Quality-Preserving Extreme Image Compression: Using Interpretable Conditioning Inputs with Diffusion Models

Shayan Ali Hassan* (✉), Danish Humair*, Ihsan Ayyub Qazi, and Zafar Ayyub Qazi

Department of Computer Science, Lahore University of Management Sciences
{25100165,25100183,ihsan.qazi,zafar.qazi}@lums.edu.pk

Abstract. Diffusion models have revolutionized image synthesis, but their potential for image compression remains underexplored. We introduce PLIC (Pseudo-Lossy Image Compression), a compression framework leveraging diffusion models and conditioning inputs to achieve high compression ratios while maintaining strong perceptual similarity and superior image quality. Unlike traditional neural compressors using abstract latent representations, our approach uses interpretable conditioning inputs (text prompts, canny edges, color palettes) to guide diffusion-based image reconstruction. Grounded in rate-distortion-perception theory, PLIC prioritizes minimizing bitrate and distortions over pixel-perfect reconstruction, allowing diffusion models to fill in plausible details during decompression which still results in high perceptual similarity. Evaluating on 490 real-world images, we demonstrate superior compression ratios (0.004 bits per pixel and 0.197 bits per pixel on average) while maintaining excellent image quality (mean BRISQUE=23.36, mean CPBD=0.60) and high perceptual similarity. Our approach scales effectively with increasing image resolution, with compression advantages growing at the most common image resolutions. We analyze practical implications including benefits for internet affordability, archival storage, and deployment considerations. Our project code can be found at: `https://github.com/PseudoLossy/PLIC`.

Keywords: Image Compression · Diffusion Models · Conditioning Inputs.

1 Introduction

Image compression is widely used for reducing the data footprint of images while maintaining acceptable visual quality. Without image compression, many digital applications such as web browsing, video streaming, and cloud storage would not be possible at scale. Traditional lossless image compression algorithms, such as those used by PNG, can achieve compression ratios of up to 3:1 (66%

* These authors contributed equally to this work.

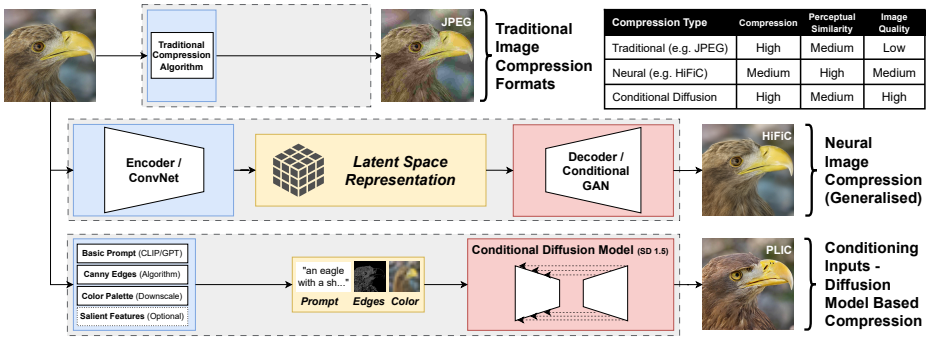

Fig. 1. Overview of how using conditioning inputs in tandem with diffusion models to compress images (our methodology) differs from traditional codecs and neural image compressors. Blue regions indicate compression steps, while red regions indicate decompression steps, while the width of each region provides an estimate of the computational demands. Notice the symmetrical and asymmetrical computational demand of neural compressors and our method respectively.

compression), while lossy formats like JPEG and WebP can reach up to 20:1 (95% compression) [11]. However, at such high compression levels, image quality often degrades significantly due to severe compression artifacts and distortions.

With the rise in popularity of deep learning models during the last decade, neural network architectures have been extensively applied to image compression tasks, achieving state-of-the-art performance. Generally, neural image compressors involve extracting high-level features from an image using a network and transforming them into a latent space representation. This representation achieves compression by exploiting spatial and semantic redundancies to distill only the most important information. The latent space representation is then provided as input to another network in order to obtain a reconstruction of the original image. Leveraging image features in this manner to reconstruct images has been shown to produce images with fewer visual distortions and greater perceptual fidelity compared to algorithmic codecs [26].

More recently, denoising probabilistic diffusion models such as Stable Diffusion have shown remarkable performance in generating high-quality, realistic images [9,15]. Architectures such as ControlNet [43] use "conditioning inputs" to control diffusion model outputs by conditioning them to adhere to structural guides. These inputs inherently encode perceptually important information about an image, and unlike abstract latent space representations, conditioning inputs explicitly extract interpretable spatial information such as color, structure, and depth, suggesting their potential utility beyond mere generation control.

We propose to re-imagine these conditioning inputs as components of a novel compression paradigm. We term our framework as PLIC (Pseudo-Lossy Image Compression). PLIC is grounded in rate-distortion-perception theory, strategically selecting ControlNet conditioning inputs that minimize bitrate and min-

imize distortions (maximizing quality), deliberately deprioritizing pixel-perfect reconstruction. This allows diffusion models to leverage their learned priors to "fill in" perceptually plausible details during reconstruction, effectively outsourcing part of the reconstruction process to the model's understanding of visual reality. This can result in exceptional compression ratios and higher image quality while still maintaining high perceptual similarity, a significant departure from the predominant focus on latent space representations in current compression literature. The differences between PLIC and traditional image compression formats and neural image compressors are illustrated in Fig. 1.

Through extensive experimentation on a carefully curated dataset of 490 real-world images from the top 1,000 globally visited domains, we demonstrate that using just three interpretable conditioning inputs (text prompts, canny edges, and color palettes) is sufficient to outperform neural compression baselines. Our evaluation includes three key experiments: (1) assessing the perceptual benefits of increasing the number of conditioning inputs, (2) analyzing image quality, perceptual similarity, and compression ratios against existing approaches, and (3) evaluating the feasibility of our framework with current technology. We evaluate our method across diverse image types and various resolutions (333×687 to 4032×2030), revealing that compression advantages scale better at common image sizes, despite computational overhead during decoding.

Taken together, we make three key contributions in this work:

- We introduce a novel compression framework grounded in rate-distortion-perception theory, which uses conditioning inputs to deliberately prioritize image quality and compression efficiency over pixel-perfect reconstruction.
- We demonstrate that our conditioning-inputs-based compression method, using three conditioning inputs (text prompts, canny edges, and color palettes) achieves exceptional *compression ratios* of 0.004 bits per pixel at minimum and 0.197 bits per pixel on average while having the best average *image quality* (mean BRISQUE=23.36, mean CPBD=0.60). While PLIC currently incurs higher *computation times*, it produces reconstructions with high *perceptual similarity*. Through extensive experiments against a neural framework that maximizes perceptual similarity and a neural framework that maximizes compression, we show our approach outperforms most previous baselines despite its significantly smaller data footprint.
- We provide a detailed analysis of the method's practical implications, including its scalability benefits across common image resolutions, its potential to increase internet affordability without degrading user experience, advantages for long-term archival storage, and considerations regarding computational requirements and ease of deployment.

2 Related Work

2.1 Neural Image Compression

Most traditional image compression techniques based on neural networks employ some form of either a variational autoencoder (VAE) [19] or generative adver-

sarial network (GAN) [12]. VAEs use a predefined network known as an encoder to transform the input, in this case an image, into a probabilistic latent space. This distribution of the image within the latent space is the compressed form, and a predefined decoder is used to transform the latent space distribution back into the input space. GANs utilize a generator and discriminator network to generate new data indistinguishable from the original input data's distribution. Since both architectures try to reconstruct an image from a smaller latent representation, there is some loss in information during the encoding and decoding steps, thus both suffer from the rate-distortion-perception tradeoff [8,5]. The rate-distortion-perception tradeoff states that in low bitrate contexts, such as compression, minimizing the distortions in images will lead to less perceptually pleasing images due to noise and other factors, and vice-versa. VAEs are known to induce blurriness in images for this reason [45].

There have been many proposed improvements and modifications to such neural networks in order to tailor them for image compression with higher realism, such as using less computationally expensive decoding activations in VAEs [40,37], semantically decoupling an image into multiple independent regions before encoding them [10], using Conditional-GANs trained on labelled data instead [25], and using a text encoder to inform the image encoder which details are perceptually the most important [21].

2.2 Diffusion Models for Image Compression

Diffusion models, also referred to as diffusion denoising probabilistic models (DDPMs) [15], have gained prominence as a powerful class of generative models, known for their high-quality image synthesis. Such models rely on the denoising autoencoder, which is repeatedly sampled while supervised by input features (such as text features in the case of text-to-image diffusion models) in order to incorporate random noise into an image, allowing for iterative generation of a high quality image. Latent diffusion models which shift the diffusion process to a lower-dimensional latent space have helped to substantially reduce computational demands [34]. Furthermore, ControlNet [43]; a neural network architecture designed to add spatial conditioning controls to large, pre-trained text-to-image diffusion models allows for even greater fidelity and controllability in generations. Using image conditioning inputs that are easily understood by humans such as edge maps, depth maps, segmentation masks, among others, users are able to modify the image structure and style as needed.

Unsurprisingly, using diffusion models to compress images is emerging as an area of research interest due to their learned knowledge about both high-level and low-level visual concepts, allowing them to reconstruct image details at higher fidelity and perceptual quality for a given bitrate [46]. However, these models are known for their generative diversity [1,16] therefore in the context of image compression where the generated image must obey a ground truth, recent work has focused on controlling the output of these models while optimizing them for compression. This includes introducing additional latent variables to guide the denoising process [39], removing redundant processes in the denoising steps to

increase performance [33] and using short text embedding generated from the original image itself instead of a prompt to generate the image [30]. While all of these methods score well on perceptual metrics, their use of latent vectors as the condition for the diffusion model prevents them from outperforming existing methods in terms of compression. Furthermore, the easily interpretable conditioning inputs typically employed in ControlNet are not explored as a means of compression. While [22] investigates the use of sketches as a conditioning input to preserve structural information for compression, and [7] uses simple image descriptions in conjunction with latent image representations, the similarity of the reconstructed images is lacking with respect to the originals, especially for perceptually important details.

To the best of our knowledge, this is the first work that thoroughly investigates the potential of using simple, non-vector conditioning inputs as a means of compressing images.

3 Methodology

Supported by the rate-distortion-perception theory, we propose to choose conditioning inputs that aim to minimize bitrate and maximize perceptual similarity with the original image, thus forgoing minimization of pixel-level differences and spatial distortions [4]. While this would render reference-based pixel-wise metrics unsuitable for evaluation, this is preferable as most conditioning inputs supported by ControlNet versions of diffusion models capture a facet of the most perceptually meaningful information only, tending to ignore finer details at the pixel-level. Furthermore, diffusion models will be able to judiciously "fill in" the gaps at the pixel-level, such as texture, due to their learned knowledge about the world, allowing for the reconstructed images to still be of higher quality even if there are pixel-level differences. With the prioritized optimizations in mind, we extract the semantic, structural and color information from the original image via the following conditioning inputs:

1. Text prompt: The prompt serves as a description of the image and is sufficient to provide semantic information. We opt to use the GPT-4-vision API to generate the prompt from the original images, as this allows us to tweak human-friendly zero-shot instructions which can help prevent misuse as well as leverage a Large Language Model (LLM) to only describe perceptually and contextually important information.

2. Canny Edges: Canny edges can provide the structural information [6], ensuring that reconstructions have the same composition as the original image. This approach strikes a balance between minimal bitrate and improved accuracy, as the canny edgemaps are compact due to being monochrome.

3. Color Palettes: A color palette further contextualizes the information provided to the diffusion model by providing general information of how the colors were distributed in the image, ensuring accurate color replications. It can be obtained by downsizing an image to a small resolution such as 32×32.

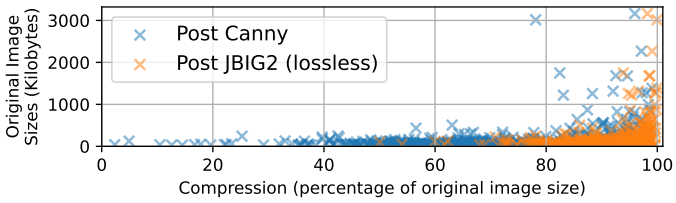

Fig. 2. Using JBIG2 on canny edges provides significant lossless compression.

3.1 Additional Optimizations

In order to increase perceptual similarity in a compression maximizing manner, we propose three further optimizations to the aforementioned conditioning inputs:

Due to edge detection, most pixels in a Canny edge bitmap have the same value, representing negative space between thin, edge outlines, allowing further optimization. We exploit this by using lossless JBIG2, the industry standard compression algorithm for bi-level images, typically used in fax machines and black-and-white PDFs. JBIG2 outperforms other algorithms by reducing bi-level images by a factor of 2-5 [17]. Testing on 490 images scraped from the Web (Sec. 4), we see that JBIG2 provides significant lossless compression, allowing the canny edges to be up to 99.95% and on average 90% smaller than the original image. The results are presented in Fig. 2.

The color map is encoded via WebP, which provides significantly reduced compression artifacting compared to JPEG [11], ensuring that the colors extracted from the original image are provided to the diffusion model accurately.

Lastly, we propose using segmentation masks to preserve finer perceptual details across compression and reconstruction. Stable Diffusion Inpainting allows using a black and white mask to indicate which sections of an image should be recreated, and which sections should be left untouched, thus allowing for the preservation of critical, *salient features*, such as faces, small text and small logos [42]. The cropped salient features are stored unmodified as part of the compressed form, hence the total compression becomes indirectly proportional to the area of salient features in the original image. However, since only small, important details are at risk of being distorted and thus marked as salient [33,22], compression is expected to still be significant. Salient features can be masked very quickly and automatically by leveraging Meta's Segment Anything Model [20] to identify various segments in an image and using Grounding DINO [23] to identify segments with small, salient features.

Our approach extracts only the most high-level and perceptually important features as conditioning inputs to serve as the compressed state of an image. This method guarantees the preservation of information in these conditioning inputs during reconstruction while allowing a diffusion model to 'fill in' missing details autonomously–a technique we term as *'Pseudo-Lossy' Image Compression*. Therefore, for the sake of brevity we shall refer to our proposed methodology

of compressing images via conditioned diffusion models as **PLIC**. Fig. 1 provides an overview of the conditioning inputs in PLIC and how it differs from traditional methods.

4 Experiments & Results

To evaluate the efficacy of PLIC in real-world scenarios, we scraped images from the top 1000 globally visited domains using the Google Chrome UX Report (CRUX) lists [36] and manually removed pages with inappropriate content, resulting in a final set of 600 pages. Given the imbalance in image quantity across websites, each website was categorized into one of seven main categories derived from Cloudflare's domain categorizations [35]: E-Commerce, Informational, Business / Company, News, Social Media, Video Streaming, and other.

From each category, we randomly sampled 70 images, resulting in a dataset of 490 images. [1] A minimum resolution of 512×256 in either orientation was set because (i) lower-resolution images lack the necessary detail and structure needed by current text-to-image models to generate high-quality, coherent outputs [32] and (ii) such images do not provide much compression due to their already small size. Thus, in our dataset, the resolution of collected images ranged from 333×687 to 4032×2030.

We conduct 3 different types of experiments:

- Evaluating the perceptual benefits of increasing conditioning inputs.
- Detailed analysis of image quality, perceptual similarity and compression ratios against existing approaches.
- Exploring how the generation costs scale up in two different use cases.

4.1 Ablation Study of Conditioning Inputs

We validate the choice of conditioning inputs by compressing all images in the dataset to their conditioning inputs, reconstructing them and evaluating the perceptual similarity of the reconstructions with respect to the original images. We repeat this experiment 4 times, each time adding another conditioning input to evaluate the contribution of each to the perceptual similarity, (prompts only, prompts + canny edges, prompts + canny edges + color maps, prompts + canny edges + color maps + salient feature preservation enabled). We used Stable Diffusion 1.5 (SD1.5) as the diffusion model in conjunction with two ControlNet pipelines to condition the input, where one conditioned the image generation based on structure and color and the other used inpainting to generate all parts of the image other than the selected region.

[1] Images with transparent pixels were excluded to prevent processing inconsistencies. Many image models are optimized for RGB channels and may mishandle the alpha channel for transparency. This ensures consistent data processing and avoids potential artifacts during model training and inference.

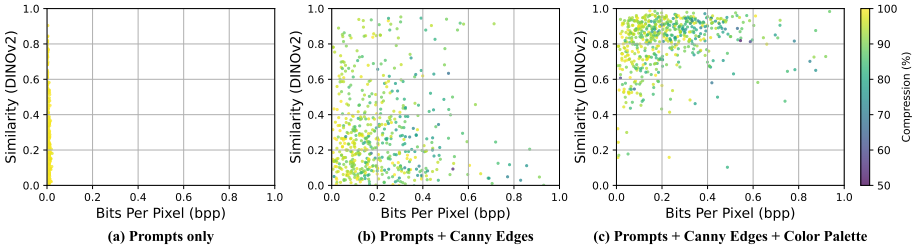

Fig. 3. Tradeoff between bpp and perceptual similarity as conditioning inputs increase.

EXP	DINOv2			Compression (%)			Compression (bpp)		
	Median	Min	Max	**Median**	Min	Max	**Median**	Min	Max
Prompt	**0.16**	0.00	0.90	**99.78**	84.00	99.998	**0.0004**	0.0001	0.02
+ Canny	**0.24**	0.00	0.95	**91.14**	54.00	99.95	**0.17**	0.0001	0.93
+ Color	**0.85**	0.10	0.99	**90.47**	52.91	99.94	**0.18**	0.01	0.94
+ Salient	**0.89**	0.58	0.98	**77.85**	0.00	96.82	**0.33**	0.05	1.45

Table 1. Summary statistics of the ablation study.

The extent of compression is measured in two ways; (i) via comparing the size of the original image against the compressed state (i.e. the sum of prompt size, canny edges size after JBIG2 compression, color palette size, segmentation mask size and salient features size) and (ii) bits per pixel (bpp). For the perceptual similarity evaluation metric, we use DINOv2, a self-supervised Vision Transformer that achieves state-of-the-art performance on many computer vision tasks [29]. DINOv2 creates visual feature embeddings which are well-suited to capture important semantic information about images. Furthermore, since DINOv2 is self-supervised, it has learned visual features that can generalize across various image distributions, even outside of its training set. The cosine similarity between the embeddings of the original and the reconstructed image can be used as an accurate measure of perceptual similarity [44]. The results are presented in Fig. 3 and the summary statistics are provided in Tab. 1.

As we add more conditioning inputs, we see weaker compression ratios as the median bpp increases from 0.0004 to 0.33. However, looking at compression as the percentage reduction in size instead we see that the images are still being heavily compressed in all cases. Furthermore, adding more conditioning inputs results in a consistent increase in perceptual similarity score, increasing from 0.16 as a baseline to 0.89, referencing the rate-perception tradeoff. Due to salient preservation targeting small regions from an image, the perceptual similarity score does not improve much between the salient feature preservation test and the previous test. Fig. 4 visually shows the effects of our salient feature step on critical details, as well as the effects of increasing conditioning inputs generally.

Fig. 4. Perceptual gains of increasing the number of conditioning inputs.

4.2 Detailed Image Analysis

Next, we evaluate PLIC against two neural image compression frameworks:

- **HiFiC [25]:** HiFiC is a popular conditional-GAN-based image compressor that achieves state-of-the-art results in many image compression on various metrics such as PSNR (Peak signal-to-noise ratio), MS-SSIM (Multiscale - structural similarity index measure) and FID (Fréchet inception distance) [14] at very low bitrates. In order to ensure a fair comparison, we aim to keep the compression ratios between PLIC and HiFiC as close as possible hence we specifically use HiFiC-Low as it achieves the highest compression ratios as compared to HiFiC-Medium or HiFiC-High, at the cost of some loss in similarity.
- **TACO [21]:** TACO is an image compression framework which utilizes a text encoding to guide a diffusion encoder for creating a latent representation of an image, instead of directly using a text prompt to guide the encoder instead. It also achieves state-of-the-art results in LPIPS (Learned Perceptual Image Patch Similarity) [44], a widely used perceptual similarity metric. Similarly to HiFiC, we use TACO with the hyper-parameter $\lambda = 0.015$ which results in the most aggressive compression.

To ensure a more robust evaluation, we use PLIC with not just SD1.5 but also Flux.1, a recent transformer based text-to-image diffusion model that has gained considerable popularity, in order to draw any insights between consistent behavior in diffusion models (we did not provide color palettes as a conditioning input to Flux as it was recently released at the time of experimentation and a well-trained canny + color conditioned version of the model did not exist yet). Furthermore, instead of just using DINOv2 as the perceptual similarity metric, we also evaluate all images using the LPIPS metric, due to its high alignment with human ratings [44].

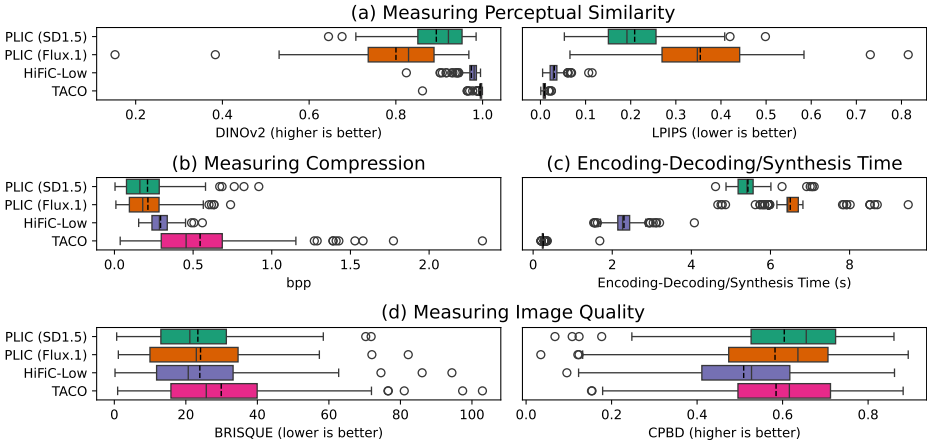

Fig. 5. Comparison of perceptual similarity, compression strength, computational time and image quality between PLIC (SD1.5, Flux.1), HiFiC-Low and TACO.

As mentioned previously in Sec. 3, while our methodology produces largely perceptually similar reconstructions, they are different at the pixel-level since fine, noisy details are not captured at any point. Since PLIC is not constraint to exactly reconstruct pixel-level textures, the reconstructed images do not suffer from compression artifacts / blockiness / blurriness as the diffusion model makes no attempt to reconstruct exact details from compressed data and is free to generate the output from it's own learned distributions. To measure this increase in *image quality*, a no-reference (NR) image quality assessment (IQA) model is required since providing the original image as a reference would lead to poor scores due to the pixel-wise differences being interpreted as reconstruction loss. To measure reconstruction quality, we use BRISQUE (blind / referenceless image spatial quality evaluator) [27], which uses luminance coefficients to quantify possible losses of 'naturalness' in the image due to the presence of compression distortions, and CPBD (cumulative probability of blur detection) [28], which uses a probabilistic model to measure levels of blur at each edge in an image.

We evaluate all methodologies against each other across 6 dimensions: perceptual similarity with DINOv2 and LPIPS, compression, encoding–decoding / synthesis times, and image quality with BRISQUE and CPBD, on a 150 image subset of our dataset for each sub-experiment. The results are provided in Fig. 5. While HiFiC-Low and TACO achieve better perceptual similarity scores, the scores for PLIC with SD1.5 are still reasonably good, with an average of 0.90 DINOv2 score and 0.21 LPIPS score (perceptual similarity for PLIC with Flux is slightly lower as expected due to the lack of color palettes). However, the compression ratios achieved by PLIC are noticeably better, with an average bpp of 0.20 for SD1.5, which is lower than the $25th$ percentile bpp for both HiFiC and TACO. As for image quality, the BRISQUE score distributions show average scores of 23.36, 23.91 and 29.90 for SD1.5, HiFiC and TACO respectively, indi-

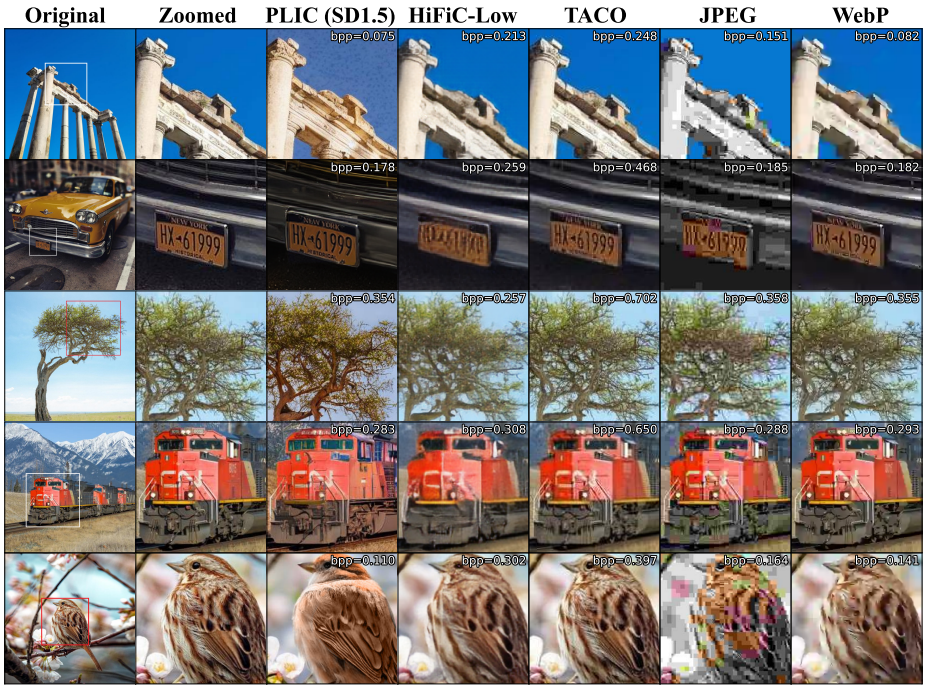

Fig. 6. Visual comparison showing the superiority of PLIC and conditioning inputs + diffusion model compression in general to reconstruct higher quality and fidelity of images that are still perceptually similar in the extremely low bitrate regime. JPEG & WebP images were compressed with maximum strength in order to bring the compression extents as close as possible to PLIC. HiFiC and TACO were also used at maximum compression strength as previously mentioned.

cating TACO has more spatial distortions than the other two. Meanwhile, the CPBD score distributions show average scores of 0.60, 0.51 and 0.58 for SD1.5, HiFiC and TACO respectively, indicating HiFiC images to be more blurry compared to the other two. While HiFiC images suffer from blurriness and TACO images suffer from distortions and noise, PLIC with SD1.5 minimizes both distortions and bluriness, performing the best when evaluated by the BRISQUE and CPBD metrics. Hence, PLIC demonstrates the best image quality and visually pleasing reconstructions (given that the viewer is fine with a non-exact pixel-level yet still perceptually similar overall reconstruction). This becomes evident in Fig. 6 too, as the PLIC reconstructions look much sharper and defined, free of compression artifacts. In the computational time analysis, a clear weakness of PLIC is shown as HiFiC and TACO are much faster at encoding and decoding images, while the diffusion models are slower due to their highly iterative nature.

Lastly for the detailed image analysis, we investigate how the average bpp changes as image dimensions / resolution changes. High-resolution images (at

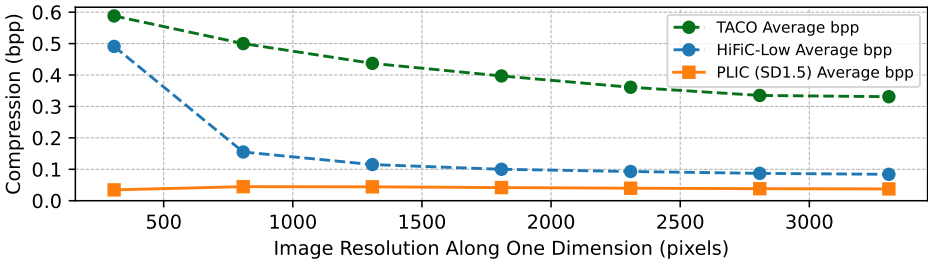

Fig. 7. Comparison of average bpp after compression across image resolutions.

least 3000 pixels along any dimension) were sampled from our dataset, downsized versions of various resolutions were created for each and each version was compressed and decompressed using PLIC with SD1.5, HiFiC and TACO. The results are shown in Fig. 7. We find that the PLIC average bpp is not dependent upon the original / reconstructed image size. This makes sense as prompts are negligible in size and generated according to what the image is depicting thus they are similar irrespective of resolution. Color palettes are always obtained by reducing the image to a constant 32×32 resolution, thus they have negligent impact on the bpp. Canny edges meanwhile are the same size as the original dimensions and are thus the amount of bits needed to encode the information for one pixel remains proportionately the same, especially after JBIG2 compression. Only varying level of salient features cause small deviations for PLIC. Overall, for image dimensions of about 500 to 1500 pixels especially, PLIC scales better at compressing images when compared to a GAN or diffusion-encoder based compressors such as HiFiC or TACO, which proportionately require a larger latent vectors to represent an image at lower resolutions.

4.3 Cost Analysis Across File Sizes and Storage Duration

In this section, we evaluate the overall cost implications of our framework by examining two distinct scenarios: (i) on-demand image delivery, where both network egress and GPU decoding costs matter, and (ii) long-term archival storage, where storage fees and time before access dominate.

On-Demand Transfer & Decoding: Fig. 8a shows the ratio of original transfer cost to our method's total cost (i.e., PLIC transfer *plus* decoding). A ratio above 1 indicates that using PLIC is cheaper. We bin images by their original file size, and apply a common egress fee ($0.09/GB on AWS [2]), as well as a GPU rental rate ($1/hour for an A100 on Vast.ai [38]). Notably, our GPU VRAM was heavily underutilized and rental prices include a significant premium, so the per-image decoding cost could be even lower when amortized over larger batches or shared GPU usage. Across bins, we observe higher cost savings (ratio up to $\times$4–5) as image size increases. This stems from the fact that our compressed conditioning data does not grow significantly with resolution, as seen in Fig. 7. Furthermore, larger image files in our dataset were likely to be

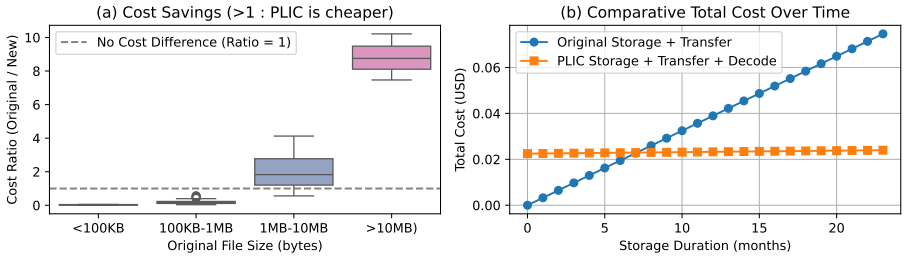

Fig. 8. (a) *Immediate cost ratio* of storing/transferring original images vs. using PLIC, binned by file size. Ratios above 1.0 indicate that PLIC is cheaper. (b) *Cumulative cost over time* for long-term storage, comparing original images to PLIC's reduced representations.

less optimized beforehand, granting our framework more competitive savings, whereas smaller image files were already heavily compressed and showed a ratio lower than 1, but at the cost of much lower fidelity.

Long-Term Storage & Decoding: In many real-world scenarios, images remain stored for extended periods, making the *storage footprint* at least as important as immediate transfer and decoding cost. Fig. 8b compares the cumulative cost of retaining and eventually transferring the original images versus storing PLIC's smaller conditioning inputs and decoding them on-demand. We apply a common storage rental rate ($0.023/GB/month for Amazon S3 [3]), and assume the same egress fees and GPU rental as before. Over time, the modest decode overhead is overshadowed by significant storage savings, especially as the image set grows or is retained for periods longer than 6 to 12 months. Therefore, PLIC provides substantial operational benefits for archival or on-demand use cases (high-write, low-read scenarios), even under conservative GPU cost assumptions, due to it's assymetrical computational demands and the compression step being quite inexpensive as the diffusion model remains un-involved.

5 Discussion

In this section, we would like to discuss other aspects, contributions and limitations of our method that were not explored in the previous sections:

Bandwidth Savings & Internet Affordability: Webpage sizes have increased by roughly 13 times over the last decade, in large part due to the use of more and higher quality images [13]. Simultaneously, 94 developing countries fail to meet the target for affordable broadband services due to Internet plan costs exceeding 2% of the monthly Gross National Income (GNI) per capita [18]. Given the extreme levels of compression, PLIC-based image compression can be a viable pathway towards increasing bandwidth savings for end users, thus making the internet much more affordable. PLIC can also enhance the user experience during web browsing given that in most cases, image quality is enhanced com-

pared to traditional image codecs and neural compression frameworks, while still preserving high perceptual similarity with respect to the original images. Due to the combination of high image quality and perceptual similarity, but lower pixel-level similarity, we expect such compression to excel in 'semantic imagery contexts', where images serve primarily to evoke or illustrate broad concepts, rather than to provide detailed visual information. Examples of such contexts include news articles, educational materials, and stock photos.

High-Write, Low-Read Scenarios: PLIC is attractive for workloads where images are written or uploaded frequently but accessed only sporadically. Backups, archives, long-tail media libraries, and even "cold" portions of personal photo galleries fall into this category. In such settings, traditional storage tiers often force a painful trade-off: either keep full-fidelity images and incur recurring capacity fees, or delete them outright to reclaim space. PLIC offers a third option. Because encoding is lightweight while decoding is heavyweight, write time remains minimal yet stored footprint is reduced significantly compared to other compression methods. Reads remain possible albeit with a compute penalty. This "compress instead of delete" path allows users to retain access to rarely viewed or lower-value images that would otherwise be purged, deferring costly storage upgrades for cloud providers and individual users alike.

Client-Side Reconstruction: Currently, due to the computational demands of diffusion models, we assume the image to be transmitted over a network in it's compressed state and being reconstructed at the edge, such as through a content delivery network (CDN) server, where adequate compute is available. While this can help to reduce costs for CDN providers and generally decrease ingress bandwidth over the internet, end-users can not currently receive the affordability benefits. However, recent trends indicate running deep neural networks on client devices such as smartphones and laptops may soon become feasible. Apple has recently made improvements to iPhone hardware allowing them to run language models such as OpenELM [24]. Additionally, one-step image generators such as DMD2 [41], based on knowledge distillation, allow high-quality images to be synthesized at extremely low inference times.

Ease of Deployment: As long as the image model supports text, canny, color inputs and inpainting, similar extents of compression and image quality can be achieved regardless of the image model at the core. Thus PLIC offers flexibility for a wide range of tasks where one model may be better than another. Furthermore, unlike existing neural compression frameworks, PLIC requires no additional model training to use whatsoever. We were able to experiment with the proposed methodology immediately because versions of popular image models that accept conditioning inputs will already be trained by community contributors for artistic purposes and use in a variety of image generation tasks. This 'out-of-the-box' approach can allow developers to immediately deploy PLIC, unlike GAN and encoder-based methodologies such as HiFiC and TACO which must be extensively trained first, are only limited to the specific task of image compression, and must be re-trained again to incorporate improvements in neural network understanding.

Ethical Considerations: PLIC based compression, if used properly, is also able to account for any societal biases in diffusion models, as evident by the example in Fig. 4. As more conditioning inputs are provided, the model gains context and a better understanding of the features such as in this case, the person's skin tone, refining the output to match the original. Conversely, this also means a malicious actor could falsify canny edges / color palettes or add misleading words to prompts to intentionally produce inaccurate or negatively biased reconstructions. Past instances show that people are understandably sensitive to offensive content generated by image models, such as when Google's Gemini model in an attempt to curb societal biases ended up generating images with obvious historical inaccuracies, causing controversy [31].

6 Conclusion

In this paper, we investigated a novel image compression approach using conditioning inputs for diffusion models, providing a compelling alternative to traditional neural compression methods. By leveraging text prompts, canny edges, color palettes, and salient feature preservation, our method achieves extreme compression while maintaining high perceptual similarity, effectively navigating the rate-distortion-perception tradeoff. Experiments on real-world images demonstrate superior compression ratios and image quality, with scalability benefits at common resolutions. While decoding remains computationally intensive, the flexibility and ease of deployment make this approach promising for future compression solutions, with potential applications in bandwidth optimization, energy / storage efficiency, and real-time image reconstruction.

Acknowledgments. We thank Rania Yakub Khan, Zaeem Khan and Ayesha Mirza for assisting with the experiments, verifying our results and for providing invaluable feedback.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aithal, S.K., Maini, P., Lipton, Z., Kolter, J.Z.: Understanding hallucinations in diffusion models through mode interpolation. Advances in Neural Information Processing Systems **37**, 134614–134644 (2025)
2. Amazon Web Services: Amazon EC2 On-Demand Pricing. `https://aws.amazon.com/ec2/pricing/on-demand/`, last accessed 2025/03/17
3. Amazon Web Services: Amazon S3 Pricing. `https://aws.amazon.com/s3/pricing/`, last accessed 2025/03/17
4. Bachard, T., Bordin, T., Maugey, T.: Coclico: Extremely low bitrate image compression based on clip semantic and tiny color map. In: 2024 Picture Coding Symposium (PCS). pp. 1–5. IEEE (2024)

5. Blau, Y., Michaeli, T.: Rethinking lossy compression: The rate-distortion-perception tradeoff. In: International Conference on Machine Learning. pp. 675–685. PMLR (2019)
6. Canny, J.: A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence **PAMI-8**(6), 679–698 (1986)
7. Careil, M., Muckley, M.J., Verbeek, J., Lathuilière, S.: Towards image compression with perfect realism at ultra-low bitrates. In: The Twelfth International Conference on Learning Representations (2023)
8. Cover, T.M.: Elements of information theory. John Wiley & Sons (1999)
9. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10850–10869 (2023)
10. Feng, R., Gao, Y., Jin, X., Feng, R., Chen, Z.: Semantically structured image compression via irregular group-based decoupling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17237–17247 (2023)
11. Ginesu, G., Pintus, M., Giusto, D.D.: Objective assessment of the webp image coding algorithm. Signal processing: image communication **27**(8), 867–874 (2012)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Habib, R., Tanveer, S., Inam, A., Ahmed, H., Ali, A., Uzmi, Z.A., Qazi, Z.A., Qazi, I.A.: A framework for improving web affordability and inclusiveness. In: Proceedings of the ACM SIGCOMM 2023 Conference. pp. 592–607 (2023)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Hoogeboom, E., Agustsson, E., Mentzer, F., Versari, L., Toderici, G., Theis, L.: High-fidelity image compression with score-based generative models. arXiv preprint arXiv:2305.18231 (2023)
17. Howard, P.G., Kossentini, F., Martins, B., Forchhammer, S., Rucklidge, W.J.: The emerging jbig2 standard. IEEE Transactions on Circuits and Systems for Video Technology **8**(7), 838–848 (1998)
18. ITU: Measuring digital development: Ict price trends 2019 (2020)
19. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
21. Lee, H., Kim, M., Kim, J.H., Kim, S., Oh, D., Lee, J.: Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity. arXiv preprint arXiv:2403.02944 (2024)
22. Lei, E., Uslu, Y.B., Hassani, H., Bidokhti, S.S.: Text+ sketch: Image compression at ultra low rates. arXiv preprint arXiv:2307.01944 (2023)
23. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38–55. Springer (2024)
24. Mehta, S., Sekhavat, M.H., Cao, Q., Horton, M., Jin, Y., Sun, C., Mirzadeh, S.I., Najibi, M., Belenko, D., Zatloukal, P., et al.: Openelm: An efficient language model

family with open training and inference framework. In: Workshop on Efficient Systems for Foundation Models II@ ICML2024 (2024)
25. Mentzer, F., Toderici, G.D., Tschannen, M., Agustsson, E.: High-fidelity generative image compression. Advances in Neural Information Processing Systems **33** (2020)
26. Mishra, D., Singh, S.K., Singh, R.K.: Deep architectures for image compression: a critical review. Signal Processing **191**, 108346 (2022)
27. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)
28. Narvekar, N.D., Karam, L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Transactions on Image Processing **20**(9), 2678–2683 (2011)
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
30. Pan, Z., Zhou, X., Tian, H.: Extreme generative image compression by learning text embedding from diffusion models. arXiv preprint arXiv:2211.07793 (2022)
31. Raghavan, P.: Gemini image generation got it wrong. we'll do better (2024). URL https://blog. google/products/gemini/gemini-image-generation-issue (2024)
32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
33. Relic, L., Azevedo, R., Gross, M., Schroers, C.: Lossy image compression with foundation diffusion models. In: European Conference on Computer Vision. pp. 303–319. Springer (2024)
34. 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)
35. Ruth, K., Fass, A., Azose, J., Pearson, M., Thomas, E., Sadowski, C., Durumeric, Z.: A world wide view of browsing the world wide web. In: Proceedings of the 22nd ACM Internet Measurement Conference. pp. 317–336 (2022)
36. Ruth, K., Kumar, D., Wang, B., Valenta, L., Durumeric, Z.: Toppling top lists: Evaluating the accuracy of popular website lists. In: Proceedings of the 22nd ACM Internet Measurement Conference. pp. 374–387 (2022)
37. Tao, L., Gao, W., Li, G., Zhang, C.: Adanic: Towards practical neural image compression via dynamic transform routing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16879–16888 (2023)
38. Vast.ai: GPU Pricing for A100 SXM4. `https://cloud.vast.ai/?gpu_option=A100%20SXM4`, last accessed 2025/03/17
39. Yang, R., Mandt, S.: Lossy image compression with conditional diffusion models. Advances in Neural Information Processing Systems **36**, 64971–64995 (2023)
40. Yang, Y., Mandt, S.: Computationally-efficient neural image compression with shallow decoders. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 530–540 (2023)
41. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T.: Improved distribution matching distillation for fast image synthesis. In: NeurIPS (2024)
42. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023)

43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
44. 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)
45. Zhao, S., Song, J., Ermon, S.: Towards deeper understanding of variational autoencoding models. arXiv preprint arXiv:1702.08658 (2017)
46. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5729–5739 (2023)