Conditional Hallucinations for Image Compression

Till Aczel
ETH Zürich
taczel@ethz.ch

Roger Wattenhofer
ETH Zürich
wattenhofer@ethz.ch

Abstract

In lossy image compression, models face the challenge of either hallucinating details or generating out-of-distribution samples due to the information bottleneck. This implies that at times, introducing hallucinations is necessary to generate in-distribution samples. The optimal level of hallucination varies depending on image content, as humans are sensitive to small changes that alter the semantic meaning. We propose a novel compression method that dynamically balances the degree of hallucination based on content. We collect data and train a model to predict user preferences on hallucinations. By using this prediction to adjust the perceptual weight in the reconstruction loss, we develop a **Con**ditionally **Ha**llucinating compression model (**ConHa**) that outperforms state-of-the-art image compression methods. Code and images are available at https://polybox.ethz.ch/index.php/s/owS1k5JYs4KD4TA.

1 Introduction

Lossy compression is characterized by a trade-off between reducing the number of communicated bits (rate) and the undesirable introduction of content changes (distortion), presenting a constrained optimization challenge. For image compression measuring the distortion or "how close a compressed image is to the original image" is a challenging task. A distortion metric aligned with human perception should assess not only pixel-wise similarity but also how much content changes and the overall realism of the reconstructed image. Applying a slight blur to an image containing small text can make it unreadable, drastically increasing distortion, while blurring an image of a uniform blue sky has minimal impact. For texture-like content, such as grass, freckles, and stone walls, generating pixels that realistically match a given texture is more important than reconstructing precise pixel values; generating any sample from the distribution of a texture is generally sufficient. Conversely, for images featuring small text, straight lines, and small faces, data reconstruction should be pixel-by-pixel to preserve fidelity [1]. Distortion metrics often struggle to capture these subtleties of human perceptual preference.

Learned image compression methods directly optimize the rate-distortion trade-off. Thus the limitations of distortion metrics do not only impact their evaluation but also affect the training process. To overcome these shortcomings, state-of-the-art methods [1; 2] optimize the rate-distortion-perception trade-off, rather than just the rate-distortion trade-off. Here, perception refers to how closely the compressed image resembles the original in terms of realness or visual fidelity. This can be achieved with a GAN-like discriminator, encouraging models to generate samples that match the distribution of real images, enhancing perceptual quality. However, optimizing this trade-off can introduce hallucinations, where generated images contain details not present in the original data. This is undesirable, particularly when small hallucinations alter the semantic content. In Figure 1, we present two images: one with small text and another with grass, compressed by two models—one avoiding hallucinations and the other generating in-distribution samples. The optimal rate-distortion-perception trade-off varies based on image content and compression rate. For text, hallucinations should be avoided, while for grass, generating in-distribution samples replicating texture, even with hallucinations, is preferred.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

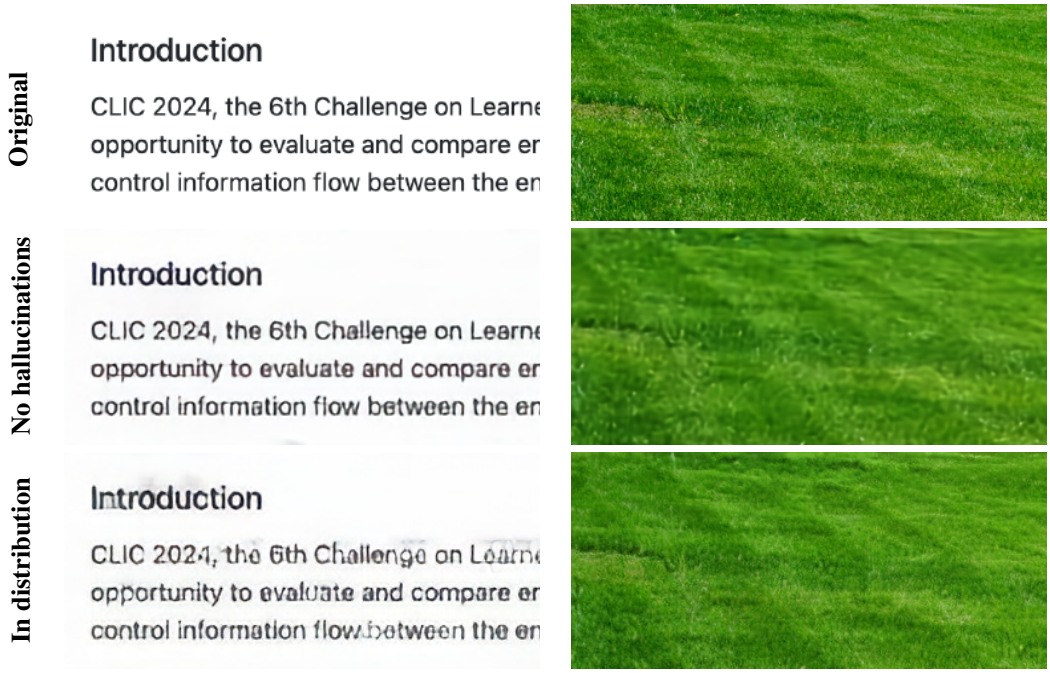

Figure 1: 1) Original image 2) compressed with an MSE distortion optimized model 3) compressed with a hallucinating MSE+GAN optimized model. For the left image containing text, hallucinations degrade the image quality. For the right image containing grass, an in-distribution sample with hallucinations produces a higher perceptual quality.

In this paper, we gather labels and train a classifier to predict user preferences for the level of detail to hallucinate, based on the original uncompressed image. By using the classifier's prediction as a weight for the GAN discriminator in the reconstruction loss, we train a **Con**ditionally **Ha**llucinating compression model, **ConHa**. This model automatically adjusts the reconstruction process based on image content, determining whether to minimize hallucinations or generate in-distribution samples. To the best of our knowledge, this is the first work to introduce automatic balancing between distortion and perception in image compression. The degree of hallucination varies not only between different images but also within a single image. By dynamically adjusting hallucination levels, we surpass state-of-the-art compression methods.

2 Related work

Traditional lossy image compression models like JPEG [3], JPEG2000 [4], WebP [5], and BPG [6] are widely used. Learned lossy image compression, which optimizes the relaxed rate-distortion trade-off, is conceptually similar to Variational Autoencoders (VAEs) [7; 8], but requires latent quantization for compression [9]. Using this similarity, [9] developed a non-linear codec that outperforms traditional methods. Building on this, they introduced Hyperprior [10], a hierarchical VAE that jointly learns and compresses both the latent variable and its prior. More recently, ELIC [11] explores architectural improvements to propose a high-performing and computationally efficient learned image codec.

In the rate-distortion trade-off, the rate has an accurate approximation, but measuring distortion is much harder. Handcrafted metrics like PSNR and MS-SSIM often poorly align with user studies and can even negatively correlate with human preferences among the best-performing models [12]. The hidden representations of trained CNNs correlate strongly with human preferences [13], which the Learned Perceptual Image Patch Similarity (LPIPS) metric [13] leverages by fine-tuning a CNN on two-alternative forced choice (2AFC) data. Distortion metrics often fail to accurately capture human perception, leaving models vulnerable to adversarial attacks and resulting in out-of-distribution samples. The Fréchet Inception Distance (FID) [14] measures alignment between image distributions, evaluating the realism of generated images. Generative Adversarial Networks (GANs) [15] address

Figure 2: Samples from the CLIC 2024 image test set with w as their x coordinate. Images predicted to perform better without hallucinations are on the left, while those predicted to excel with in-distribution samples are on the right.

this by using a generator and a discriminator to ensure in-distribution sample generation. Variants of GAN-like adversarial losses [16; 17; 18; 19; 20; 1; 2] effectively rectify compression artifacts and generate desired distribution samples. Notably, [20] introduces the rate-distortion-perception trade-off, using perception as a metric for realism assessment. HiFiC [1] is a state-of-the-art image compressor favored in user studies over previous codecs.

Attaining perfect realism often leads to undesired distortion [21] and introduces hallucinations into generated content. Our research aims to automatically determine the optimal balance between distortion and perception. While some methods allow manual control over the distortion-perception trade-off [22; 23], enabling users to set realism levels for the entire image, our approach automates this process. Automation is beneficial as users typically prefer convenience and may not clearly understand their desired level of hallucinations. Instead of relying on user input for adjustments, our method selects realism levels for individual image parts based on content, allowing users to enjoy in-distribution samples with minimized hallucinations without altering the semantic meaning.

3 Methodology

Human preferences in the trade-off between rate, distortion, and perception vary with image content. A compression model reflecting human perception must balance avoiding hallucinations and maintaining fidelity to in-distribution samples. To capture this nuance, we collected a dataset of 5,408 preference choices between a Mean Squared Error (MSE)-based model and a Generative Adversarial Network (GAN)-based model. First, we learn a preference model that predicts which compression method is preferred based on image content. The perceptual compression loss is then scaled according to this prediction. For images where preserving exact content is crucial and hallucinations are undesirable, the compression loss emphasizes rate and distortion. Conversely, for images where generating an in-distribution sample suffices and exact pixel-wise accuracy is less critical, the model optimizes for rate, distortion, and also perceptual quality. This approach yields a compression model that aligns better with human perceptual preferences.

3.1 Hallucination-distribution preference model

To align the compression model's loss term with human preferences, obtaining human labels for all crops of all training images is too expensive. Instead, we train a preference model M_P on the labeled data, namely a binary classifier, which can then be utilized during the compression model training. We provide further training details in Appendix C. In Figure 2, we visualize the preference

model M_P predictions. Images with text, small faces, or straight lines (where hallucinations can alter semantic content) are on the left, while images of grass, skin, cloth, and other textures are on the right. Unlike LPIPS, which has two inputs with variable distortion types, M_P only needs to predict distortion-realism human preferenc. This constraint allows the preference model to learn human preferences more accurately.

3.2 Compression model

Rate-distortion optimized compression: We follow the autoencoder-based learned lossy image compression that optimizes the rate-distortion trade-off. This method comprises a learned encoder E and generator G. An image x from distribution p_x is encoded by E into a quantized representation y, which is then decoded by G to reconstruct the image x'. The distribution of the quantized representation y is learned by a probability model P, which is then used to further compress y using an entropy coding algorithm. The rate-distortion trade-off can be jointly optimized by:

$$\mathcal{L}_{rd} = \mathbb{E}_{x\sim p_x}\left[\lambda r(y) + d(x, x')\right], \tag{1}$$

where λ controls the trade-off, r is an approximation of the bit-stream length $r(y) = -log(P(y))$, and d is a distortion metric (in our case MSE).

Rate-distortion-perception optimized compression: Models optimizing the rate-distortion trade-off with imperfect distortion metrics generate out-of-distribution samples with visible artifacts. To address this, the trade-off can be extended to include perceptual quality, leading to a rate-distortion-perception trade-off. However, current perceptual metrics like LPIPS are vulnerable to adversarial attacks, causing models to struggle with accurate image reconstruction. Previous works [16; 17; 18; 19; 20; 1; 2] improve the variational autoencoder by integrating a perceptual metric and a GAN-like discriminator, resulting in the objective:

$$\mathcal{L}_{rdp} = \mathbb{E}_{x\sim p_x}\left[\lambda r(y) + d(x, x') + \beta(d_{LPIPS}(x, x') - log(D(x', y)))\right], \tag{2}$$

where β is the perception weight, d_{LPIPS} is the LPIPS metric and D is a GAN-like discriminator. The compression model and discriminator are trained alternately, with the discriminator optimizing:

$$\mathcal{L}_D = \mathbb{E}_{x\sim p_x}\left[-log(1 - D(x', y)))\right] + \mathbb{E}_{x\sim p_x}\left[-log(D(x, y)))\right] \tag{3}$$

We have grouped D_{LPIPS} and D together since these terms control the perceptual quality and enforce that the reconstructed image is within the image distribution p_x.

Content-dependent rate-distortion-perception optimized compression: The optimal balance in the rate-distortion-perception trade-off varies with the image content. By adjusting the weight assigned to perceptual quality in the rate-distortion-perception loss function, the compression model learns when to add details to create an in-distribution sample:

$$\mathcal{L}_{wrdp} = \mathbb{E}_{x\sim p_x}\left[\lambda r(y) + d(x, x') + \beta w(d_{LPIPS}(x, x') - log(D(x', y)))\right], \tag{4}$$

where w is the weight predicted by the hallucination-distortion preference model M_P for image x. For images of at least 64×64, there is no information flow between distant pixels. Thus, while w remains static during training, the model determines hallucination levels based on a limited context window during inference, resulting in varying amounts of hallucinations within an image. Although we use HiFiC as our baseline, this correction method can be applied to any model utilizing a VAE with a GAN discriminator. Training details are in Appendix C.

4 Experiments

Datasets: Due to our two-stage training process, we require two distinct training sets. The preference model, is trained using the DIV2K training set [24], while the compression models are trained on the Vimeo90K dataset [25]. The user study is performed on the CLIC 2024 image test set [12], which consists of 32 diverse high-resolution (above 2 megapixels) images.

Baselines: Our approach integrates the Mean & Scale Hyperprior [10] and HiFiC [1] compression models, both sharing the same architecture. While the Hyperprior model optimizes solely optimizes Mean Squared Error (MSE) as the distortion metric, HiFiC combines MSE with perceptual loss using LPIPS [13] and a GAN discriminator. We modify the HiFiC model by making the hyperparameter that controls the perception loss weight image conditional. To assess the impact of this image conditionality, we also train a model with a fixed weight, ConHa-fixed, calculated as the average weight across the entire training dataset.

Figure 3: Comparison of compression methods. Our approach represents a middle ground between the Hyperprior and HiFiC models. For images with pavement (first column), our model adds details to create in-distribution samples. It avoids excessive hallucination for images with small faces and text (middle column) or objects with straight edges (right column).

4.1 Qualitative assessment

Our method strikes a balance between optimizing the rate-distortion trade-off, which can result in blurry images with no hallucinations, and optimizing the rate-distortion-perception trade-off, which can produce in-distribution samples but with too many hallucinations. In Figure 3, we compare our compression model to baseline models using an image featuring diverse objects. For small faces, straight lines, and text on a shirt, our method avoids excessive generating details that could obscure recognition. Conversely, it produces in-distribution samples for textures like road surfaces, larger faces, and rust on a bridge. Depending on the image content, our method generates samples nearly indistinguishable from either the Hyperprior model or HiFiC. While any small crop may have a similarly performing compression model, models optimizing for a fixed point on the rate-distortion-perception trade-off tend to underperform across many images. Thus, our approach consistently generates high-quality images regardless of content, achieving a superior balance in the rate-distortion-perception trade-off.

4.2 Quantitative assessment

Computational distortion metrics often fail to predict human preferences accurately, highlighting the need for user studies. In our study, we collected 1531 comparisons from 40 participants, each completing a maximum of 50 comparisons, with a median time of 7.8 seconds per comparison.

Details aboutthe user study can be found in Appendix A. Figure 4 presents the results, showing models on the x-axis and bootstrapped Elo scores on the y-axis across three bitrates: low, medium, and high. Across all bitrates, our model consistently outperforms HiFiC, the previous state-of-the-art compression model. At low bitrates, we compare our model to the extremes of the compression spectrum, Hyperprior and HiFiC, as well as our model with a fixed ratio. Our method, a middle ground between Hyperprior and HiFiC outperforms both extremes. When the weight of perceptual losses is fixed instead of being image conditional, performance drops to near HiFiC levels. This result underscores the importance of conditioning the weight of perceptual losses on the image for optimal performance in image compression.

In Appendix B, we provide a comprehensive comparison of our method and the baseline models using various computational metrics on the CLIC 2024, CLIC 2020 [12] and Kodak dataset [26]. ConHa demonstrates comparable performance to Hyperprior and HiFiC across all metrics, with no definitive best model. As observed in the CLIC [12] challenges, automated metrics do not align with human evaluations.

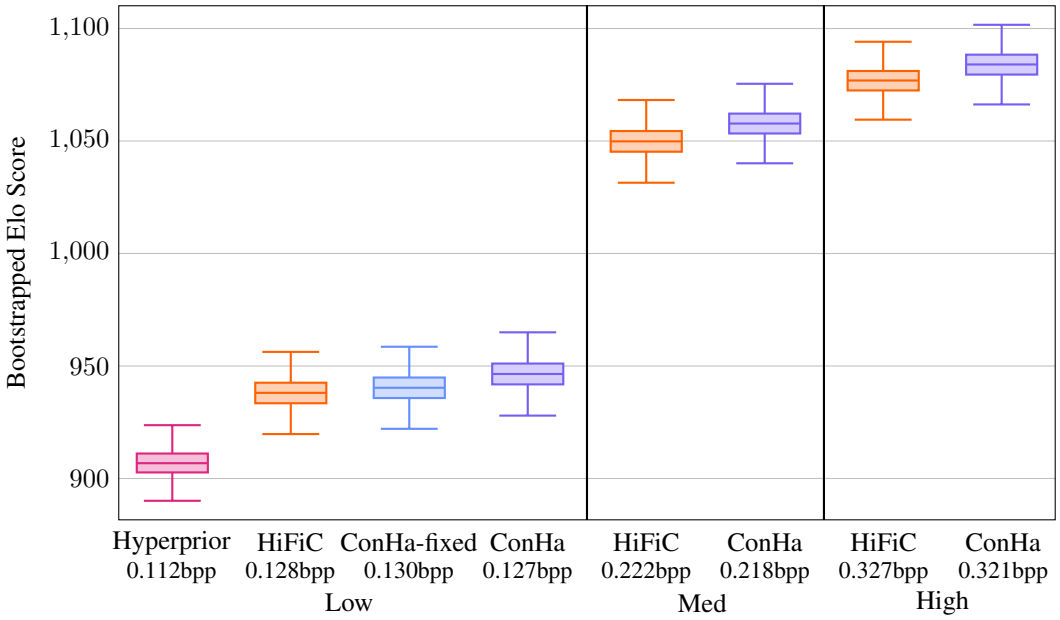

Figure 4: Bootstrapped Elo Scores box plot on the CLIC 2024 image test set. Each box represents the distribution of Elo Scores, with the horizontal line indicating the median, the box extending from the first quartile (Q1) to the third quartile (Q3), and the whiskers extending to 1.5 times the interquartile range (Q3-Q1). At low bitrates, ConHa (ours) is compared against three baselines: Hyperprior, HiFiC, and ConHa-fixed. For medium and high bitrates, ConHa is compared solely against the previous state-of-the-art, HiFiC. Remarkably, at all bitrates, ConHa consistently outperforms the baseline models, as evidenced by the higher median Elo Scores.

5 Conclusion

In the realm of lossy image compression, striking a balance between minimizing distortion and maintaining perceptual fidelity poses a significant challenge. Traditional distortion metrics often fail to align with human perceptual preferences, resulting in compression models that either hallucinate excessive details or generate out-of-distribution samples. By incorporating a classifier trained to predict user preferences regarding detail hallucination, we introduce automatic balancing between distortion and perception in image compression. This novel approach allows our model to adapt its compression strategy dynamically, optimizing performance across a wide range of image content and compression rates. ConHa carefully chooses when and what parts of images to hallucinate in a way that aligns with human perception. By addressing the limitations of distortion metrics and introducing dynamic content-aware balancing between distortion and realism, our method outperforms state-of-the-art compression models.

References

[1] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020.

[2] Dailan He, Ziming Yang, Hongjiu Yu, Tongda Xu, Jixiang Luo, Yuan Chen, Chenjian Gao, Xinjie Shi, Hongwei Qin, and Yan Wang. Po-elic: Perception-oriented efficient learned image coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1764–1769, 2022.

[3] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[4] David S Taubman and Michael W Marcellin. Jpeg2000: Standard for interactive imaging. *Proceedings of the IEEE*, 90(8):1336–1357, 2002.

[5] Google Developers. Webp. `https://developers.google.com/speed/webp/`. Accessed: May 17, 2024.

[6] Fabrice Bellard. BPG (Better Portable Graphics). `https://bellard.org/bpg/`. Accessed: May 17, 2024.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[8] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[9] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

[10] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

[11] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan 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*, pages 5718–5727, 2022.

[12] Compression.cc. Compression.cc. `https://www.compression.cc/`. Accessed on May 17, 2024.

[13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[16] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, pages 2922–2930. PMLR, 2017.

[17] Shibani Santurkar, David Budden, and Nir Shavit. Generative compression. In *2018 Picture Coding Symposium (PCS)*, pages 258–262. IEEE, 2018.

[18] Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. *Advances in neural information processing systems*, 31, 2018.

[19] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019.

[20] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[21] George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. *Advances in Neural Information Processing Systems*, 34:11517–11529, 2021.

[22] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22324–22333, 2023.

[23] Shoma Iwai, Tomo Miyazaki, and Shinichiro Omachi. Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2900–2909, 2024.

[24] Eirikur Agustsson and Radu Timofte. Div2k dataset, 2017. Accessed: 2024-05-22.

[25] Tianfan Xue, Jiajun Wu, Katherine L. Bouman, and William T. Freeman. TOFlow: Temporal output learning for video frame interpolation. `http://toflow.csail.mit.edu/`, 2019.

[26] Kodak graphics. `http://r0k.us/graphics/kodak/`.

[27] Mark E Glickman. A comprehensive guide to chess ratings. *American Chess Journal*, 3(1):59–102, 1995.

A User study

The user study is used in both data collectoin for the preference model training and as evaluation. Following previous studies [12; 1], we utilize the two-alternative forced choice (2AFC) labeling method. We decided to implement a web application for our user study, as shown in the screenshot in Figure 5. Users begin by seeing the task, "Select the compressed image that looks closer to the original." followed by instructions. The participant can see the original image in the center of the interface. Using the arrow keys, they can select an area of interest. By pressing and holding keys 1 and 2, participants can switch between an image compressed by Model A and Model B. To restrict the decision area to 786×786 pixels, panning is disabled once a compressed image is displayed. To submit their decision, participants hold down either key 1 or 2 and then press the space bar. The instructions are displayed line by line to avoid overwhelming the participant and can be toggled on and off using the 'I' key. The task "Select the compressed image that looks closer to the original" is always displayed on the screen. Additionally to the label, the interface tracks the number of key presses and the time spent on each image. The source code is provided in the supplementary material.

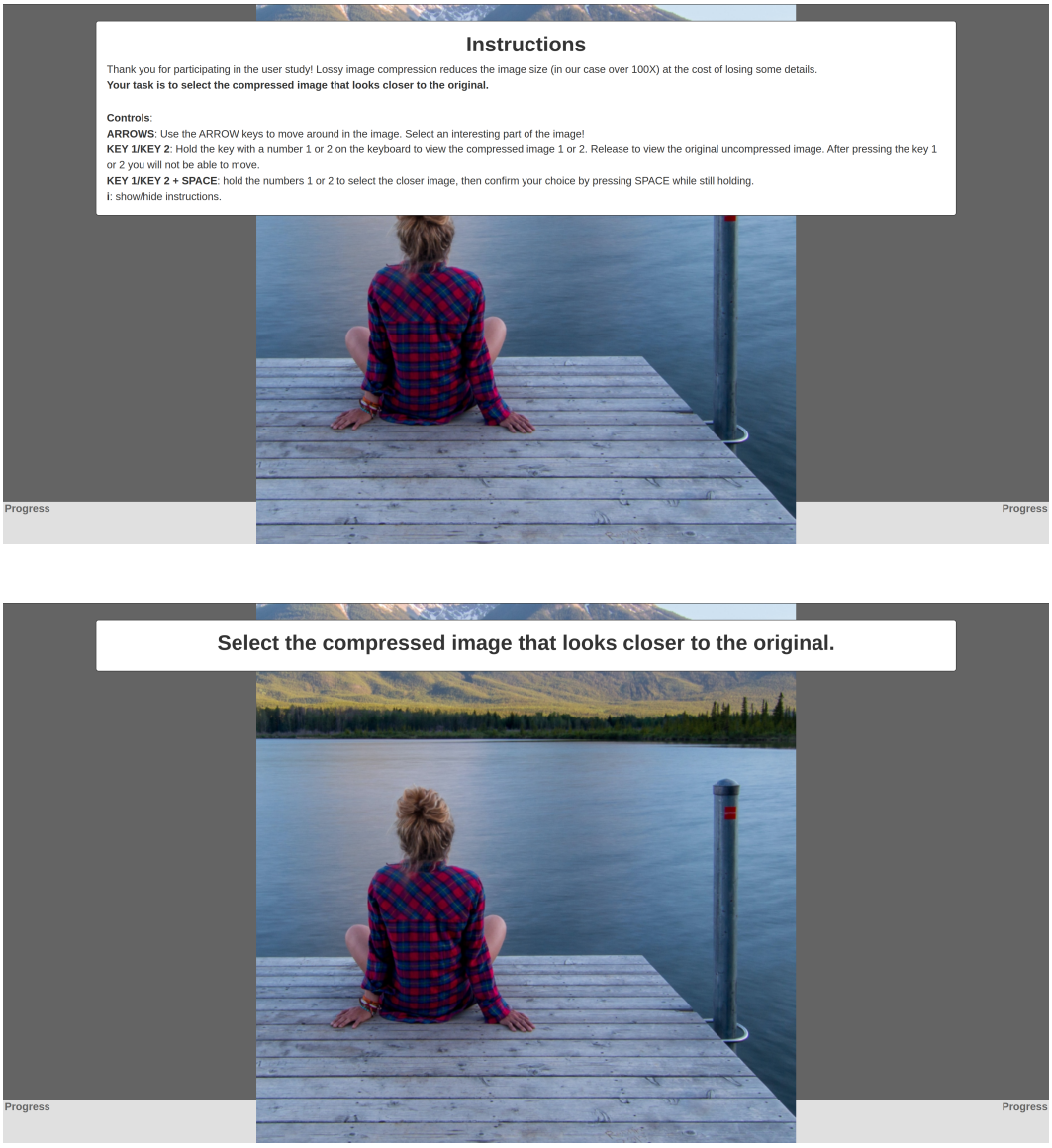

Figure 5: User study interface with instructions (top) and after closing the instructions (bottom).

A.1 Data collection

To align a compression model with human preferences, training data is essential. For training data, we compressed images using only two models: the rate-distortion optimized model (Hyperprior) and the rate-distortion-perception optimized model (HiFiC). The key difference between these two models is that Hyperprior focuses solely on rate-distortion without a perceptual loss term, while HiFiC incorporates perceptual losses. As a result, the labels reflect user preferences regarding the trade-off between exact reconstruction with artifacts and hallucinations in in-distribution samples. Label collection occurred at the 'low' bitrate defined in the HiFiC paper, yielding 5,408 comparisons for the DIV2K training and validation set. Labels were gathered through crowd-sourcing with volunteers, and these comparisons are available in the supplementary material.

A.2 Evaluation

During the evaluation phase for each comparison, a 786×78 crop of the original image is presented alongside two compressed images created using different compression models. To aggregate binary comparisons into model rankings, we use the Elo ranking system [27]. However, directly applying the Elo algorithm is challenging due to its permutation invariance. To overcome this, we implement bootstrapping, sampling 10,000 times with replacement and reporting statistics from the final score distribution.

B Further results

In Figures 6, 7 and 8, we provide a comprehensive comparison of our method and the baseline models using various computational metrics. BPG performs best in terms of PSNR, followed by Hyperprior, with perceptual compression models trailing behind. However, in terms of MS-SSIM, other methods catch up to BPG. Notably, GAN-based methods such as HiFiC, ConHa-fixed, and ConHa exhibit lower (thus better) LPIPS and FID scores. Our method shares similarities with HiFiC, exhibiting slightly higher PSNR, comparable MS-SSIM and LPIPS scores, and marginally worse FID scores. This suggests that our method achieves a balance between traditional fidelity metrics and perceptual

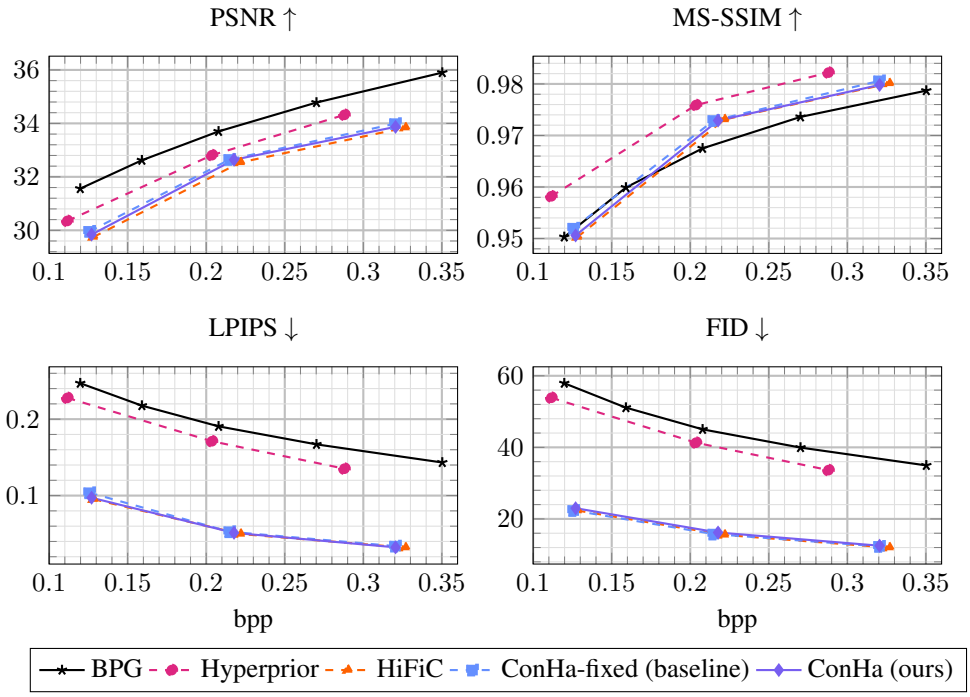

Figure 6: Rate-distortion and -perception curves on CLIC 2024 image test set. Arrows indicate whether higher (↑) or lower (↓) values are preferable.

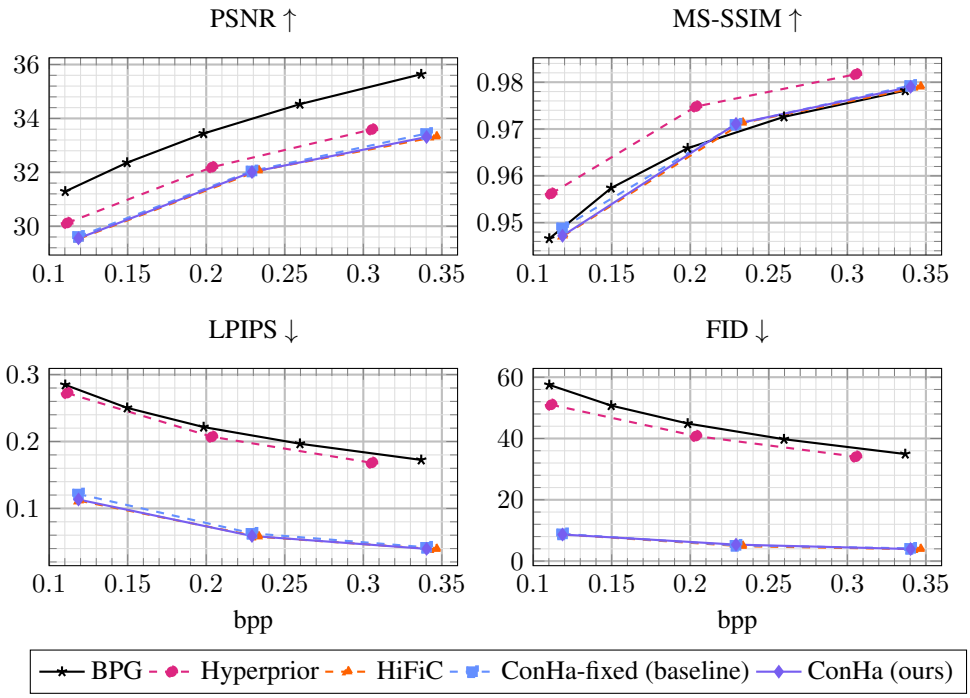

Figure 7: Rate-distortion and -perception curves on CLIC 2020 test set. Arrows indicate whether higher (↑) or lower (↓) values are preferable.

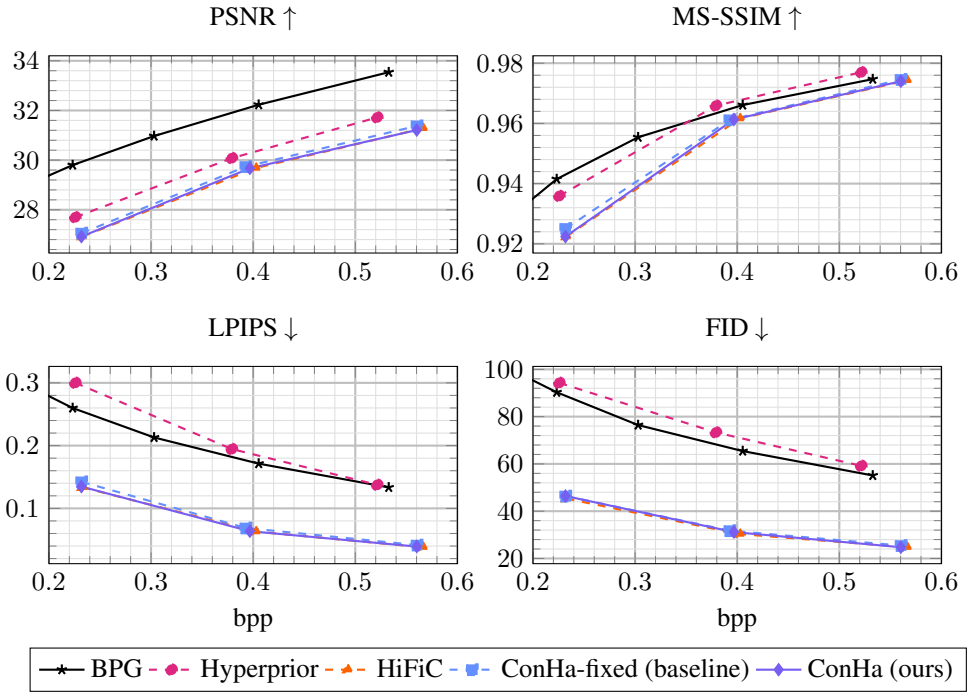

Figure 8: Rate-distortion and -perception curves on the Kodak dataset. Arrows indicate whether higher (↑) or lower (↓) values are preferable.

quality metrics. The shortcomings of the distortion metrics are evident, as they do not capture the performance gap between ConHa and other baselines, as indicated by the user study. It's worth noting that the FID score quantifies the disparity between the original and distorted distributions.

C Training details

Both models are trained on random 256 $\times$ 256 crops, the preference model on the DIV2K [24], while the image compressor on the Vimeo90k dataset [25]. Data collection for the preference model training is described in Appendix A.1.

Preference model:

The features are extracted from image x using a frozen pre-trained ResNet50, and a small head is trained on these features. For most images, the M_P model predicts weights in the range of [0.5, 0.6]. To push the weight w to closer to 0 or 1 during compression model training, we multiply the logit before the sigmoid function by 100. The architecture is illustrated in Figure 9. The configuration

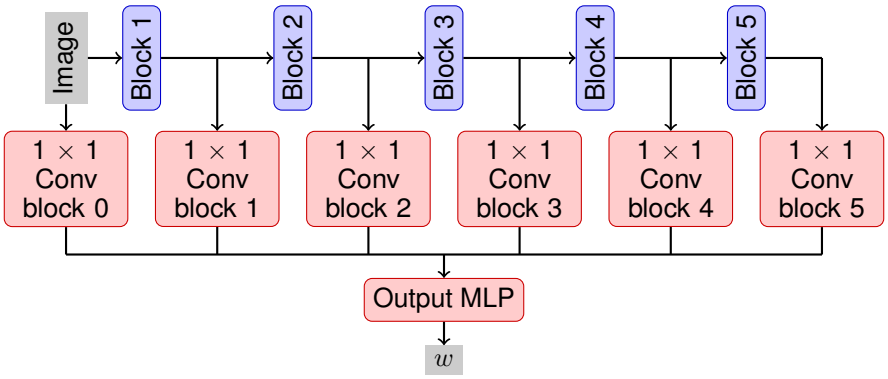

Figure 9: The hallucination-distribution preference model architecture. Features are extracted with a pre-trained frozen ResNet50, then fed into a block of convolutional layers with 1 $\times$ 1 kernel. The 1 $\times$ 1 convolutional block outputs are average pooled, then concatenated and fed into 2 feed-forward layers with ReLU activation in-between.

of layers in the 1$\times$1 Conv blocks is presented in Table 1. We collected labels at the 'low' bitrate, following the definition in the HiFiC paper, resulting in 5408 pairwise comparisons for the DIV2K training and validation dataset. Training took 30 epochs, employing a batch size of 8. The Adam optimizer is utilized with a learning rate of 1×10^{-4} and cosine learning rate annealing. We employ the binary cross-entropy loss function. Training the preference model takes less than an hour on an Nvidia A100 GPU.

layer	block 0	block 1	block 2	block 3	block 4	block 5
1	16	64	64	256	256	256
2	32	32	32	64	64	64
3	16	16	16	16	16	16

Table 1: The 1$\times$1 Conv block number of neurons per layer.

Compression model: We adopt the HiFiC approach [1] for our compression model training, with one notable adjustment regarding the training dataset. While HiFiC utilized their proprietary dataset, we employ the Vimeo90k dataset [25] due to accessibility constraints. As HiFiC we train the M&S Hyperprior model for 2M iterations and HiFiC, ConHa-fix, and ConHa follow a two-stage training. First, we train the model for 1 million iterations with the rate, MSE, and LPIPS losses. Then we continue training for another 1 million iterations with the GAN component incorporated, together the two-stage training takes 2 million steps as well. Notably, training our compression model for low, medium, or high bitrates is efficient, requiring less than 3 days on an A100 GPU.

D Limitations

Our work builds on top of rate-distortion and rate-distortion-perception optimized VAEs. We aim to strike a balance by automatically selecting the most suitable compression method based on image content. Human preferences between the two compression models may not vary significantly

depending on the compressed image. Our model closely resembles the rate-distortion-perception compression model, differing only in cases where the state-of-the-art underperforms the rate-distortion optimized model. Our improvements are noticeable only through user studies, as existing metrics struggle to capture such nuanced human preferences. Enhancing distortion metrics may enable them to detect subtle differences, but it may also render our work obsolete. Distortion metrics would then consider visible compression artifacts and whether the image appears aesthetically pleasing.