VLIC: Vision-Language Models As Perceptual Judges for Human-Aligned Image Compression

Kyle Sargent[1,2], Ruiqi Gao[3], Philipp Henzler[2], Charles Herrmann[3], Aleksander Hołyński[3]
Li Fei-Fei[1], Jiajun Wu[1], Jason Y. Zhang[2]

[1]Stanford University, [2]Google Research, [3]Google DeepMind

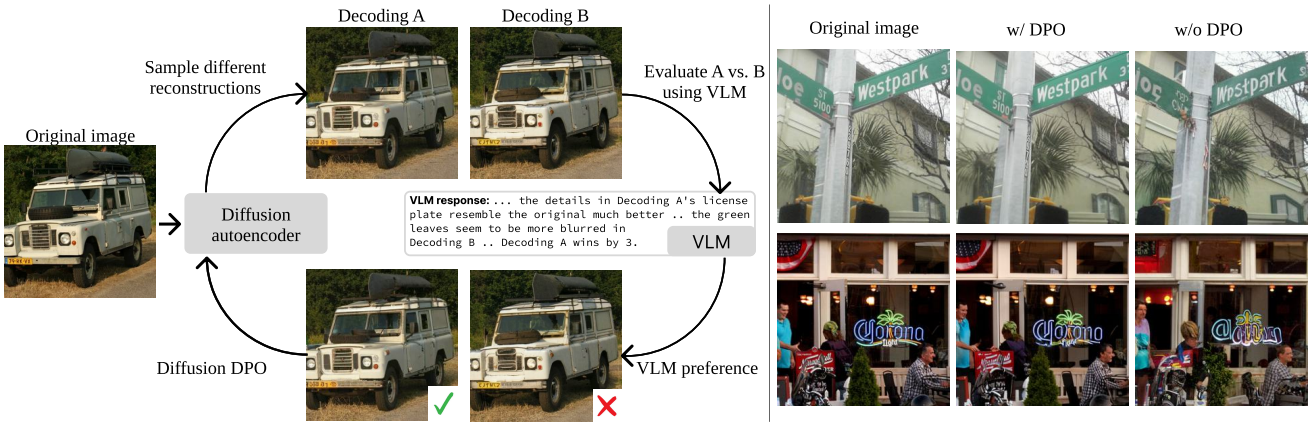

Figure 1. **Left.** We propose a new post-training technique for diffusion autoencoders which uses Vision-Language Models to judge different decodings of the same image and leverage these judgements to improve the autoencoder through Diffusion DPO. **Right.** Our method, VLIC, demonstrates substantial improvements in the overall reconstruction quality, as well as better alignment to human perception.

Abstract

*Evaluations of image compression performance which include human preferences have generally found that naive distortion functions such as MSE are insufficiently aligned to human perception. In order to align compression models to human perception, prior work has employed differentiable perceptual losses consisting of neural networks calibrated on large-scale datasets of human psycho-visual judgments. We show that, surprisingly, state-of-the-art vision-language models (VLMs) can replicate binary human two-alternative forced choice (2AFC) judgments zero-shot when asked to reason about the differences between pairs of images. Motivated to exploit the powerful zero-shot visual reasoning capabilities of VLMs, we propose **V**ision-**L**anguage Models for **I**mage **C**ompression (VLIC), a diffusion-based image compression system designed to be post-trained with binary VLM judgments. VLIC leverages existing techniques for diffusion model post-training with preferences, rather than distilling the VLM judgments into a separate perceptual loss network. We show that calibrating this system on VLM judgments produces competitive or state-of-the-art performance on human-aligned visual compression depending on the dataset, according to perceptual metrics and large-scale user studies. We additionally conduct an extensive analysis of the VLM-based reward design and training procedure and share important insights. More visuals are available on our website.*

1. Introduction

Compressing images and videos is necessary for storing and transmitting the rich sensory multimedia captured every day. This process inherently involves making trade-offs between compression rate (*i.e.* file size) and visual quality. Ideally, the assessment of visual quality should align well with human perception, prioritizing details to which humans are more sensitive. For instance, humans are sensitive to perturbations to faces and text, but not to high-entropy natural textures such as grass or fur.

Historically, visual quality has been assessed via classical metrics such as PSNR and SSIM [47]. However, these metrics are poorly aligned with human perception and often contradict human judgments of visual quality [12, 55]. To address this, prior work has focused on learning neural

Original | Ours: 0.196 bpp | HiFiC: 0.199 bpp | PerCo: 0.5 bpp

Original | Ours: 0.198 bpp | HiFiC: 0.287 bpp | PerCo: 0.12 bpp

Original | Ours: 0.203 bpp | HiFiC: 0.281 bpp | PerCo: 0.12 bpp

Figure 2. **Qualitative results on standard image compression datasets.** *Top*: We compare VLIC with HiFiC [31] and PO-ELIC [17] on a CLIC 2022 image [44] at various bits per pixel (bpp). *Bottom*: We compare our approach with HiFiC and PerCo on MS-COCO [26]. We find that our approach represents perceptually relevant fine details, faces, and textures more faithfully

network approximations to human judgment of visual similarity, with varying approaches that target low-level similarity [12, 55] and semantic or high-level similarity [15]. While these learned perceptual metrics enable training neural compression methods that are better aligned with human perception, they are not without issues. Directly optimizing on these perceptual metrics can exploit their null-space, improving network performance with only limited gains in actual human-perceived quality [7, 21]. Moreover, trained perceptual metrics do not necessarily generalize beyond the datasets of human judgments used to calibrate them. For instance, networks calibrated on low-level visual differences may not agree with human judgments on images with high-level semantic differences [15].

In this work, we propose an alternative to the dominant paradigm of learned and differentiable perceptual metrics used during image compression network training. First, we show that Vision-Language Models (VLMs) are effective zero-shot perceptual judges of visual similarity (Figure 1). Specifically, we show that an off-the-shelf VLM (Gemini 2.5-Flash [9]) can replicate human judgments on multiple human visual judgment datasets, namely on BAPPS [21] and our own dataset of human judgments collected on images compressed by various compression baselines.

The fact that VLM reasoning can replicate human perceptual judgments is an encouraging finding because, as VLMs are improved through considerable investments, improved automatic perceptual judges may result without additional effort to collect human judgment data and train perceptual metrics. However, it is not clear how to convert the binary 2AFC judgments produced by VLMs into an optimizable perceptual metric which can be exploited by existing GAN-based perceptually oriented compression systems.

Therefore, motivated by a desire to maximally exploit VLMs for human-oriented visual compression, we instead design a diffusion-based visual compression system similar to recent diffusion-based approaches for visual compression [6, 20, 39]. Since diffusion-based visual techniques can leverage the rich existing literature on diffusion model post-training with preferences [45], we can benefit from VLM perceptual judgments without having to use them to train a separate perceptual metric.

Concretely, we make the following contributions:

1. We show that an off-the-shelf VLM (Gemini 2.5-Flash [9]) can replicate human judgments of visual similarity on multiple human judgment datasets zero-shot.
2. We present a diffusion-based visual compression system based on FlowMo [39] extended with an additional entropy coder. We show that VLM-generated preferences, can be used to post-train this system via Diffusion DPO [45], improving performance. Moreover, we show that ensembling VLM preferences with those of a traditional perceptual metric, LPIPS, adds additional benefits and exceeds the performance of post-training with either reward alone.
3. We quantitatively study VLIC and show it achieves either competitive or state-of-the art performance relative to strong existing compression baselines, depending on the dataset, and conduct several large-scale user studies. We additionally provide several empirical analyses outlining best practices for VLM-based compression post-training, conduct ablation studies on reward design, and analyze and discuss important failure modes.

2. Related Work

Perceptually-oriented Image Compression. Prior work has considered GAN-based [18] and diffusion-based [19] perceptually-oriented image compression. GAN-based techniques such as HiFiC [31] and PO-ELIC [17] are quite popular and are often fast to decode, particularly if a factorized entropy model is used [17, 33]. Diffusion-based visual compression can be separated into two distinct styles. The first style, based on diffusion autoencoders [36], attempts to learn perceptually-oriented visual compression end-to-end with a discrete latent bottleneck [1, 2, 7, 39, 51, 52] or with the decoder conditioned on a prior learnt representation [20]. This style sometimes has the added benefit of producing a tractable latent space for downstream generative modeling. The second style involves using a trained diffusion model as an entropy coder over the diffusion model's reverse process [34, 42]. Our model architecture belongs to the first style of diffusion-based visual compression, and is derived from FlowMo [39] with a few modifications, while our training scheme is a novel combination of diffusion autoencoder training with Diffusion DPO [3].

Approximations to Human Perception of Visual Similarity. Various work has explored designing proxies for human perception, such as LPIPS [55], E-LPIPS [21], DreamSim [15], and DISTS [12]. These models have been used to calibrate learned image compression models [17, 31] but have also been used to generate different human visual content which humans will perceive similarly [14]. In visual compression, other work has argued for the use of text in the encoding process together with a generative decoder [48], with the intuition that text better captures human-relevant information in visual data. Moreover, practical implementations of text-aligned visual compression have been realized [24, 25, 29], though techniques leveraging self-supervised learning backbones [53] and generative foundation models [6] have also found success. Different from these works, we directly train a compression system using a VLM as a zero-shot proxy for human judgment of visual similarity, eschewing learned metrics calibrated on human preferences [15, 55] or heuristics for which data to encode [24, 25].

Aligning Diffusion Models to Preferences. Various works have considered how to align diffusion models to

preferences. Differentiable techniques include DRAFT [8] and VADER [35], while reinforcement learning-based techniques, such as Diffusion Direct Preference Optimization (DDPO) [45] and Denoising Diffusion Policy Optimization [3], can support non-differentiable preferences as rewards.

Several works have explored aligning diffusion models to rewards produced by VLMs, including RewardDance [49] and HSPv3 [30]. Our choice to use VLMs to produce rewards for image compression is contextually novel in the context of image compression aligned to human preferences, and is motivated by our finding that VLMs can replicate human visual similarity judgments zero-shot, but is inspired by prior works in diffusion model post-training.

3. Method

We will now provide an overview of the VLIC system. We will first review the architecture and training scheme. Then we will explain the process by which a VLM is guided to produce perceptual judgments. A full method diagram is shown in Figure 3.

Architecture and training. Our architecture and first training stage are identical to FlowMo [39], with the only architectural difference being that we use finite scalar quantization (FSQ) [32] in lieu of lookup-free quantization [54] for simplicity and to eliminate the commitment and entropy losses. We similarly adopt the rectified flow framework [27, 28] and use an LPIPS loss on the 1-step denoised prediction of the diffusion decoder following prior work [7, 39, 52].

Different from prior work, for our second stage of training we elect to post-train the model via Diffusion DPO [45] to align the diffusion model to arbitrary (potentially non-differentiable) preferences. Diffusion DPO is a variant of Direct Preference Optimization (DPO) [37] adapted to the unique setting of diffusion. The Diffusion DPO objective, adapted to the discrete diffusion autoencoder setting where the denoising network encodes and quantizes the original image $\mathbf{x}$ internally, is

$$L_{\text{DDPO}}(\theta) = -\mathbb{E}\log\sigma\left(-\beta\omega(\lambda_t)(\Delta^w - \Delta^l)\right), \quad (1)$$

where

$$\Delta^w = \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_t^w, \mathbf{x}, t)\|_2^2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\hat{\mathbf{x}}_t^w, \mathbf{x}, t)\|_2^2, \quad (2)$$

$$\Delta^l = \left\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_t^l, \mathbf{x}, t)\right\|_2^2 - \left\|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\hat{\mathbf{x}}_t^l, \mathbf{x}, t)\right\|_2^2. \quad (3)$$

The expectation is taken over sampled reconstructions from the model and ranked as winner $\hat{\mathbf{x}}_\mathbf{0}^w$ and loser $\hat{\mathbf{x}}_\mathbf{0}^l$, and with β a KL-weight controlling the degree to which the learned policy can deviate from the original reference policy. ϵ and ϵ_θ are the noise and noise estimator respectively, t the timestep, and $\omega(\lambda_t)$ an SNR-dependent weighting factor. Δ^w represents the difference between the loss on the winning example between the current and reference policy, which is intended to be decreased, while the loss difference on the losing example Δ^l is increased.

Unlike techniques requiring differentiable rewards [8, 35], this formulation of Diffusion DPO can be used with our VLM-defined rewards for image quality assessment. Moreover, unlike methods such as Denoising Diffusion Policy Optimization [3] which require a carefully tuned value function or baseline, Diffusion DPO trains stably across diverse datasets as winning samples are contrasted against losing samples with the same latent code as conditioning, which is similar to GRPO [40] with $n = 2$. We additionally co-train with the original flow matching training loss, since we find this allows us to post-train for longer without divergence. This loss is:

$$L_{\text{Flow}}(\theta) = \mathbb{E}_{\epsilon,x,t}(\|\mathbf{v} - \mathbf{v}_\theta(\mathbf{x}, \mathbf{x}_t, t)\|_2^2), \quad (4)$$

with $\boldsymbol{v} = \boldsymbol{\epsilon} + \boldsymbol{x}$ the flow matching velocity and $\mathbf{v}_\theta(\mathbf{x_t}, \mathbf{x}, t)$ the velocity estimate of the diffusion autoencoder (note that we may predict either $\boldsymbol{\epsilon}$ or $\boldsymbol{v}$ via reparameterizing the network output, which is by default in $\boldsymbol{v}$-parameterization [38] in our case) given the noisy original image $\mathbf{x_t}$, timestep t and original image $\mathbf{x}$ which is quantized within the network. Our final training loss is:

$$L(\theta) = L_{\text{DDPO}}(\theta) + \lambda_{\text{Flow}} L_{\text{Flow}}(\theta) \quad (5)$$

with hyperparameter λ_{Flow}. We train with the encoder unfrozen which led to slightly better performance and may enable the encoder to acquire features necessary to improve the reward in $L_{\text{DDPO}}(\theta)$, since the VLM reward is unseen during pretraining.

We further compress the discrete tokens from FSQ via a secondary entropy coder, which is trained separately. The entropy coder takes the form of a simple autoregressive transformer over the 1-dimensional latent sequence. After this entropy coder is trained, we use it to compress the latent code via arithmetic coding, similar to prior work [10].

The VLIC reward function. VLIC is a diffusion autoencoder, meaning that a given image is compressed deterministically to a discrete latent code c, but then decompressed stochastically. Post-training with Diffusion DPO involves sampling two decompressions of the same latent code and ranking them via a reward function to produce a winning and losing sample $\hat{\boldsymbol{x}}_0^w$ and $\hat{\boldsymbol{x}}_0^l$.

Any reward function can in principle be used to determine the winning and losing sample. In our work, we use an off-the-shelf VLM (Gemini 2.5-Flash [9]) to judge decompressed images. An overview of the reward computation is shown on the right hand side of Figure 3. Essentially, we

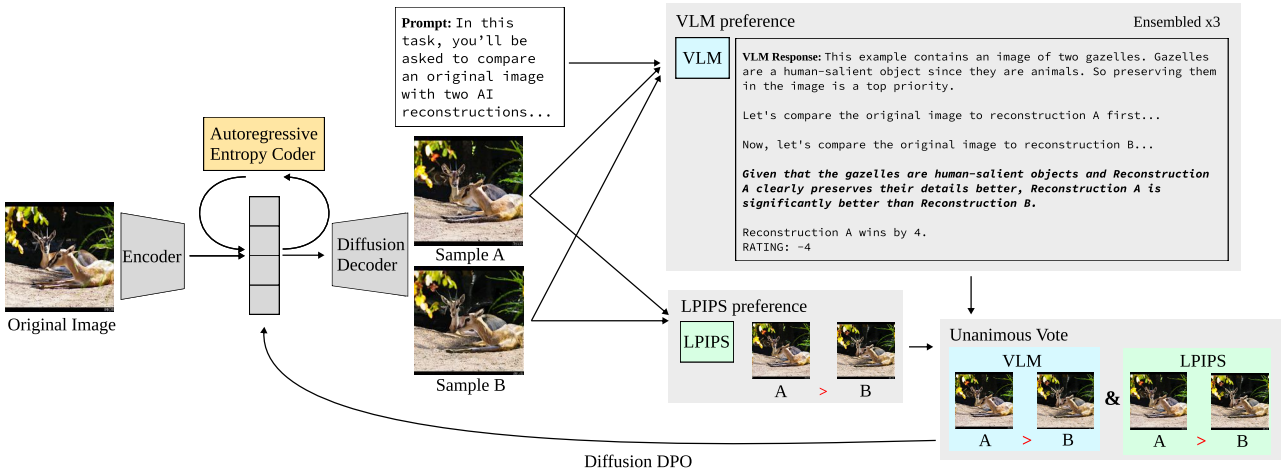

Figure 3. **Method.** An original image is encoded to a one-dimensional discrete latent code via an encoder. The discrete code is entropy coded by an auto-regressive language model. The diffusion decoder samples two reconstructions conditioned on the latent code, which are ranked via a VLM. The resulting preference is used to train the full diffusion autoencoder via Diffusion DPO [45].

pass the VLM three images: original $\boldsymbol{x}$, reconstruction A denoted $\hat{\boldsymbol{x}}_0^{\boldsymbol{A}}$, and reconstruction B denoted $\hat{\boldsymbol{x}}_0^{\boldsymbol{B}}$. We prompt the VLM to produce a numerical rating between -5 and 5 indicating whether reconstruction A or reconstruction B is closer to the original image. Negative numbers indicate A is superior. Prior to producing the numerical rating, the VLM is asked to provide detailed reasoning explaining the contents of each image and noting artifacts or inconsistencies in both reconstructions. The full text of the prompt is given in the supplementary material.

Since VLMs are prone to hallucination and to ignoring the contents of the provided images [16], we apply several mitigation strategies to improve the reliability of the reward signal. We do the following:

1. For a given random seed i, we rate each pair of reconstructed images in two orders, reversing the order the second time, so the VLM produces rewards

$$r^i_{B,0} = -r^i_{A,0} = \text{VLM}(\boldsymbol{x}, \hat{\boldsymbol{x}}_0^{\boldsymbol{A}}, \hat{\boldsymbol{x}}_0^{\boldsymbol{B}}, i)$$

and

$$r^i_{A,1} = -r^i_{B,1} = \text{VLM}(\boldsymbol{x}, \hat{\boldsymbol{x}}_0^{\boldsymbol{B}}, \hat{\boldsymbol{x}}_0^{\boldsymbol{A}}, i).$$

The final ratings per seed are then given by

$$r^i_A = \text{sign}(r^i_{A,0} + r^i_{A,1}), \quad r^i_B = \text{sign}(r^i_{B,0} + r^i_{B,1}).$$

2. Ensemble the reward over n random seeds of the VLM, so that

$$r_A = \sum_{i=1}^{n} r^i_A, \quad r_B = \sum_{i=1}^{n} r^i_B$$

where r^i_A is the rating assigned to image A for seed i, and similarly for r^i_B respectively.

3. Ensemble the VLM reward with LPIPS [55], a traditional perceptual metric. We require LPIPS and the VLM to produce a unanimous judgment in order to use a preference pair for training. If they disagree, the example is discarded.

These modifications help to reduce the noise in the VLM reward computation and provide a more consistent training signal. Importantly, ensembling with LPIPS provides superior performance to using LPIPS or the VLM reward alone, as we show later in experiments.

4. Experiments

In the following experiments, we provide some technical details on our model training and setup. Then we present our main results comparing against state-of-the-art compression baselines on standard benchmark datasets. We contextualize and interpret the results as they are presented. Finally, we provide some additional ablation studies and analysis experiments.

Our setup. We train VLIC models at two BPP values: 0.07, and 0.21. We pretrain all VLIC models for $1,000,000$ steps with Adam [22] learning rate 10^{-4}. DPO posttraining runs for $8,000$ steps with learning rate 5×10^{-7}. The batch size is 256 in both stages. Training was completed on 256 TPUv4. All models were trained in JAX [5] with bfloat16 precision. We leverage Diffusion DPO in an online fashion, sampling a preference buffer of approximately $2,560$ examples every 250 steps (some examples may be discarded for the ensemble reward). We found online training with updated buffers provided superior results compared with synthesizing an offline preference dataset. We train using VLM

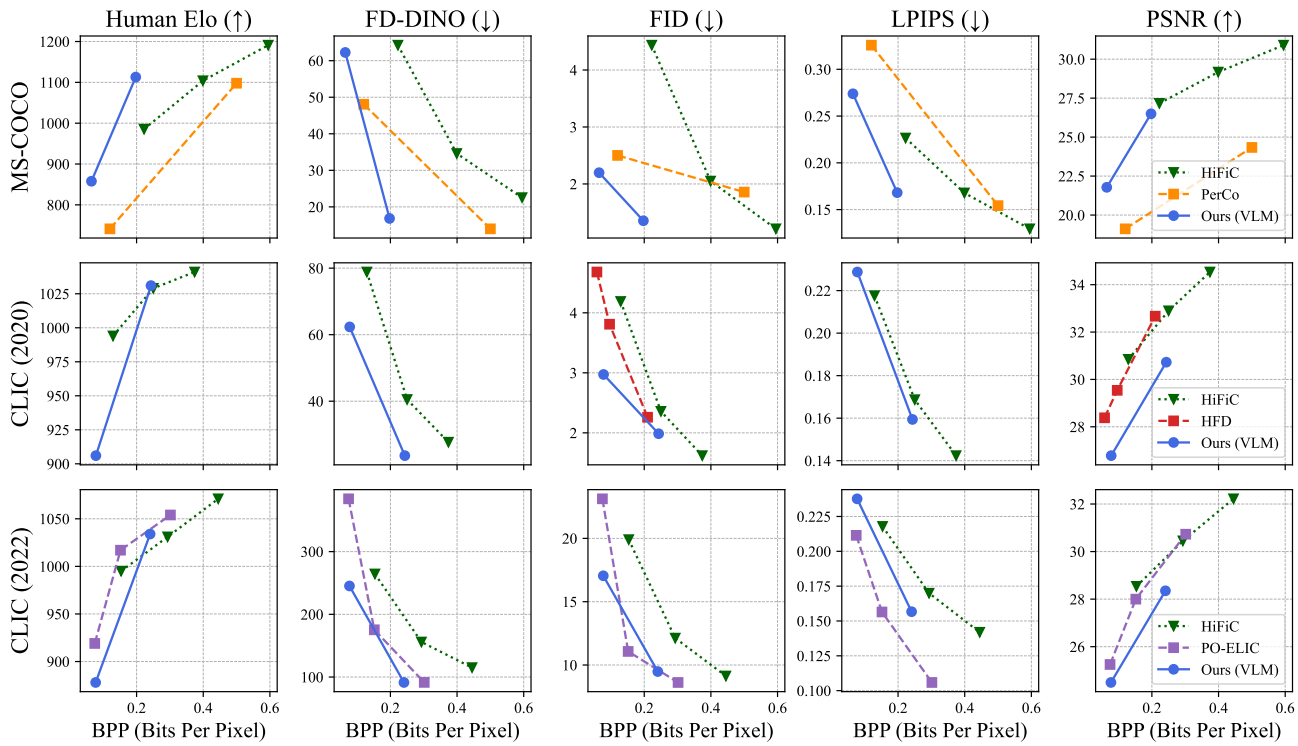

Figure 4. **Quantitative Evaluation on Image Compression Datasets.** Overall, VLIC achieves competitive or state-of-the-art performance. VLIC performs particularly well on perceptual metrics and particularly well on MS-COCO, which contains a high percentage of images with human-relevant characteristics such as text and faces.

rewards in an asynchronous fashion, simultaneously querying the VLM over the network using an updated sample buffer and performing DPO training on a slightly out-of-date sample buffer, so the latency of the VLM computation can be overlapped with DPO training.

Our models were trained at 256×256 resolution on ImageNet [11]. To accommodate variable resolutions, we use a tiled inference procedure similar to prior work [20], the details of which can be found in the supplementary material. We use a shifted schedule during inference and implement classifier-free guidance by dropping out the discrete latent code 10% of the time, following prior work [39].

4.1. Main results.

First, we will present the baselines, datasets, and metrics. Then we will analyze and explain the main results.

Baselines. We compare against HiFiC [31], PerCo [6], HFD [20], and PO-ELIC [17]. Since not all baselines have released code or reconstructions, and not all baselines can operate at all resolutions, we provide numbers only where possible for each baseline. We provide metrics for HiFiC for all datasets and metrics since the code is public. PerCo does not have released code, and after multiple attempts to correspond with the authors, we have instead chosen to rely on an open source replication [23] which matches the performance on most perceptual metrics. A limitation of PerCo is that it cannot handle high resolution data such as CLIC 2020 and CLIC 2022. HFD has not released code or reconstructions, but we have taken their CLIC 2020 numbers from the paper after exactly replicating their evaluation pipeline on CLIC 2020. PO-ELIC has not released code, but has released reconstructions on CLIC 2022, so we directly compare against the released reconstructions.

Datasets. MS-COCO [26] is a standard image compression benchmark which consists of a selection of $30,000$ images from the MS-COCO 2014 validation set. We crop these images to 256×256 using the DALL-E crop protocol following common practice [20, 31]. We compare against PerCo and HiFiC on MS-COCO.

The CLIC 2020 [43] train set is a standard image compression benchmark containing 428 high-resolution images of up to 4 megapixels. We compare against HiFiC and HFD on CLIC 2020.

The CLIC 2022 [44] test set is a standard compression benchmark of 30 high-resolution images of up to 4 megapixels. We compare against PO-ELIC and HiFiC on CLIC 2022.

Metrics. For all datasets, we evaluate with standard image quality metrics, namely LPIPS [55] and PSNR. We additionally compute two distributional image quality metrics, FID [18], and FD-DINO [41]. Finally, we compute human

MS-COCO	Human Elo ↑	FD-DINO ↓	FID ↓	LPIPS ↓	PSNR ↑
Ours (0.07bpp)	**858**	**62.25**	**2.20**	**0.274**	**21.78**
– LPIPS post-training only	838	63.63	2.33	**0.274**	21.77
Ours (0.21bpp)	**1112**	**16.83**	1.35	**0.168**	26.50
– LPIPS post-training only	1103	16.96	**1.30**	0.169	**26.54**

Table 1. **Importance of VLM.** At multiple BPP (prior to entropy coding), post-training with the VLM + LPIPS objective provides gains over post-training the compression model with LPIPS alone.

Accuracy	BAPPS-Val	Compressed Images
Human	73.99	72.15
LPIPS	69.56	92.32
DreamSim	68.13†	-
VLM	69.44	83.80

Table 2. **Human 2AFC benchmarks.** Gemini 2.5-Flash can replicate human judgments on 2AFC datasets zero-shot. †*Number taken from paper*.

Elo via large-scale user-studies.

For high-resolution datasets such as CLIC 2020 and CLIC 2022, we evaluate distributional metrics such as FID [18] and FD-DINO [41] on square 256×256 random crops, following prior work [20, 31]. To maintain a sufficient sample size, we use 100 random crops per image for CLIC 2022 and 10 random crops per image for MSCOCO.

For each dataset, we conduct large-scale user studies of the compressed images, in which users are asked to visually compare two compressed versions of the same image. We collect 1,812 pairwise ratings for CLIC 2020, 2,523 pairwise ratings for CLIC 2022, and 15,705 pairwise ratings for MS-COCO. Since Elo [13] is order-dependent but our ratings per dataset are collected in a single parallelized job, we report the Elo averaged over 10,000 random re-shufflings of the rating order. For CLIC 2020 and CLIC 2022, due to the very high resolution, raters are shown random crops rather than full images. More details on the rating process are available in the supplementary material.

We emphasize that for the evaluation of perceptually oriented visual compression, prior work has shown perceptually oriented distortion metrics and simple distortion metrics like PSNR are at odds with each other given a fixed rate [4]. Therefore, it is theoretically challenging to achieve improvements on all metrics simultaneously, so metrics correlated with human perception should be given greater weight in our analysis, even if they come at the expense of less correlated metrics. In general, we consider Elo to be the gold standard, since we directly compute it from thousands of human ratings of the test images. Prior work has also noted that FD-DINO is more predictive of human judgments than FID [41]. Finally, the poor correlation of PSNR with human assessment of perceptual quality at a given bitrate is well-known [46], and we regard it as the least important metric in our analysis.

Results. Overall, VLIC achieves very strong performance, as shown in Figure 4. It generally achieves stronger performance than HiFiC, PerCo, and HFD. It achieves particularly strong performance on MS-COCO, which contains a high percentage of images with human-relevant features such as text and faces. VLIC under-performs relative to PO-ELIC on CLIC 2022, but without released code, the performance of PO-ELIC on low-resolution data or other datasets is unclear, and it is important to note that CLIC 2022 contains only 30 images.

Additionally, VLIC tends to achieve relatively stronger metrics on human-correlated metrics (i.e., ELO, FD-DINO) compared with PSNR. For instance, VLIC achieves superior perceptual metrics in general relative to HFD and HiFiC, but worse PSNR. We deem this appropriate and even desirable, since our goal is to maximize human perception of reconstruction quality and these metrics are at odds given a fixed BPP and assuming enough model capacity [4].

4.2. Analysis

Replicating human judgments. To provide motivation for using a VLM to approximate a human perceptual judge, we show in Table 2 that Gemini 2.5-Flash can be used to replicate human perceptual judgments on the BAPPS dataset [55] and on our own collected dataset of human preferences ("Compressed images"). Compressed Images contains 5401 tuples of compressed images from two random baselines or our method; each tuple is rated at least 3 times by human raters via a custom interface. Inter-rater agreement (the "Human" row) is measured by holding out one human judgment and measuring its agreement with the average of the remaining human judgments for that tuple. It is somewhat surprising that both LPIPS and the VLM exceed the performance of a single human on our compressed images, but this is attributable to the compressed images being highly similar to the original image (much more similar than distorted BAPPS images, on average), meaning human judgments are noisier on average.

Importance of the VLM. Since our final VLM reward is ensembled with LPIPS, it is important to verify that the VLM provides gains over post-training with a binary LPIPS-determined reward alone. We provide this comparison in Table 1. Adding in the VLM reward provides a noticeable performance boost. At lower bitrates, the VLM

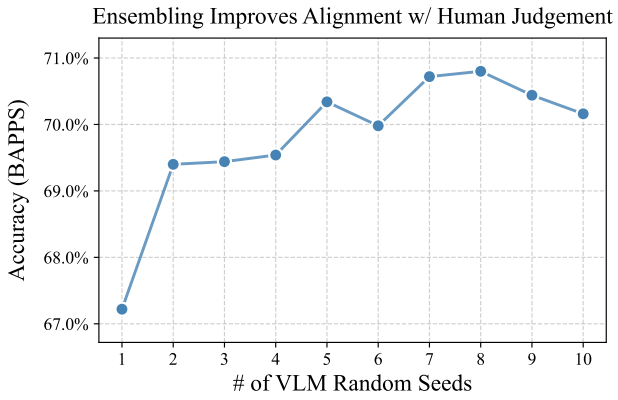

Figure 5. **Scaling self-ensembling.** The VLM becomes more predictive of human judgment on BAPPS [21] as test-time compute (number of VLM seeds) is scaled.

MS-COCO	FD-DINO ↓	FID ↓	LPIPS ↓	PSNR ↑
Ours (VLIC)	67.83	2.31	**0.278**	**21.68**
– No ensemble w/ LPIPS	**67.68**	**2.10**	0.280	21.29
– No post-training	82.31	2.40	0.300	21.27
– No self-ensembling	68.36	2.15	0.280	21.53

Table 3. **Ablation study.** Various components of reward design are necessary for best performance.

provides a stronger reward signal since VLM judgments are less noisy when the images are more different.

Mitigating noise. The noisiness of the VLM reward on highly similar images is a serious issue. Figure 6 shows a characteristic failure mode where, for highly similar reconstructions, the VLM fails to be self-consistent when the order of reconstructed images is reversed.

Indeed, reducing the noise in the VLM judgment is critical, especially since DPO struggles with noisy judgments [50]. Thankfully, the noisiness of the VLM reward can be mitigated in several ways. Most helpful is self-ensembling, *i.e.*, computing the reward as the majority vote of multiple captioning requests to the VLM, ensembled with itself over multiple random seeds. The importance of self-ensembling is reflected in Figure 5, where we observe that performance on human judgment data increases as the number of VLM random seeds is increased, though eventually saturating. For our main experiments, we use $n = 3$ seeds, though more seeds would improve performance at the cost of more VLM queries.

Ablation study. In order to verify various other components of the system, we provide a large-scale ablation study in Table 3. This study is conducted on MS-COCO. Interpreting the table, we see that only using the VLM to rank reconstructed images yields comparable performance to ensembling LPIPS and the VLM together ("No ensembling with LPIPS"), with the main difference being worse pixel-aligned metrics, namely PSNR and LPIPS, though distributional metrics actually improve. Failing to post-train the model with DPO yields poorer performance on every single metric. Not ensembling the VLM reward with itself ("No self-ensembling") yields poorer performance on the majority of metrics, as the reward becomes noisier.

VLM prompted with (Original, 1, 2): **"... Overall, Reconstruction B is better because it preserves the whiskers better than Reconstruction A and also produces a more accurate color for the bridge of the nose and has less blurring for the floral pattern. It is a slight improvement over Reconstruction A ..."**

VLM prompted with (Original, 2, 1): **"... In reconstruction B, we have less smudging on the textures of the flower, which are less defined in Reconstruction A, so this is another key win for Reconstruction B. Reconstruction B has less discoloration, so it is better. Therefore, Reconstruction B is superior ..."**

Figure 6. **Failure modes.** VLMs can hallucinate an incorrect ranking when the images are highly similar, such as in this case when the VLM fails to be self-consistent when the order of reconstructed images is reversed.

Limitations. The diffusion decoder adds additional latency compared with GAN-based compression methods, though this limitation is not unique to our method and is shared by other diffusion-based approaches [6, 20, 42, 52]. Additionally, VLM-based rewards are more expensive to compute than evaluations of a small perceptual network.

5. Conclusion

In this paper, we showed that off-the-shelf VLMs have learned a visual prior that is highly correlated with human perception. When prompted to reason about the differences between images, we showed that VLMs can replicate human similarity judgments. Motivated by this, we designed a diffusion-based compression system, VLIC, designed to be trained with VLM preferences. We then post-trained the system with VLM preferences, and achieved competitive or state-of-the-art performance on human-aligned image compression depending on the dataset.

The quality of VLIC is dependent on the accuracy of the VLM used as a perceptual judge, and as VLMs are improved through considerable research and investment, image compression techniques such as VLIC may benefit from their stronger zero-shot perceptual priors and achieve further improvements in human-aligned compression performance.

Acknowledgments. We thank Ben Poole, David Minnen, and Dina Bashkirova for helpful discussions. This work is in part supported by ONR N00014-23-1-2355 and ONR MURI N00014-22-1-2740.

References

[1] Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. FlexTok: Resampling Images into 1D Token Sequences of Flexible Length. In *ICML*, 2025. 3

[2] Vighnesh Birodkar, Gabriel Barcik, James Lyon, Sergey Ioffe, David Minnen, and Joshua V Dillon. Sample What You Can't Compress. *arXiv preprint arXiv:2409.02529*, 2024. 3

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In *ICLR*, 2024. 3, 4

[4] Yochai Blau and Tomer Michaeli. Rethinking Lossy Compression: The Rate-distortion-perception Tradeoff. In *ICML*, 2019. 7

[5] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable Transformations of Python+NumPy Programs, 2018. 5

[6] Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards Image Compression with Perfect Realism at Ultra-low Bitrates. In *ICLR*, 2023. 3, 6, 8

[7] Yinbo Chen, Rohit Girdhar, Xiaolong Wang, Sai Saketh Rambhatla, and Ishan Misra. Diffusion Autoencoders are Scalable Image Tokenizers. *arXiv preprint arXiv:2501.18593*, 2025. 3, 4

[8] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In *ICLR*, 2023. 4

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3, 4

[10] Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language Modeling Is Compression. In *ICLR*, 2024. 4

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *CVPR*, 2009. 6

[12] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020. 1, 3

[13] Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York, 1978. 7

[14] Jeremy Freeman and Eero P Simoncelli. Metamers of the Ventral Stream. *Nature Neuroscience*, 14(9):1195–201, 2011. 3

[15] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *NeurIPS*, 2023. 3

[16] Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in Plain Sight: VLMs Overlook Their Visual Representations. In *COLM*, 2025. 5

[17] 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 *CVPR*, 2022. 2, 3, 6

[18] 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. *NeurIPS*, 2017. 3, 6, 7

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *NeurIPS*, 2020. 3

[20] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis. High-Fidelity Image Compression with Score-based Generative Models. *arXiv preprint arXiv:2305.18231*, 2023. 3, 6, 7, 8

[21] Markus Kettunen, Erik Härkönen, and Jaakko Lehtinen. E-LPIPS: Robust Perceptual Image Similarity via Random Transformation Ensembles. *arXiv preprint arXiv:1906.03973*, 2019. 3, 8

[22] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 5

[23] Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. PerCo (SD): Open Perceptual Compression. *arXiv preprint arXiv:2409.20255*, 2024. 6

[24] Hagyeong Lee, Minkyu Kim, Jun-Hyuk Kim, Seungeon Kim, Dokwan Oh, and Jaeho Lee. Neural Image Compression with Text-guided Encoding for both Pixel-level and Perceptual Fidelity. In *ICML*, 2024. 3

[25] Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text+Sketch: Image Compression at Ultra Low Rates. In *ICML Workshop on Neural Compression: From Information Theory to Applications*, 2023. 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2, 6

[27] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *ICLR*, 2023. 4

[28] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2022. 4

[29] Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. AToken: A Unified Tokenizer for Vision. *arXiv preprint arXiv:2509.14476*, 2025. 3

[30] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards Wide-Spectrum Human Preference Score. In *CVPR*, 2025. 4

[31] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-Fidelity Generative Image Compression. *NeurIPS*, 2020. 2, 3, 6, 7

[32] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite Scalar Quantization: VQ-VAE Made Simple. In *ICLR*, 2024. 4

[33] David Minnen and Saurabh Singh. Channel-wise Autoregressive Entropy Models for Learned Image Compression. In *ICIP*, 2020. 3

[34] Guy Ohayon, Hila Manor, Tomer Michaeli, and Michael Elad. Compressed Image Generation with Denoising Diffusion Codebook Models. In *ICML*, 2025. 3

[35] Mihir Prabhudesai, Russell Mendonca, Zheyang Qin, Katerina Fragkiadaki, and Deepak Pathak. Video diffusion alignment via reward gradients. *arXiv preprint arXiv:2407.08737*, 2024. 4

[36] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *CVPR*, 2022. 3

[37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023. 4

[38] Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models. In *ICLR*, 2022. 4

[39] Kyle Sargent, Kyle Hsu, Justin Johnson, Li Fei-Fei, and Jiajun Wu. Flow to the Mode: Mode-Seeking Diffusion Autoencoders for State-of-the-Art Image Tokenization. In *ICCV*, 2025. 3, 4, 6

[40] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024. 4

[41] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing Flaws of Generative Model Evaluation Metrics and Their Unfair Treatment of Diffusion Models. In *NeurIPS*, 2023. 6, 7

[42] Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy Compression with Gaussian Diffusion. *arXiv preprint arXiv:2206.08889*, 2022. 3, 8

[43] George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Ballé, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. CLIC 2020 Workshop and Challenge on Learned Image Compression Dataset. https://archive.compression.cc/2020/, 2020. Accessed: November 13, 2025. 6

[44] George Toderici, Radu Timofte, Johannes Ballé, Eirikur Agustsson, Nick Johnston, Fabian Mentzer, Zeina Sinno, Andrey Norkin, Krishna Rapaka, Erfan Noury, Ross Cutler, Luca Versari, and Fabien Racapé. CLIC 2022 Workshop and Challenge on Learned Image Compression Dataset. https://archive.compression.cc/2022/, 2022. Accessed: November 13, 2025. 2, 6

[45] B Wallace, M Dang, R Rafailov, L Zhou, A Lou, S Purushwalkam, S Ermon, C Xiong, SR Joty, and N Naik. Diffusion Model Alignment Using Direct Preference Optimization. In *CVPR*, 2023. 3, 4, 5

[46] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale Structural Similarity for Image Quality Assessment. In *Asilomar Conference on Signals, Systems, and Computers*, 2003. 7

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TIP*, 13(4):600–612, 2004. 1

[48] Tsachy Weissman. Toward Textual Transform Coding. *IEEE BITS the Information Theory Magazine*, 3(2):32–40, 2023. 3

[49] Jie Wu, Yu Gao, Zilyu Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, et al. RewardDance: Reward Scaling in Visual Generation. *arXiv preprint arXiv:2509.08826*, 2025. 4

[50] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization. In *ICLR*, 2026. 8

[51] Yilun Xu, Gabriele Corso, Tommi Jaakkola, Arash Vahdat, and Karsten Kreis. DisCo-Diff: Enhancing Continuous Diffusion Models with Discrete Latents. In *ICML*, 2024. 3

[52] Ruihan Yang and Stephan Mandt. Lossy Image Compression with Conditional Diffusion Models. *NeurIPS*, 2023. 3, 4, 8

[53] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. Generation: Taming Optimization Dilemma in Latent Diffusion Models. In *CVPR*, 2025. 3

[54] Lijun Yu, José Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language Model Beats Diffusion – Tokenizer is Key to Visual Generation. In *ICLR*, 2024. 4

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. 1, 3, 5, 6, 7