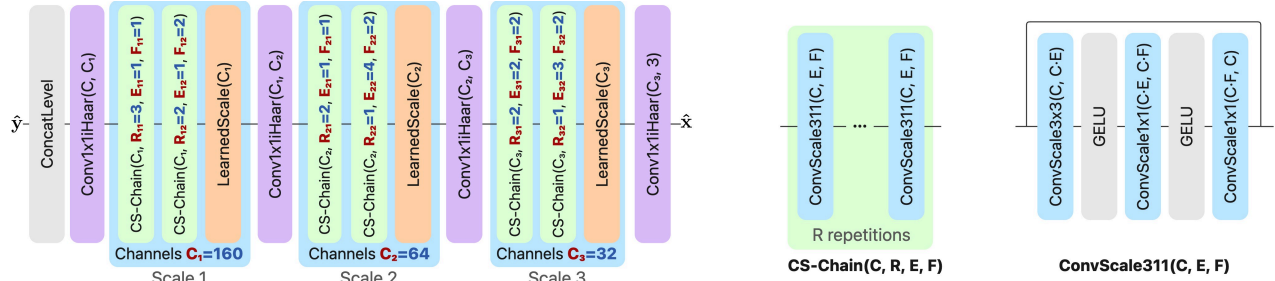


Figure 4. Detailed architecture of the outer decoder (see Appendix B for specifications of other model components). **Left:** We searched over millions of configurations from this model family, as defined by the **hyperparameters in red** with **optimal values in blue**, to achieve target iPhone runtimes while maximizing perceptual compression efficiency (see Sec. 4). **Right:** The architecture of the  $\text{ConvScale311}(C, E, F)$  module, with  $C$  channels, and expansion factors  $E, F$ . The base  $\text{ConvScale}$  layer is a reparametrization of a convolution with additional learned scales, and is described in Sec. 4.1. **Middle:** the  $\text{CS-Chain}(C, R, E, F)$  module simply repeats this block  $R$  times.

quantization scaling factors. Though these FP operations cannot be reordered by the compiler — a primary culprit for nondeterministic output — we cannot be sure how different hardware architectures may handle the FP arithmetic (*i.e.* precision and rounding modes). Thus, to achieve cross-platform determinism, we opt to run the scale decoder on CPU for compliance with the IEEE FP standard.

**Quality level control** We use a single model to represent the entire bitrate range, at negligible costs to both computation and model size. To do so, we condition the encoder and decoder networks, as well as loss definitions, on a scalar quality level  $l$  signaled in the bitstream. We follow the level embedding recipe described in Appendix E of [5] as our starting point, to which we apply several enhancements. The details can be found in Appendix F.

**Tile processing and pipelining** We introduce spatial tiling to improve computational efficiency. This enables pipelined execution, where the entropy coding and scale decoding of one tile run on the CPU while the neural components of another tile run concurrently on the accelerator. Each image is partitioned into non-overlapping tiles of size  $504 \times 504$ . During encoding, each tile is padded to  $512 \times 512$  with a 4-pixel contextual margin sourced from neighboring tiles. Including neighboring context helps maintain feature continuity across the tile boundary, partially mitigating tiling artifacts. Residual inconsistencies are further reduced by incorporating training losses which emphasize consistency across independent tile reconstructions (see Section 4.2).

### 3.3. Loss & training procedure

**Loss** Our combined rate-distortion loss function used for training is as described by Eq. 1. Similar to other perceptual-oriented learned codecs [*e.g.* 10, 19, 20], we use a combination of pixel-matching losses, perceptual losses, GAN-based losses, and losses to surgically mitigate specific artifacts. We ablate on different choices in detail in Section 4.2.

**Training procedure** We adopt the following training procedure for all experiments. The codec is trained on an inter-

nal dataset comprising approximately 90k generic images, analogous to ImageNet, supplemented with 2.3k images of text content, and another 28k high resolution open-sourced dataset from Div2K [30], CLIC [31], and Flickr2K [32]. We use the Adam optimizer [33]. The training is split into two phases: to start, the codec is trained solely on MSE; afterwards, the various perceptual losses introduced (see Sec. 4.2 and Appendix C for further detail).

## 4. Studying the codec design space

We comprehensively explore the codec design space, specifically focusing on directions that would not increase computational complexity. We explore large architectural changes in Section 4.1; perceptual optimizations in 4.2; and comprehensively search over how to best configure the backbone hyperparameters in 4.3. For all these experiments, we keep the training procedure (end of Sec. 3) constant.

### 4.1. Model Architecture enhancements

We present in detail modeling enhancements that are geared towards obtaining improved expressivity and capacity without impact on speed. Each enhancement is separately validated in the ablation studies (Sec. 5.2 and Tab. 1).

**Backbone and learned scales** Our starting point for the backbone of the encoder/decoder models is an inverted residual [34] with several modifications, which we call **ConvScale311** (Fig. 4). As validated by ablation studies, it provides a strong tradeoff between computational efficiency and expressivity. The architecture features different types of learned elementwise scales, which we find significantly improve the stability and performance of the model, at a negligible computational overhead:

1. Consider a convolution with  $C$  input channels,  $K$  output channels,  $G$  groups, and kernel size  $Y \times X$  with weight  $\mathbf{W}$  and bias  $\mathbf{b}$  of sizes  $[K, C//G, Y, X]$  and  $[K]$ . We define a new variant of the convolution layer we call **ConvScale**, which we supplement with two additional learned parameters: an input scale  $s_{\text{in}}$  and output scale  $s_{\text{out}}$  with shapes  $[1, C//G, 1, 1]$  and  $[K, 1, 1, 1]$ . We param-

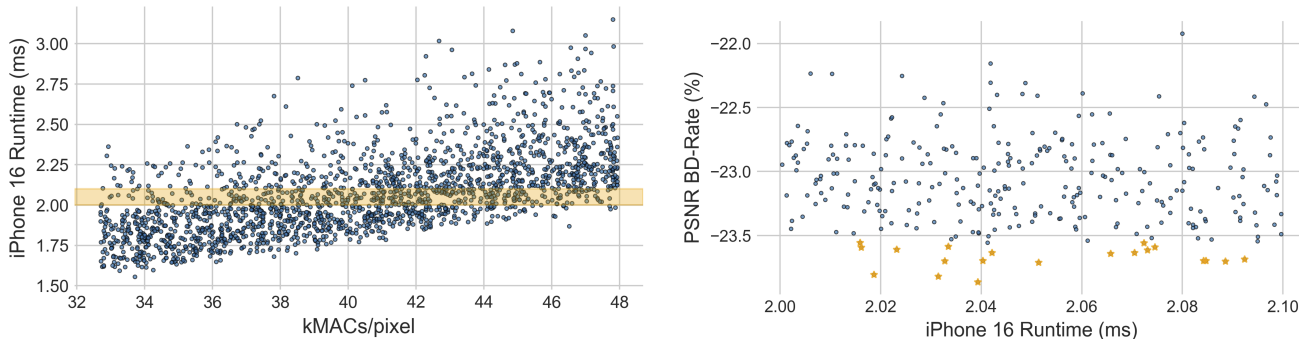


Figure 5. We perform neural architecture search for the outer decoder, progressively filtering the search space down from 1.4M model candidates to 20 models which are trained to completion (Sec. 4.3). Note that the runtime reflects the time taken to decode a single  $512 \times 512$  tile. **Left:** We benchmark the runtimes of 10,000 decoder candidates on-device, and show kMACs/pixel vs. iPhone 16 Pro runtimes (visualized for a sample of 2k models). These are further filtered by runtime, range highlighted in yellow, to choose a subset of 1,000 models to perform partial-training based filtering. **Right:** On-device runtime vs. PSNR BD-Rate for the 1,000 models trained (small subset visualized). Highlighted are the final shortlisted 20 models chosen to train to completion using the full perceptual recipe.

eterize the weight and bias to explicitly learn the scales as  $\mathbf{W}' = s_{\text{in}}s_{\text{out}}\mathbf{W}$  and  $\mathbf{b}' = \text{squeeze}(s_{\text{out}})\mathbf{b}$ . During inference we reparameterize  $\mathbf{W}'$  and  $\mathbf{b}'$  by collapsing the scales into them, leading to identical computational costs as for a normal convolution. We use ConvScale in place of all convolutions in the model.

2. We further introduce learned elementwise scaling factors that modulate activations near the end of each processing block corresponding to each spatial resolution (Fig. 4).

**Learned quantization width** It is a common methodology in learned compression for the hyperprior decoder to predict an elementwise location parameter  $\mu$  to shift the distribution used to code  $\mathbf{y}$  [24]. In our work, this is accomplished by the context decoder (Sec. 3); in addition, we find that it is helpful for the context decoder to also produce an input-specific elementwise learned quantization width  $\mathbf{q} > 0$  to adaptively modulate the width of the quantization bins. In practice, the context decoder (Fig. 3) produces a prior  $\mathbf{p}$  which is then mapped to  $\mu, \mathbf{q}$  by the context model (see below). We then quantize the main latent by rounding to the nearest integer  $\hat{\mathbf{y}} = \lfloor \frac{\mathbf{y} - \mu}{\mathbf{q}} \rfloor$ , which we then entropy-encode. After entropy-decoding, we invert the operations as  $\mathbf{q}\hat{\mathbf{y}} + \mu$ .

**One-shot context model** While learned codecs benefit significantly from autoregressive (AR) coding [e.g. 24–27], it results in slowdowns due to repeated back-and-forth memory transfers between the CPU and ML accelerator as entropy coding is interlaced with prediction. We observe, however, that this shortcoming is only a product of applying AR specifically to the *scale*  $\sigma$  which is required for entropy decoding. That is, if we decode the scale in a one-shot fashion, then we can freely apply iterative AR strategies to  $\mu, \mathbf{q}$  while keeping the computation exclusively on the ML accelerator. We refer to this as a *one-shot context model* (Fig. 3), which enjoys the benefits of AR at a negligible speed penalty. The iterative prediction structure can be chosen analogously to true

AR: for example, as channel-wise steps [35], checkerboard [25], and so on. We note that JPEG-AI [9] independently developed a component in a similar spirit, albeit applied to the  $\mu$  only, and with twice the AR prediction steps.

**Conv + Haar Resampling** Motivated by the Cosmos tokenizer [36], we employ 2D Haar wavelets for all resampling operations in the codec. Haar wavelets decompose the input into partially de-correlated channels in an invertible manner, with an analogous inverse transform. This can be interpreted as imposing an inductive bias on each learned resampling operation, promoting structured multi-scale representations and effectively increasing model capacity.

In this work, we introduce a reparametrization trick to add Haar/iHaar wavelets into the codec at *zero additional* computational cost; see Appendix H for full details.

## 4.2. Training loss enhancements

Keeping the model architecture constant, we notice that difference in training loss could lead to significant improvements to the model performance. Our cumulative distortion loss term used to train PICO is as follows:

$$D = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + w_1 \text{LPIPS}(\mathbf{x}, \hat{\mathbf{x}}) + w_2 \text{MS-SSIM}(\mathbf{x}, \hat{\mathbf{x}}) \\ + w_3 \text{TilingArtifactLoss}(\mathbf{x}, \hat{\mathbf{x}}) + w_4 \text{TextFidelityLoss}(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{m}) \\ + w_5 \text{GAN}(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{m})$$

We describe the rationale behind each loss term below.

**Pixel-matching & perceptual losses** In general, while the GAN significantly improves the visual realism, we notice that without appropriate pixel-matching + perceptual terms, it generates artifacts and hallucinates details. We moreover observe that a combination of pixel-matching and perceptual losses (MSE, LPIPS [37], MS-SSIM [38]) allows for better regularization of the GAN, as it can no longer exploit specific weaknesses within a single loss.

**Text artifact mitigation** The human visual system is extremely sensitive to distortions to text, where even the small-

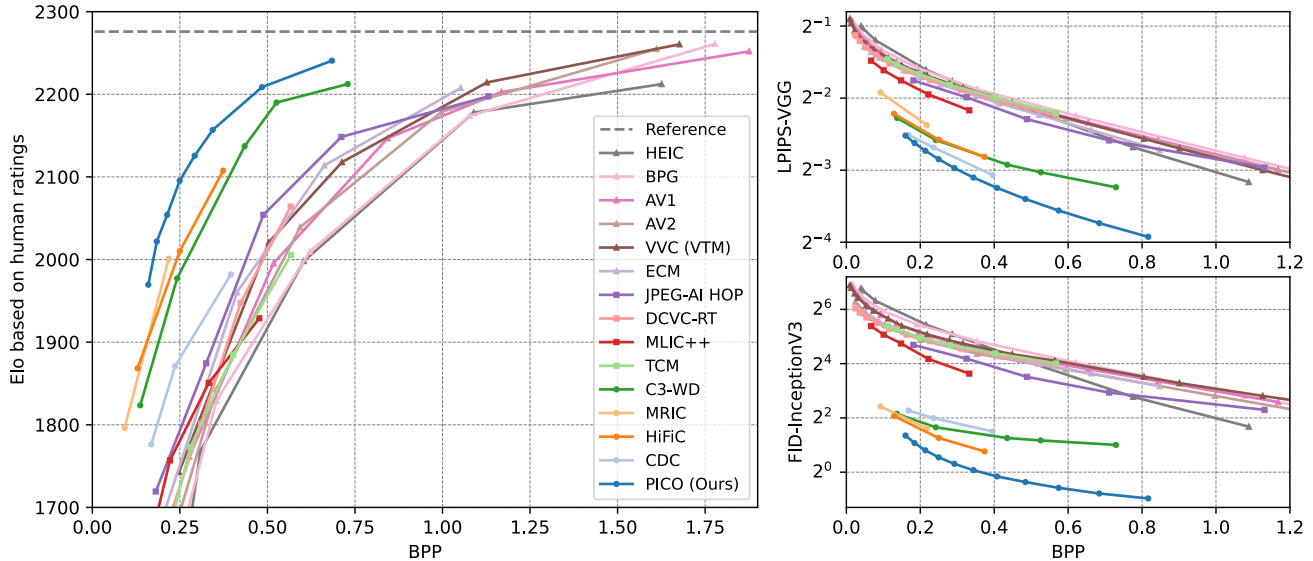


Figure 6. Rate-distortion curves of top traditional and learned codecs, based on Elo scores (higher is better) from a large-scale subjective study, and perceptual objective metrics (lower is better) on the CLIC 2020 test dataset. Traditional codecs are indicated with  $\blacktriangle$  markers, learned codecs with  $\blacksquare$ , and perceptual+learned codecs with  $\bullet$ . Evaluations on additional metrics and datasets can be found in Appendix A.

est hallucinations would render it unreadable. To this end, we augment the perceptual training with the *TextFidelityLoss* term. We use an off-the-shelf text detector [39] to generate a saliency mask. In the salient regions, a heavy L1 loss is then applied, while the GAN-based losses are subdued. In Section 5.2 we show the effectiveness of this approach.

**Low-frequency tiling artifact mitigation** PICO runs in a tiled fashion (Section 3.2), which leads to tiling artifacts in the absence of targeted mitigation. Specifically, perceptual losses and GAN in general ignore low spatial frequency components in the reconstruction, leading to color mismatch between neighboring tiles. To this end, we introduce *TilingArtifactLoss* (TAL), a multi-resolution L1 loss which imposes fidelity supervision on multiple spatial frequencies. We show ablations of this loss term in Section 5.2.

**GAN training & discriminator design** Consistent with the observations of [10, 20], we find that GAN-based training significantly improves the perceptual quality. Typically, a stronger discriminator provides better supervision to the generator (the codec), resulting in improved generation quality. We use a patch-wise discriminator architecture similar to [40], but boost the discriminator capacity by increasing the number of channels and convolution layers.

However, a larger discriminator leads to training instabilities, given that the lightweight decoder has limited capacity. We employ various strategies to stabilize GAN training. First, we utilize a two-stage training recipe. The first stage uses MSE as the only distortion loss. In the second stage, the perceptual fine-tuning stage, all distortion loss terms in Eq. 1 are added to optimize the perceptual quality. This approach improves stability by allowing the GAN-based

training to start with a reasonable initialization. We also follow a warm-up schedule by gradually increasing the weight of discriminator supervision as the training proceeds. This mitigates the risk of the compression model being misled while the discriminator is in the early stage of training.

### 4.3. Neural architecture search

On top of the high-level modeling decisions introduced in Section 4.1, we further conduct neural architecture search (NAS) to optimize over the large space of backbone hyperparameter (HP) choices. We search for models that maximize compression performance, while abiding by a target on-device runtime. We describe the process we followed for the decoder NAS; we follow similar processes for the other sub-models with details found in Appendix D.

We optimize over the decoder model family presented in Fig. 4 with the NN runtime target of 100ms for a 12MP image on an iPhone 16 Pro. This runtime threshold was chosen as the decoding speeds acceptable for real-life use. Naïvely taking the Cartesian product of the value sets for each HP results in  $\sim 1.4M$  candidate models. Given the huge number of candidate models, we proceed systematically to narrow the search space in a multi-step filtering process:

- kMACs/pixel filtering:** Given that computing operation counts is cheap, we use kMACs/pixel as a coarse form of filtering to eliminate candidates that are clearly out of bounds. Based on a preliminary analysis of typical runtimes as function of operation counts, we filter out any models with kMACs/pixel counts outside of [32.7, 48.0], reducing the search space to  $\sim 500k$  candidates.
- On-device runtime filtering:** Since MACs only loosely reflect runtime (see Fig. 5), we benchmark the actual

runtimes of randomly-sampled 10k models on an iPhone 16 Pro and filter models more than 5% away from the target runtime, resulting in  $\sim 1,000$  models.

- Compression performance filtering:** To reduce computational cost, we partially train the selected models for the first phase only (Sec. 3.3), and for 30% of the epochs. The results can be found in Fig. 5. We choose the top 20 models based on PSNR BD-rate.
- Full training of the final candidates:** Finally, we train the 20 models fully, and pick the top model based on performance on perceptual metrics and visual evaluation.

In Appendix D, we discuss the discovered architecture and provide intuition on why it provides a good tradeoff between capacity and speed. The encoder/decoder respectively have 15.2M/9.6M parameters, are 30.4MB/19.4MB on disk, and have peak memory use of 38.8MB/25.4MB on device.

## 5. Results

We consolidate insights from our exploration of the codec design space to develop PICO — a practical learned image codec optimized for alignment with human perception. In this section, we evaluate PICO’s performance in depth.

### 5.1. Evaluation procedure

**Datasets** We evaluate all the codecs on the commonly-used CLIC 2020 Test dataset [31], consisting of 428 images of varying resolutions. In Appendix A, we share subjective and objective results on the Kodak and DIV2K [30] datasets.

**Baselines** We comprehensively compare to state-of-the-art codecs; their specific configurations can be found in Appendix E. From the traditional codecs, we compare to HEIC, and the reference implementations of BPG [14], AV1 [16], VVC (VTM) [15] and of next-generation codecs AV2 [18] and ECM [17]. In terms of learned codecs, we compare to HiFiC [10], JPEG-AI [9], MLIC++ [41], CDC [11], TCM [42], MRIC [20], C3-WD [12, 23], and DCVC-RT [8]. For JPEG-AI, we evaluate the quality of the stronger-but-slower High Operation Point (HOP), and for completeness share speed benchmarks of also the Base Operation Point (BOP).

**Metrics** In this work, we focus exclusively on perceptual quality, and as such report on popular perceptually-aligned metrics: CMMD [43], FID [44] and LPIPS [37]. We report PSNR results in Appendix A, and observe that it poorly reflects perceptual quality — a well-known shortcoming.

**Subjective study** We conduct a large-scale subjective study using Mabyduck [45], an independent external platform for user preference studies. The study consists of pairwise blind A/B image comparison against a reference image, adopting the same standardized evaluation methodology employed by the CLIC compression challenge [31, 46]. We evaluate on the CLIC 2020 Test, Kodak, and DIV2K datasets and collect a total of 74,925 pairwise comparisons from 610 unique reviewers, independently screened by Mabyduck to

assure quality. Bayesian Elo scores [47, 48] are computed for each quality level of each codec based on all the pairwise comparisons, as reported in Figure 6. Extended description of the methodology can be found in Appendix G.

**Speed benchmarks** We report all baseline speed numbers as quoted in their original papers or repositories, other than the iPhone runtimes. For these, we implement the exact neural architectures of the approaches, and to ensure fair apples-to-apples comparisons, we deploy them on-device with all the optimizations we applied to PICO. We benchmark all approaches on the iPhone 17 Pro Max using the tiling strategy mentioned in Sec. 3 and report the neural runtimes. For PICO, we additionally report the end-to-end runtimes including all other codec components.

### 5.2. Findings

**Comparisons to baselines** We show quantitative comparisons based on subjective user studies and objective metrics in Figures 6, 9, 10 with a summary in Fig. 1. Qualitative comparisons can be found in Fig. 2 and Appendix J.

We observe that PICO significantly outperforms all prior traditional and learned codecs across both human ratings and perceptual quality metrics, and these gains generalize across datasets. Notably, compared with today’s best standardized codecs HEIC, AV1, and VVC (VTM), PICO has a BD-rate of over  $-60\%$  based on human ratings, suggesting a bitrate reduction of more than  $2.5\times$  for the same quality as

Property ablated	Option	CMMD-CLIP BD-Rate
One-shot autoregressivity	None	10.28%
	Channel-wise (4 groups)	3.10%
	Checkerboard	14.67%
Learned quantization width	2x2 grid	0%
	No*	8.16%
Learned Scale	Yes	0%
	None**	9.58%
	ConvScale only*	3.76%
	Per spatial scale only*	1.21%
Resampling	ConvScale + per spatial scale	0%
	Pixel reshuffling	19.51%
	Stride-2 Conv & Deconv*	8.90%
All above properties	Haar-based resampling	0%
	Disabled	31.69%
	Enabled	0%

Table 1. Architectural ablations, as evaluated on the CLIC 2020 testset. For each property, the BD-rate was computed with the anchor being the final chosen setting, in the last row. Every \* indicates halving of the learning rate to stabilize training.

Evaluation metric	Property ablated	Option	Value
L1 in text regions	Text fidelity loss	Off	0.0093
		On	0.0046
Low-frequency error across tile boundaries	Tiling artifact loss	Off	0.0020
		On	0.00097

Table 2. Artifact-specific loss ablations, as evaluated by specific metrics constructed to quantify the artifacts as described in Sec. 4.2. Visual examples can be found in Fig. 7.



Figure 7. Ablations on artifact-specific mitigation strategies. All images are encoded at a BPP 0.20. **Top:** Perceptual training leads to distortions in text, while adding the TextFidelityLoss enhances its fidelity. On the right we show the text saliency masks. **Bottom:** TilingArtifactLoss (TAL) mitigates color mismatch artifacts at tile boundaries (please zoom-in to better visualize). In the middle we show a slice in the green channel across the tile boundary, where reconstruction without TAL exhibits discontinuity. On the right we show a histogram of error around tile boundaries over the CLIC 2020 Professional Validation Set, which is significantly reduced by TAL.

evaluated by viewers. PICO also achieves a bitrate reduction of more than  $3\times$  as compared with BPG. The subjective Elo curves in Fig. 6 also suggest that HiFiC [10], MRIC [20], and C3-WD [12] are the three codecs which are the closest to PICO with respect to compression performance. However, they are all significantly slower and less practical, while achieving 20-40% larger file sizes for the same quality (Fig. 1). In general, we observe that codecs employing GANs or diffusion significantly perceptually outperform those without (e.g. JPEG-AI [9], MLIC++ [41]).

Qualitatively (see Fig. 2), PICO preserves considerably more detail than all other codecs, and produces more faithful reconstructions as compared with the original. More reconstruction examples are provided in Appendix J.

**Network architecture ablations** We conduct systematic network architecture ablations, as shown in Tab. 1, to isolate the contribution of each component to the overall compression performance. The benefit of adapting quantization width to local content is evident, as its removal results in a BD-rate increase of 8.16%. Similarly, replacing standard convolutions with our proposed ConvScale layers yields better stability and expressivity at no extra inference cost, while adding learned scaling provides additional performance gains — removing both results in a BD-rate increase of 9.58%. In our one-shot context model ablations, we find that removing the component altogether causes a large performance drop of 10.28%. Spatial AR strategies such as  $2\times 2$  grids or checkerboards deliver large improvements with minimal decoding overhead, while the minimal gains from

purely channel-wise AR suggest that spatial dependencies are rather more important to capture. Conv+Haar resampling emerges as the most effective strategy for downsampling and upsampling, outperforming both pixel shuffle/unshuffle and stride-2 convolutional alternatives, while introducing no additional computational cost. Removing all ablated properties results in a BD-rate degradation of 31.69%.

**Artifact mitigation ablations** We ablate on the text and tiling artifact mitigations. As shown in Fig. 7 top, decoded texts are not legible in the baseline, while adding TextFidelityLoss enhances text fidelity. For a quantitative comparison, we use a test set with  $\sim 100$  images with small texts. We use the same text detector [39] to label text regions, which are human-verified, and then calculate the absolute error within them (Tab. 2). The model trained with TextFidelityLoss achieves  $2\times$  lower error. In Fig. 7 bottom, we show that when TilingArtifactLoss (TAL) is missing in the training recipe, low-frequency color values visibly mismatch across tile boundaries. On the right, we show a histogram of errors across tile boundaries. The model trained with TAL has more than  $2\times$  lower cross-tile error (Tab. 2).

## 6. Conclusion

In this work, we introduce PICO, a new image codec designed for real-life use and optimized specifically for high perceptual quality. It is the product of systematic explorations of various architectural and training recipe choices, coupled with an architecture search over millions of backbone candidates to identify models that achieve optimal tradeoffs between speed and quality.

## References

- [1] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” *arXiv preprint arXiv:1611.01704*, 2016.
- [2] O. Rippel and L. Bourdev, “Real-time adaptive image compression,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 2922–2930.
- [3] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=rkcQFMZRb>
- [4] J. Ballé, N. Johnston, and D. Minnen, “Integer networks for data compression with latent-variable models,” in *International Conference on Learning Representations*, 2018.
- [5] O. Rippel, A. G. Anderson, K. Tatwawadi, S. Nair, C. Lytle, and L. Bourdev, “Elf-vc: Efficient learned flexible-rate video coding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 479–14 488.
- [6] K. Tian, Y. Guan, J. Xiang, J. Zhang, X. Han, and W. Yang, “Towards real-time neural video codec for cross-platform application using calibration information,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7961–7970.
- [7] J. Pang, M. A. Lodhi, J. Ahn, Y. Huang, and D. Tian, “Towards reproducible learning-based compression,” in *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2024, pp. 1–6.
- [8] Z. Jia, B. Li, J. Li, W. Xie, L. Qi, H. Li, and Y. Lu, “Towards practical real-time neural video compression,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-25, 2024*, 2025.
- [9] “JPEG AI Reference Software,” <https://gitlab.com/wg1/jpeg-ai/jpeg-ai-reference-software>, 2025.
- [10] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” *Advances in neural information processing systems*, vol. 33, pp. 11 913–11 924, 2020.
- [11] R. Yang and S. Mandt, “Lossy image compression with conditional diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 971–64 995, 2023.
- [12] J. Ballé, L. Versari, E. Dupont, H. Kim, and M. Bauer, “Good, cheap, and fast: Overfitted image compression with wasserstein distortion,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 259–23 268.
- [13] 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,” 2022. [Online]. Available: <https://arxiv.org/abs/2205.14501>
- [14] F. Bellard, “libbpg: BPG (Better Portable Graphics) image library,” <http://bellard.org/bpg/libbpg-0.9.5.tar.gz>, 2015, version 0.9.5, released 2015-01-11.
- [15] Joint Video Experts Team (JVET), “VVCSoftware\_VTM: VVC VTM Reference Software,” [https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware\\_VTM/-/releases/VTM-23.11](https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/releases/VTM-23.11), 2025, version 23.11, released 2025-07-03.
- [16] Alliance for Open Media, “AOM: AV1 codec library, version 3.12.1,” Alliance for Open Media, Git repository, April 2025, tag: v3.12.1, commit 10aece4, tagged fc5cf6a. [Online]. Available: <https://aomedia.googlesource.com/aom/+refs/tags/v3.12.1>
- [17] Joint Video Experts Team (JVET), “Enhanced Compression Model (ECM) Reference Software,” Fraunhofer HHI, GitLab repository, September 2025, branch: master, commit dcc311af, accessed 2025-09-25. [Online]. Available: <https://vcgit.hhi.fraunhofer.de/ecm/ECM>
- [18] Alliance for Open Media, “AVM: AV2 codec research anchor, version research-v11.0.0,” Alliance for Open Media, GitLab repository, September 2025, tag: research-v11.0.0, commit 3a5da21a. [Online]. Available: <https://gitlab.com/AOMediaCodec/avm/-/tags/research-v11.0.0>
- [19] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, “Generative adversarial networks for extreme learned image compression,” *arXiv preprint arXiv:1804.02958*, 2018.
- [20] 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*, 2023, pp. 22 324–22 333.
- [21] 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*, 2024, pp. 26 088–26 098.
- [22] L. Relic, R. Azevedo, Y. Zhang, M. Gross, and C. Schroers, “Bridging the gap between diffusion models and universal quantization for image compression,” in *Machine Learning and Compression Workshop@ NeurIPS 2024*. OpenReview, 2024.
- [23] [Online]. Available: <https://orange-opensource.github.io/Cool-Chic/>

- [24] D. Minnen, J. Ballé, and G. Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” *arXiv preprint arXiv:1809.02736*, 2018.
- [25] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, “Checkerboard context model for efficient learned image compression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 771–14 780.
- [26] 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, 2022*, pp. 5718–5727.
- [27] J. Li, B. Li, and Y. Lu, “Neural video compression with feature modulation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 17-21, 2024*, 2024.
- [28] H. Kim, M. Bauer, L. Theis, J. R. Schwarz, and E. Dupont, “C3: High-performance and low-complexity neural compression from a single image or video,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9347–9358.
- [29] X. Lu, H. Wang, W. Dong, F. Wu, Z. Zheng, and G. Shi, “Learning a deep vector quantization network for image compression,” *IEEE Access*, vol. 7, pp. 118 815–118 825, 2019.
- [30] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- [31] [Online]. Available: <https://clic2025.compression.cc/tasks/#image>
- [32] R. Timofte, E. Agustsson, L. V. Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, X. Wang, Y. Tian, K. Yu, Y. Zhang, S. Wu, C. Dong, L. Lin, Y. Qiao, C. C. Loy, W. Bae, J. Yoo, Y. Han, J. C. Ye, J.-S. Choi, M. Kim, Y. Fan, J. Yu, W. Han, D. Liu, H. Yu, Z. Wang, H. Shi, X. Wang, T. S. Huang, Y. Chen, K. Zhang, W. Zuo, Z. Tang, L. Luo, S. Li, M. Fu, L. Cao, W. Heng, G. Bui, T. Le, Y. Duan, D. Tao, R. Wang, X. Lin, J. Pang, J. Xu, Y. Zhao, X. Xu, J. Pan, D. Sun, Y. Zhang, X. Song, Y. Dai, X. Qin, X.-P. Huynh, T. Guo, H. S. Mousavi, T. H. Vu, V. Monga, C. Cruz, K. Egiazarian, V. Katkovnik, R. Mehta, A. K. Jain, A. Agarwalla, C. V. S. Praveen, R. Zhou, H. Wen, C. Zhu, Z. Xia, Z. Wang, and Q. Guo, “Ntire 2017 challenge on single image super-resolution: Methods and results,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1110–1121.
- [33] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [35] 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, pp. 3339–3343.
- [36] NVIDIA, :, N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai, P. Chattopadhyay, Y. Chen, Y. Cui, Y. Ding, D. Dworakowski, J. Fan, M. Fenzi, F. Ferroni, S. Fidler, D. Fox, S. Ge, Y. Ge, J. Gu, S. Gururani, E. He, J. Huang, J. Huffman, P. Jannaty, J. Jin, S. W. Kim, G. Klár, G. Lam, S. Lan, L. Leal-Taixe, A. Li, Z. Li, C.-H. Lin, T.-Y. Lin, H. Ling, M.-Y. Liu, X. Liu, A. Luo, Q. Ma, H. Mao, K. Mo, A. Mousavian, S. Nah, S. Niverty, D. Page, D. Paschalidou, Z. Patel, L. Pavao, M. Ramezanali, F. Reda, X. Ren, V. R. N. Sabavat, E. Schmerling, S. Shi, B. Stefaniak, S. Tang, L. Tchapmi, P. Tredak, W.-C. Tseng, J. Varghese, H. Wang, H. Wang, H. Wang, T.-C. Wang, F. Wei, X. Wei, J. Z. Wu, J. Xu, W. Yang, L. Yen-Chen, X. Zeng, Y. Zeng, J. Zhang, Q. Zhang, Y. Zhang, Q. Zhao, and A. Zolkowski, “Cosmos world foundation model platform for physical ai,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.03575>
- [37] 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, pp. 586–595.
- [38] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in *Signals, Systems and Computers, 2004.*, vol. 2. Ieee, 2003, pp. 1398–1402.
- [39] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character region awareness for text detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9365–9374.
- [40] 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, pp. 1125–1134.
- [41] W. Jiang, J. Yang, Y. Zhai, F. Gao, and R. Wang, “Mlic++: Linear complexity multi-reference entropy modeling for learned image compression,” *arXiv preprint arXiv:2307.15421*, 2023.
- [42] J. Liu, H. Sun, and J. Katto, “Learned image compression with mixed transformer-cnn architectures,” in *Pro-*

- ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 388–14 397.
- [43] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, “Rethinking fid: Towards a better evaluation metric for image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9307–9315.
- [44] 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*, vol. 30, 2017.
- [45] [Online]. Available: <https://www.mabyduck.com>
- [46] B. Naderi and R. Cutler, “A crowdsourcing approach to video quality assessment,” 2023. [Online]. Available: <https://arxiv.org/abs/2204.06784>
- [47] F. Caron and A. Doucet, “Efficient bayesian inference for generalized bradley-terry models,” 2010. [Online]. Available: <https://arxiv.org/abs/1011.1761>
- [48] [Online]. Available: <https://docs.mabyduck.com/experiments/metrics/elo>
- [49] [Online]. Available: <https://docs.mabyduck.com/experiments/strategies>
- [50] S. Ishihara, *Tests for color-blindness*. Tokyo, Hongo Harukicho: Handaya, 1917.