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

## Supplementary Material

This supplementary material contains details regarding models, training, inference, user studies, and the VLM prompt. For more visualizations, please consult the attached supplementary webpage. Thank you!

## A. Models and training

Our model is fully transformer-based, consisting of a transformer encoder and decoder and an autoregressive transformer entropy coder. We provide the relevant hyperparameters for the encoder, decoder, and autoregressive model below. The encoder and decoder are identical to FlowMo [7], with the exception that we use FSQ quantization [6]. The entropy coder is based on NanoGPT [4]. The sequence length for the encoder and decoder is computed as the sum of the number of image tokens and latent tokens.

| Low BPP | Encoder | Decoder | Entropy coder |
|---|---|---|---|
| Hidden dim. | 768 | 1152 | 768 |
| Number of layers | 8 | 16 | 16 |
| Patch size | 4 | 4 | - |
| 1D Token size | 6 | 6 | - |
| FSQ levels | 8 | 8 | - |
| Sequence length | 4352 | 4352 | 384 |

Table 1. Model hyperparameters for low BPP configuration. Total parameter count is 1.01B

| High BPP | Encoder | Decoder | Entropy coder |
|---|---|---|---|
| Hidden dim. | 768 | 1152 | 768 |
| Number of layers | 8 | 16 | 16 |
| Patch size | 4 | 4 | - |
| 1D Token size | 18 | 18 | - |
| FSQ levels | 8 | 8 | - |
| Sequence length | 4352 | 4352 | 1152 |

Table 2. Model hyperparameters for high BPP configuration. Total parameter count is 1.01B

Training proceeds in three stages. All trainings are done on 256 TPUv4 on ImageNet [3] with resolution 256 and with batch size 256. We use a random horizontal flip and center cropping. We provide details on each stage below:

1. Pretraining is done for $1,000,000$ steps with batch size 256. We use the Adam optimizer with learning rate $10^{-4}$ [5] and no weight decay or dropout.
2. Post-training is done via Diffusion DPO with learning rate $5 \times 10^{-7}$ using the desired reward (LPIPS, VLM, VLM ensembled with LPIPS, etc.). The DPO sample buffers are recomputed every 250 steps and contain $2,560$ samples.
3. The entropy coder is trained for $200,000$ steps with learning rate $10^{-4}$, dropout with $p = 0.1$ and weight decay $0.025$ with AdamW. The encoder and decoder are frozen during this period.

We provide latency and parameter counts in Table 3. Importantly, code and parameter counts for PO-ELIC and HFD are not public. We achieve lower latency and parameter count vs. PerCo, while both diffusion-based approaches are slower than HiFiC.

| | HiFiC | PerCo | Ours |
|---|---|---|---|
| Decode time (seconds) | 0.02 | 0.78 | 0.76 |
| # Parameters ($\times 10^9$) | 0.18 | 1.29 | 1.01 |

Table 3. Latency and parameter counts.

| | CNN | Color | Deblur | FrameInterp | SuperRes | Traditional | Avg |
|---|---|---|---|---|---|---|---|
| Gemini 2.5 Flash | 0.82 | 0.58 | 0.61 | 0.59 | 0.64 | **0.83** | 0.67 |
| GPT 4.1 Mini | 0.84 | 0.62 | 0.59 | 0.56 | 0.65 | 0.74 | 0.67 |
| GPT 5 Mini | **0.87** | 0.63 | **0.62** | **0.60** | **0.70** | 0.81 | **0.71** |
| GPT 5.1 | 0.85 | **0.65** | 0.59 | 0.58 | 0.66 | 0.77 | 0.68 |

Table 4. VLM accuracy breakdown by distortion type on BAPPS.

## B. Additional experiments

In Table 4, we break down accuracy of different VLMs by distortion type on the BAPPS dataset and show the performance of different VLMs. VLMs align with human judgments for CNN and Traditional distortions while performing worse for Color/Deblur/FrameInterp. Using a stronger VLM (5 mini vs. 4.1 mini) improves performance.

In Table 5, we analyze the VLM rating stability, determined by rating each image 5 times and computing the estimated self-agreement probability $p_{agree} = 2p^2 - 2p + 1$ (it is easy to see that for $p_1, p_2 \sim Bern(p)$, we have $P(p_1 = p_2) = p_{agree}$). The per-image $p_{agree}$ estimate is then averaged per-category and then over the dataset to produce the final numbers.

In Figure 1, we show the CLIC2020 results from the main paper updated to include JPEG and HEIC. Mainstream codecs perform well on PSNR, while performing significantly worse than learned methods on the three other per-

| | CNN | Color | Deblur | FrameInterp | SuperRes | Traditional | Avg |
|---|---|---|---|---|---|---|---|
| GPT 5.1 | 0.85 | 0.71 | 0.67 | 0.68 | 0.70 | 0.79 | 0.73 |

Table 5. VLM rating stability per-category and overall. We report an estimate of the probability $p$ that two randomly sampled VLM ratings of the same image tuple agree.

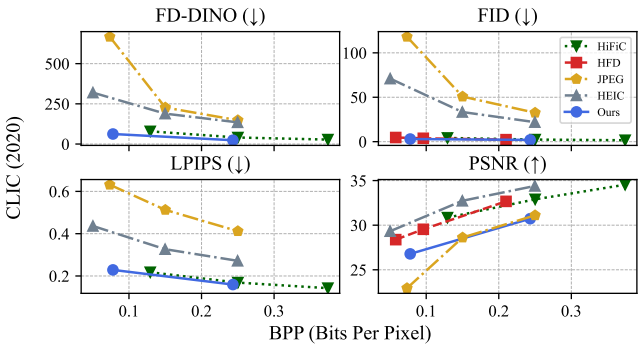


Figure 1. Classical codecs performance. JPEG and HEIC perform well on PSNR while performing more poorly on perceptually-oriented metrics.

ceptual metrics. This is in line with the rate-distortion-perception tradeoff [2], which suggests that PSNR and perceptually-oriented distance metrics are at odds given a fixed rate.

In Table 6, we provide all the raw numbers for our baseline comparison, for ease of comparisons with future work.

## C. VLM prompt

The full prompt for Gemini 2.5 Flash is shown below:

*"In this task, you'll be asked to compare an original image with two AI reconstructions of that image, Reconstruction A followed by Reconstruction B. You'll see triplets of images like (original, A, B). You'll give a relative rating of the two images. This is between -5 and 5, inclusive. Higher scores mean that Reconstruction B is relatively better. So if you give a score of -2, you think A is kind of better, and if you give a score of +5, you think B is obviously significantly better. Be sparing with the higher magnitude scores - you should expect that most triplets you see won't have an obviously better reconstruction. In your ratings, make sure to prioritize differences that are semantically important to a human observer. If a distortion changes the meaning of an image to a human observer, then it's more significant than if a distortion changes the texture of an image. Your response should conclude with "RATING: X", where X is your rating, i.e. -2 or 3. Now here is the image triplet that you need to rate. Make sure to provide a lot more reasoning and make sure to carefully look at each of the three images and provide meticulous visual justifications based on evidence from each image! Remember, negative scores mean that A is better, and positive scores mean that B is better."*

Once the response is produced, the reward is computed as described in the method section, given the final numerical rating.

## D. Inference

Since our model is only trained on $256 \times 256$ images, a zero-shot procedure is needed to support inference at higher resolutions at test time. A visualization of the tiled inference strategy we use to support any-resolution compression is shown in Figure 2. Where the image dimensions are $h, w$, $r$ is the native model resolution and $p$ is the margin size, we compute the smallest $k_h, k_w$ such that $r + k_h(r - p) \geq h$ and $r + k_w(r - p) \geq w$. The image is then resized to $r + k_h(r - p), r + k_w(r - p)$ and broken into overlapping tiles, which are separately encoded. The image tiles are then diffused jointly following MultiDiffusion [1].

## E. User studies

For policy-related reasons, we cannot share visuals of the exact user interface. Users are presented with three images: Image A, Original Image, and Image B (in that order). The user is asked to select an answer from 5 options:

*Which image is more similar to the original image?*

- *Image A is much more similar*
- *Image A is slightly more similar*
- *About the same*
- *Image B is slightly more similar*
- *Image B is much more similar*

Since users may have varying display sizes, images are resized to 480 pixels in width. We show random crops for CLIC 2020 and CLIC 2022 due to the very high resolution.

## F. Additional capabilities

Since Diffusion DPO only requires binary preferences, we may leverage the VLM to rate reconstructed images according to criteria other than perceptual quality. For instance, in Figure 3, we instruct the VLM to read out the readable text in the original and reconstructed image, and compute the reward as the edit distance between the text in the original and reconstructed image. In this case, the model degenerates to a local minimum where the image text is censored, so the readable text in the reconstruction has edit distance bounded by the length of the text in the original image. Alternative formulations of edit distance can lead to slightly improved

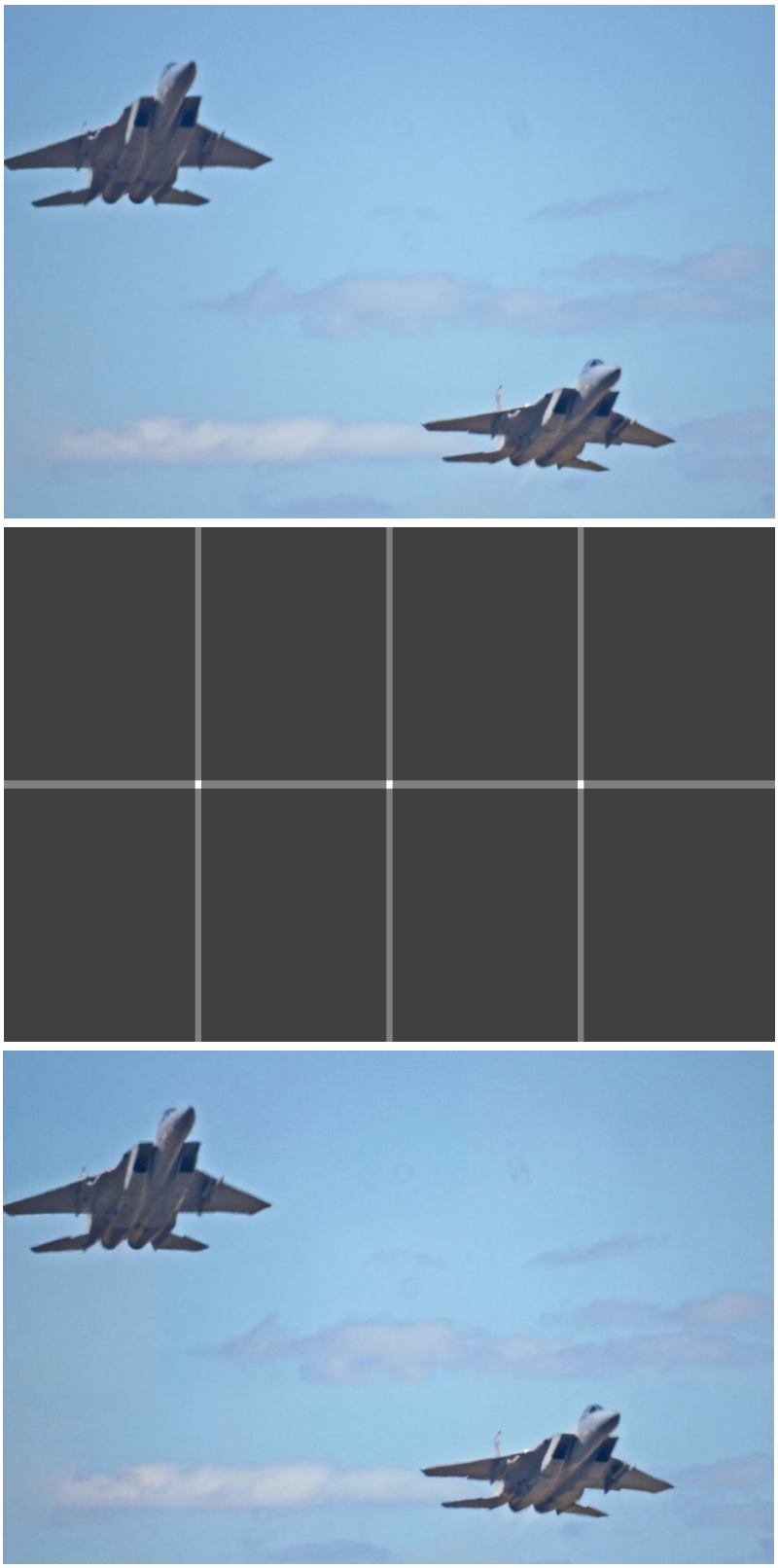

Figure 2. **Tiled inference for arbitrary resolutions.** From top to bottom: Original image, tiling strategy, reconstructed image. The margin size (we use 8 pixels in this work) must be large enough to communicate information between patches during diffusion to avoid unsightly border artifacts, but not so large as to waste BPP.

| Dataset | Model | BPP | PSNR | LPIPS | FID | FD-DINO | ELO |
|---|---|---|---|---|---|---|---|
| CLIC (2020) | **HiFiC** | 0.13 | 30.85 | 0.22 | 4.19 | 78.84 | 994 |
| | **HiFiC** | 0.25 | 32.90 | 0.17 | 2.36 | 40.51 | 1029 |
| | **HiFiC** | 0.37 | 34.54 | 0.14 | 1.62 | 27.68 | 1041 |
| | **HFD** | 0.06 | 28.38 | - | 4.68 | - | - |
| | **HFD** | 0.10 | 29.54 | - | 3.81 | - | - |
| | **HFD** | 0.21 | 32.67 | - | 2.26 | - | - |
| | **VLIC (ours)** | 0.08 | 26.78 | 0.23 | 2.97 | 62.34 | 906 |
| | **VLIC (ours)** | 0.24 | 30.73 | 0.16 | 1.99 | 23.63 | 1031 |
| MSCOCO | **HiFiC** | 0.22 | 27.16 | 0.23 | 4.44 | 64.24 | 985 |
| | **HiFiC** | 0.40 | 29.17 | 0.17 | 2.05 | 34.60 | 1104 |
| | **HiFiC** | 0.60 | 30.90 | 0.13 | 1.21 | 22.53 | 1191 |
| | **PerCo** | 0.13 | 19.12 | 0.33 | 2.50 | 48.08 | 741 |
| | **PerCo** | 0.50 | 24.34 | 0.15 | 1.86 | 13.99 | 1098 |
| | **VLIC (ours)** | 0.06 | 21.78 | 0.27 | 2.20 | 62.25 | 858 |
| | **VLIC (ours)** | 0.20 | 26.50 | 0.17 | 1.35 | 16.83 | 1113 |
| CLIC (2022) | **HiFiC** | 0.15 | 28.54 | 0.22 | 19.90 | 264.38 | 995 |
| | **HiFiC** | 0.29 | 30.45 | 0.17 | 12.12 | 155.88 | 1031 |
| | **HiFiC** | 0.44 | 32.22 | 0.14 | 9.13 | 115.01 | 1071 |
| | **PO-ELIC** | 0.07 | 25.25 | 0.21 | 23.12 | 384.20 | 919 |
| | **PO-ELIC** | 0.15 | 28.00 | 0.16 | 11.07 | 175.33 | 1017 |
| | **PO-ELIC** | 0.30 | 30.73 | 0.11 | 8.62 | 91.46 | 1054 |
| | **VLIC (ours)** | 0.08 | 24.50 | 0.24 | 17.05 | 245.23 | 878 |
| | **VLIC (ours)** | 0.24 | 28.35 | 0.16 | 9.48 | 91.34 | 1034 |

Table 6. Raw numbers for all models comparison, compare with Figure 4 of main paper.


Figure 3. **Censoring readable text.** A failure case of an edit-distance based reward on readable text determined by the VLM causes the model to degenerate to censoring all readable text in the images.

performance on average for rendering readable text, though the difference is not generally noticeable.

## G. Disclaimer

Some images in this paper contain content which is sensitive to left-right orientation such as readable text. While

some standard benchmark datasets such as MS-COCO are prepared with random flips, we may re-flip the reconstructions of both ours and the baselines for ease of visualization.

## H. Image attributions

**Figure 1, Truck**
- Sourced From: ImageNet
- Image ID: n03594945_15055
- License: ImageNet Agreement

**Figure 1, Street Sign**
- Sourced From: COCO
- Image ID: 000000001779
- Original Creator: I am R.
- License: CC BY 2.0
- URL: https://www.flickr.com/photos/isfullofcrap/2368878795/

**Figure 1, Bar**
- Sourced From: COCO
- Image ID: 000000000723
- Original Creator: Henry Zbyszynski
- License: CC BY 2.0
- URL: https://www.flickr.com/photos/hankzby/7385695522/

**Figure 2, Chicago Skyline**
- Sourced From: CLIC 2022
- Image ID: 07113e38700d3f0dab7a9f34d451298a54de3cef3bc4e03945d5fead4f513ecd
- License: Unsplash

**Figure 2, Street Sign**
- Sourced From: COCO
- Image ID: 000000000250
- Original Creator: Unknown
- License: CC BY-NC-SA 2.0
- URL: http://images.cocodataset.org/train2014/COCO_train2014_000000000250.jpg

**Figure 2, Women on Phone**
- Sourced From: COCO
- Image ID: 000000000536
- Original Creator: Unknown
- License: CC BY-NC-SA 2.0
- URL: http://images.cocodataset.org/val2014/COCO_val2014_000000000536.jpg

**Figure 2, Soldiers**
- Sourced From: COCO
- Image ID: 000000001149
- Original Creator: Picatinny Arsenal
- License: CC BY-NC 2.0
- URL: https://www.flickr.com/photos/picatinnyarsenal/5202323780/

**Figure 3, Gazelles**
- Sourced From: ImageNet
- Image ID: n02423022_31692
- License: ImageNet Agreement

**Figure 6, Mouse**
- Sourced From: ImageNet
- License: ImageNet Agreement

**Gallery, Bridge**
- Sourced From: CLIC 2022
- Image ID: 732bf474788c19c0c1fae6dd7689d4cda2f4e0632a1c7725e970b69d44d08f3e
- License: Unsplash

**Gallery, Tennis player**
- Sourced From: COCO
- Image ID: 000000001815
- Original Creator: Kate Tann
- License: CC BY-SA 2.0
- URL: https://www.flickr.com/photos/43555660@N00/6050671677/

**Gallery, Woman**
- Sourced From: COCO
- Image ID: 000000001569
- Original Creator: Tom Conger
- License: CC BY-NC 2.0
- https://www.flickr.com/photos/tomconger/3578082722/

**Gallery, Hunter**
- Sourced From: COCO
- Image ID: 000000001948
- Original Creator: Unknown
- License: CC BY-NC-SA 2.0
- http://images.cocodataset.org/val2014/COCO_val2014_000000001948.jpg

**Gallery, Dinner**
- Sourced From: CLIC 2020
- License: Unsplash

**Gallery, Lady in dress**
- Sourced From: CLIC 2020
- License: Unsplash

**Gallery, Cat**
- Sourced From: CLIC 2020
- License: Unsplash

**Gallery, Firehouse**
- Sourced From: CLIC 2020
- License: Unsplash

## References


[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *ICML*, 2023. 2

[2] Yochai Blau and Tomer Michaeli. Rethinking Lossy Compression: The Rate-distortion-perception Tradeoff. In *ICML*, 2019. 2

[3] 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. 1

[4] Andrej Karpathy. nanoGPT. 2023. 1

[5] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 1

[6] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite Scalar Quantization: VQ-VAE Made Simple. In *ICLR*, 2024. 1

[7] 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. 1