# Text + Sketch: Image Compression at Ultra Low Rates

**Eric Lei** [1] **Yiğit Berkay Uslu** [1] **Hamed Hassani** [1] **Shirin Saeedi Bidokhti** [1]

## Abstract

Recent advances in text-to-image generative models provide the ability to generate high-quality images from short text descriptions. These foundation models, when pre-trained on billion-scale datasets, are effective for various downstream tasks with little or no further training. A natural question to ask is how such models may be adapted for image compression. We investigate several techniques in which the pre-trained models can be directly used to implement compression schemes targeting novel low rate regimes. We show how text descriptions can be used in conjunction with side information to generate high-fidelity reconstructions that preserve both semantics and spatial structure of the original. We demonstrate that at very low bit-rates, our method can significantly improve upon learned compressors in terms of perceptual and semantic fidelity, despite no end-to-end training.

## 1. Introduction

Recent works from the lossy compression literature have demonstrated that when human satisfaction or semantic visual information is prioritized, compression schemes that manually encode images using human-written text descriptions as the compressed representation (Bhown et al., 2018; 2019) yield significant improvements compared to traditional compressors. These works show that when operating at such low bit-rates, high levels of human satisfaction can still be achieved despite low pixel-wise fidelity. (Weissman, 2023) argues that transmitting the compressed information directly in the form of human language, known as textual transform coding, encodes information that scales with the semantic content in the image as interpreted by a human, rather than pixel-wise content.

[1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. Correspondence to: Eric Lei <elei@seas.upenn.edu>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

(a) Ground-truth.

(b) Text only reconstr. (0.0023 bpp).

(c) Text + sketch reconstr. (0.013 bpp).

*Figure 1.* Text-only reconstruction (PIC) preserves semantic information. Adding a sketch (PICS) preserves structural components.

Concurrent work in text-to-image generative models have provided the ability to generate high-quality images that represent the semantic information of the text across many domains (Ramesh et al., 2022; Rombach et al., 2022). These models, when scaled to orders of magnitude larger parameter counts and billion-scale datasets, have achieved remarkable capabilities in terms of converting language concepts to high quality images when assessed by humans. At such scale, these foundation models provide impressive zero-shot capabilities, allowing them to be used as a backbone when designing models for tasks not explicitly trained for.

Prior neural compression paradigms, such as generative compression, attempt to align its reconstructions with human assessment at low bit-rates by enforcing a distribution matching constraint. In contrast, our work investigates neural compression schemes that target human satisfaction by directly transmitting text containing human-aligned semantic information. By leveraging the recent advances in pre-trained foundation models that operate with vision and language, we demonstrate how neural compression can benefit from the scale of such models, whereas similarly scaled neural compressors would require extensive resources to train end-to-end.

Directly using an off-the-shelf text-to-image model (with no

further training) to implement a textual transform code can yield good results in terms of preserving coarse semantic information at very low bit-rates. However, current language-vision models, typically built on top of CLIP (Radford et al., 2021), are limited in the amount of semantic concepts they can synthesize, especially pertaining to the spatial placement of objects. As shown in Fig. 1, when sending a text that is CLIP-optimized as the compressed representation, coarse semantic information is kept, but lower-level details of the image such as the placement of objects is poor. We show how transmitting limited side information in the form of a sketch can preserve lower-level structures. Our full contributions are as follows.

1) We design a neural compressor that uses text-to-image models in a zero-shot manner to implement compression schemes preserving human semantics at rates below 0.003 bits-per-pixel (bpp), which is an order of magnitude lower than previously studied regimes.

2) We show how side information in the form of a compressed spatial conditioning map can be used to provide the high-level structural information in the image along with a transmitted text caption, producing reconstructions that improve structural preservation.

3) We show that our schemes outperform state-of-the-art generative compressors in terms of semantic and perceptual quality, despite no end-to-end training.

## 2. Related Work

**Neural Compression.** The use of neural networks to design lossy compressors was initiated by merging quantization with autoencoder architectures (Toderici et al., 2016; Ballé et al., 2017; Theis et al., 2017; Agustsson et al., 2017). These models are traditionally trained with reference distortion metrics such as MSE, MS-SSIM (Wang et al., 2003), and LPIPS (Zhang et al., 2018). However, reconstructions suffer from blurriness at low bit-rates, motivating the field of generative compression (Agustsson et al., 2019; Mentzer et al., 2020). In this field, distortion can be sacrificed for perceptual quality (Blau & Michaeli, 2019), measured as alignment between source and reconstruction distributions. This improves human satisfaction in the rate regime of $<0.1$ bpp, compressors tuned for pixel-wise distortions fail to generate realistic reconstructions. At such low bitrates, pixel-wise fidelity metrics fail to align with human perception, since they largely focus on low-level details rather than the higher-level structures. Generative compression thus allows for realistic but not necessarily faithful (with regards to a distortion measure) reconstructions. However, it poses realism in terms of a distribution matching formulation which can offer some alignment with human satisfaction; textual transform coding attempts to directly encode the human-aligned semantic information in the form of language.

**Text-to-Image Models.** While many architectures have been studied for text-to-image generation, such as VAEs (Ramesh et al., 2021; Ding et al., 2021) and GANs (Gal et al., 2021), diffusion models have become the method of choice due to easier scaling to massive datasets (Rombach et al., 2022; Ramesh et al., 2022). These methods typically leverage CLIP (Radford et al., 2021), a pre-trained model that provides a shared text-image embedding space, to retrieve an embedding corresponding to the input text. The diffusion model uses this embedding as a conditional input to denoise randomly sampled noise into an image corresponding to the text. Our work does not necessarily require diffusion models per se; it can use any foundation model that can generate images from text, pre-trained at scale.

Diffusion-based neural compressors have also been investigated (Yang & Mandt, 2022; Pan et al., 2022). Rather than transmit text, these models transmit a quantized embedding as the conditional input to the diffusion-based decoder. DiffC (Theis et al., 2022) directly transmits pixels corrupted by noise in a diffusion process. Contrary to these models, our proposed compressor uses fully pre-trained text-to-image models, transmits text directly as a compressed representation for the conditional input, and utilizes a spatial conditioning input as side information.

**Human Compression.** (Bhown et al., 2018; 2019) demonstrates a hand-crafted compression scheme in which humans write down text descriptions of the image to compress; the decoder consists of another human who has access to a database of images and image editing software. Human-rated scores for this scheme were higher than WebP at similar rates, despite the fact that the reconstructions may not necessarily be faithful at the pixel-level. Building off these results, (Weissman, 2023) conjectures that human satisfaction is a function of pixel-level fidelity with a semantic fidelity, which can be interpreted via human language. At large rates, pixel-wise fidelity dominates human satisfaction; at low rates, pixel-wise fidelity becomes less meaningful when compared to the "textual" information of the image.

## 3. Transmitting Text With Side Information

### 3.1. Textual Transform Coding via Prompt Inversion

Textual transform coding (Weissman, 2023) represents the image using a text description, which gets encoded with a lossless compressor. The decoder first recovers the text, which is used to synthesize the reconstructed image. Our decoder is assumed to be some text-to-image model $G$ that is pre-trained on a large-scale dataset. In this section, we use Stable Diffusion (SD) (Rombach et al., 2022) for $G$.

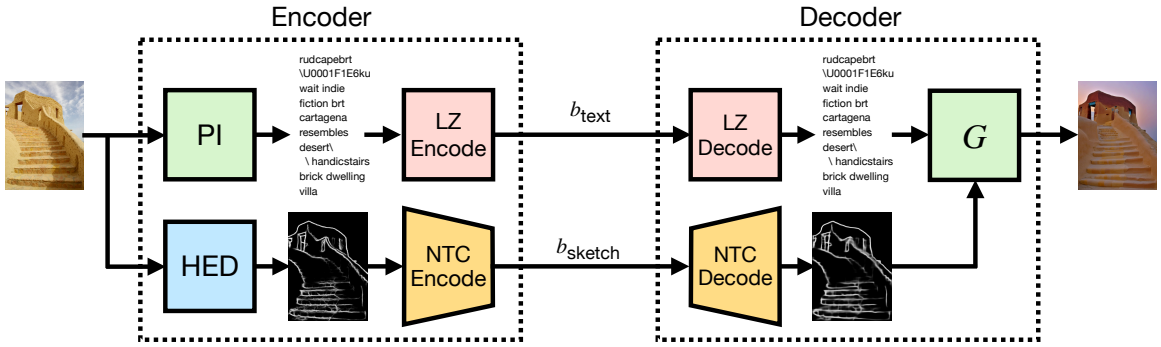



*Figure 2.* Diagram of PICS. Separate bitstreams for the text and sketch are losslessly encoded. Removing the bottom branch yields PIC.

One option to encode an image into text is via image captioning methods (Stefanini et al., 2022). However, most image captioning methods such as (Li et al., 2022) produce text captions that align with human language, but may not necessarily be optimal for the text-to-image model. Since SD uses pre-trained CLIP for text embeddings, it is more meaningful to directly search in the embedding space of CLIP in order to find text that represents an image for SD.

Following (Wen et al., 2023), we use prompt inversion (PI), which performs projected gradient search in CLIP's embedding space, using cosine similarity between the image embedding and the text embedding as the objective. To project to a hard text, the nearest CLIP embedding is found for each token being searched over. The tokens are converted to text and losslessly compressed. At the decoder, the decoded text is simply provided to $G$ which synthesizes a reconstructed image. We call this method Prompt Inversion Compression (PIC). PIC can achieve very low rates (around 0.002-0.003 bpp), yet preserve semantic information, since CLIP itself has semantic image comparison capabilities due to its vision-text merged feature space.

An interesting fact of language-vision models such as SD is that quantization is naturally built into the model, where the language to vision conversion takes place. Text, after converted to tokens, is directly mapped to a codebook of embedding vectors. Thus, one can interpret prompt inversion as the encoder searching for the best CLIP codeword.

### 3.2. Adding Spatial Conditioning Maps

One challenge with using PIC is that it is difficult to increase reconstruction quality as the bitrate of text increases. As shown in (Wen et al., 2023), increasing the number of tokens after a certain point fails to improve the CLIP score of the reconstructed image. Rather than attempting to increase the textual information in a way that $G$ can process, we instead propose to send side information in the form of a "sketch" of the original image, which contains finer structural information.

In this setting, we choose $G$ to be ControlNet (Zhang & Agrawala, 2023), a text-to-image model built on top of SD that can process spatial conditioning maps in the form of edge detection maps, segementation maps, depth maps, etc. It ensures that the reconstructed images follow the spatial structure of the input map, and the style suggested by the text prompt. We use ControlNet as our decoder by sending a compressed version of the edge detection map (i.e., the sketch) as side information in addition to the prompt inversion text. In particular, we use the variant of ControlNet trained with Holistically-nested Edge Detection (HED) maps (Xie & Tu, 2015) since those were found to have lower rate-distortion compared to Canny edge and segmentation maps. To compress the sketch, we use standard learned non-linear transform codes (NTC) (Ballé et al., 2021) trained on a small dataset of HED maps. We call this scheme Prompt Inversion Compressor with Sketch (PICS), shown in Fig. 2.

## 4. Experimental Results

### 4.1. Setup

**Datasets and Evaluation**: We use three evaluation datasets: Kodak (Franzen), CLIC 2021 (CLI) test, and DIV2K (Agustsson & Timofte, 2017) validation. Since textual transform coding operates in an order of magnitude lower regime than even "extreme" compression ($< 0.1$bpp), pixel-wise reference distortion metrics (PSNR, MS-SSIM, LPIPS) are not as meaningful. As human-aligned semantic reference metrics are still an open problem (Weissman, 2023), we use cosine similarity of CLIP embeddings as a proxy,

$$d_{\text{CLIP}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = 1 - \frac{e(\boldsymbol{x}) \cdot e(\hat{\boldsymbol{x}})}{\|e(\boldsymbol{x})\| \|e(\hat{\boldsymbol{x}})\|}, \tag{1}$$

where $e(\cdot)$ is the image encoder of CLIP. Ideally, a human study would be performed, which we leave for future work. In addition, we use standard no-reference metrics to measure realism according to distributional alignment, FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018).

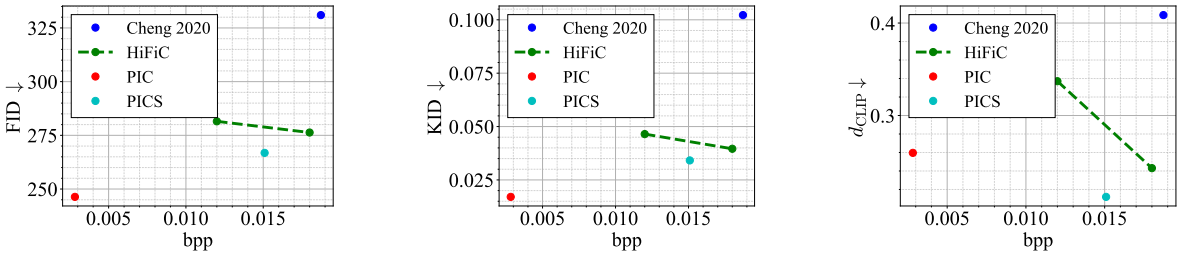



*Figure 3.* Achieved rate-perception and rate-distortion tradeoffs on CLIC 2021.

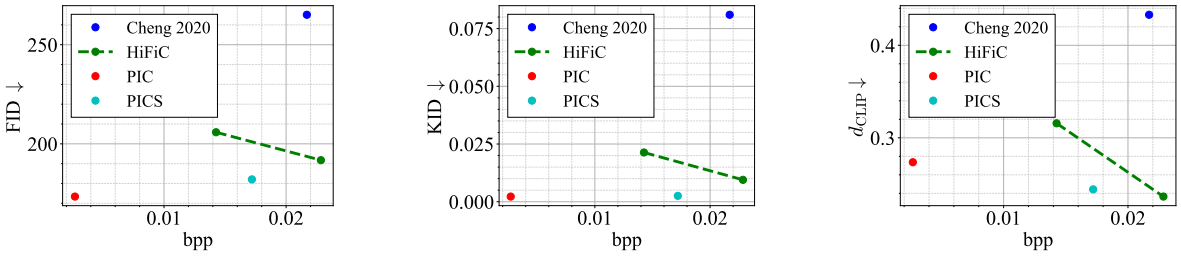
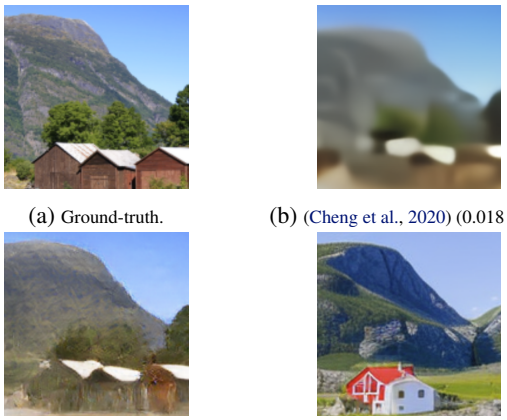



*Figure 4.* Achieved rate-perception and rate-distortion tradeoffs on DIV2K.



(a) Ground-truth. (b) (Cheng et al., 2020) (0.018 bpp). (c) HiFiC (0.016 bpp). (d) PICS (0.013 bpp).

*Figure 5.* Zoomed-in version of Fig. 1.

**Baseline Methods**: These include a generative compression baseline, HiFiC (Mentzer et al., 2020), and a NTC baseline (Cheng et al., 2020) optimized for MS-SSIM.

**PIC/PICS**: See appendix.

### 4.2. Results

**Quantitative Results**: Shown in Figs. 3, 4, we compare PIC/PICS in terms of rate-perception and distortion (in terms of $d_{\text{CLIP}}$). At such a low rate regime, HiFiC achieves better rate and semantic and perceptual quality than MS-SSIM trained NTC models. However, PICS is able to improve upon that further, with strict improvement in all tradeoffs. Interestingly, while PIC also strictly improves the rate-perception tradeoff, it performs worse in terms of semantic quality than PICS and HiFiC (albeit at lower rate). This shows that adding the sketch actually helps the generative model achieve higher semantic quality.

**Qualitative Results**: We visualize several reconstruction examples for all models and compare them with the ground-truths, in Figs. 1, 5, 6, 7, and 8. In general, PIC is able to reconstruct very coarse concepts contained in the ground-truth image. The NTC model optimized for rate-distortion yields blurry reconstructions in the low-rate regime. HiFiC improves realism, producing a sharper image with perhaps different textures than the original. In some cases, there are still compression artifacts, since HiFiC is not operating in the (near)-perfect realism regime. PICS is able to recover the high-level spatial structure of the ground-truth with superior sharpness, but synthesizes different textures or colors in the image. For example, Fig. 5 shows how PICS generates a house in front of a mountain of similar shape, but completely changes the color and style of the house as well as the composition of the mountainside. Additionally, PI-encoded prompts mostly recover semantic concepts, in Figs. 6-8.

## 5. Conclusion

In this paper, we use pretrained text-to-image models to construct a compressor that transmits a short text prompt and compressed image sketch. The only training required is to learn a lightweight learned compressor on HED sketches. Experimental results demonstrate superior performance in terms of semantic and perceptual quality. Current and future work includes a human study to evaluate human satisfaction

of reconstructed images.

## References


Clic 2021: Challenge on learned image compression. URL https://clic.compression.cc/2021/index.html.

Agustsson, E. and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.

Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., and Gool, L. V. Soft-to-hard vector quantization for end-to-end learning compressible representations. *Advances in neural information processing systems*, 30, 2017.

Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., and Gool, L. V. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 221–231, 2019.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJxdQ3jeg.

Ballé, J., Chou, P. A., Minnen, D., Singh, S., Johnston, N., Agustsson, E., Hwang, S. J., and Toderici, G. Nonlinear transform coding. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):339–353, 2021. doi: 10.1109/JSTSP.2020.3034501.

Bhown, A., Mukherjee, S., Yang, S., Chandak, S., Fischer-Hwang, I., Tatwawadi, K., Fan, J., and Weissman, T. Towards improved lossy image compression: Human image reconstruction with public-domain images. *arXiv preprint arXiv:1810.11137*, 2018.

Bhown, A., Mukherjee, S., Yang, S., Chandak, S., Fischer-Hwang, I., Tatwawadi, K., and Weissman, T. Humans are still the best lossy image compressors. In *2019 Data Compression Conference (DCC)*, pp. 558–558, 2019. doi: 10.1109/DCC.2019.00070.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Blau, Y. and Michaeli, T. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7948, 2020.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., and Tang, J. Cogview: Mastering text-to-image generation via transformers. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19822–19835. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a4d92e2cd541fca87e4620aba658316d-Paper.pdf.

Franzen, R. W. Kodak lossless true color image suite. URL https://r0k.us/graphics/kodak/.

Gal, R., Patashnik, O., Maron, H., Chechik, G., and Cohen-Or, D. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and 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.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Lefaudeux, B., Massa, F., Liskovich, D., Xiong, W., Caggiano, V., Naren, S., Xu, M., Hu, J., Tintore, M., Zhang, S., Labatut, P., and Haziza, D. xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers, 2022.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Mentzer, F., Toderici, G. D., Tschannen, M., and Agustsson, E. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020.

Obukhov, A., Seitzer, M., Wu, P.-W., Zhydenko, S., Kyl, J., and Lin, E. Y.-J. High-fidelity performance metrics

for generative models in pytorch, 2020. URL https://github.com/toshas/torch-fidelity. Version: 0.3.0, DOI: 10.5281/zenodo.4957738.

Pan, Z., Zhou, X., and Tian, H. Extreme generative image compression by learning text embedding from diffusion models. *arXiv preprint arXiv:2211.07793*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and 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.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 539–559, 2022.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy image compression with compressive autoencoders. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJiNwv9gg.

Theis, L., Salimans, T., Hoffman, M. D., and Mentzer, F. Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022.

Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., Covell, M., and Sukthankar, R. Variable rate image compression with recurrent neural networks. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06085.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Wang, Z., Simoncelli, E., and Bovik, A. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Weissman, T. Toward textual transform coding. *arXiv preprint arXiv:2305.01857*, 2023.

Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., and Goldstein, T. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023.

Xie, S. and Tu, Z. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1395–1403, 2015.

Xue, T., Chen, B., Wu, J., Wei, D., and Freeman, W. T. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019.

Yang, R. and Mandt, S. Lossy image compression with conditional diffusion models. *arXiv preprint arXiv:2209.06950*, 2022.

Zhang, L. and Agrawala, M. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00068. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068.

# A. Visual Reconstructions

We place visual reconstructions referenced in the main text here.

# B. Implementation Details

## B.1. Baselines and Evaluation

For HiFiC, we use an open-source implementation[1] pre-trained on OpenImages (Kuznetsova et al., 2020) to a target bitrate of 0.14 bpp. We then fine-tune on a subset of OpenImages with a target bitrate of 0.01 bpp, by setting $\lambda^{(a)} = 32, 64$. For the NTC baseline, we use a model pre-trained[2] on Vimeo90K (Xue et al., 2019), fine-tuned on the same dataset for a target bitrate of 0.01 bpp.

To compute FID and KID, we use the torch-fidelity[3] (Obukhov et al., 2020) package.

## B.2. PIC/PICS

For PI, we set the prompt length to 16 tokens, following the ablation study in (Wen et al., 2023). To compress the HED sketch, we train a lightweight NTC model (Cheng et al., 2020) on HED maps from Vimeo90K under MS-SSIM distortion, targeting a bitrate of 0.01 bpp. We found that using MS-SSIM yielded better reconstructions from ControlNet compared to PSNR.

We use HuggingFace's diffusers library (von Platen et al., 2022) to run inference on SD and ControlNet. Although SD and ControlNet use many more parameters than NTC or HiFiC, one does not need to train these foundation models. Furthermore, with recent advances in efficient inference of diffusion models (Lefaudeux et al., 2022), inference can be run efficiently on a single commodity GPU without using excessive memory. The code will be made available at https://github.com/leieric/Text-Sketch.

[1] https://github.com/Justin-Tan/high-fidelity-generative-compression
[2] https://interdigitalinc.github.io/CompressAI/
[3] https://github.com/toshas/torch-fidelity

(a) Ground-truth.

(b) (Cheng et al., 2020) (0.015 bpp).

(c) HiFiC reconstr. (0.013 bpp).

(d) PIC reconstr. (0.0023 bpp).

(e) PICS reconstr. (0.013 bpp).

(f) PICS sketch.

(g) PICS sketch reconstr., 0.01 bpp.

*Figure 6.* From CLIC2021 test. For PIC/PICS, encoded prompt is “rudcapebrt “U0001F1E6kuwait indie fiction brt cartagena resembles desert handicstairs brick dwelling villa”.

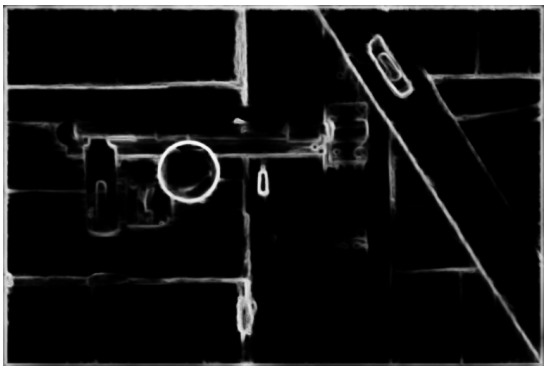
(a) Ground-truth.

(b) (Cheng et al., 2020) (0.013 bpp).

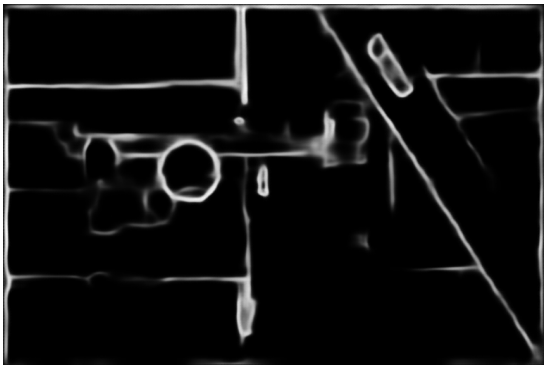
(c) HiFiC reconstr. (0.020 bpp).

(d) PIC reconstr. (0.0024 bpp).

(e) PICS reconstr. (0.011 bpp).



(f) PICS sketch.



(g) PICS sketch reconstr.

*Figure 7.* Kodim02. For PIC/PICS, encoded prompt is “gayle chases eggs eggs knob withdrawn doors textures dewey red express u043D barns farcabs hauled”.

(a) Ground-truth.

(b) (Cheng et al., 2020) (0.016 bpp).

(c) HiFiC reconstr. (0.028 bpp).

(d) PIC reconstr. (0.0023 bpp).

(e) PICS reconstr. (0.012 bpp).

(f) PICS sketch.

(g) PICS sketch reconstr.

*Figure 8.* Kodim19. For PIC/PICS, encoded prompt is “confederflipkquid rated confederfemale decorate maine k adm giggs ubunseeks lighthouse accomgigab”.