Semantic Image Compression using Textual Transforms

Lara Arikan and Tsachy Weissman
Department of Electrical Engineering
Stanford University
Stanford, CA, USA
Emails: {arikan, tsachy}@stanford.edu

***Abstract*—Image textual transforms can be much smaller than JPEGs with comparable degrees of semantic similarity to the original image. Using human semantic satisfaction scores, we demonstrate that the highest-performing textual transforms are often rated similar to JPEGs at both lower (80% by size) and higher (90% by size) degrees of compression, though the captions are orders of magnitude smaller than the smallest JPEG. AI-based captioners are competitive with humans in textual transform rate-semantic distortion tradeoffs. We compare human captions to those of two AI models (BLIP and GPT-4), accounting for human perceptions of specific semantic content and affect elements in the original and reconstructed images. GPT-4 captions are shorter on average than human captions, and also capture similar semantic elements and achieve similar semantic fidelity to the original image. Our results recommend textual transforms as a semantic compression method with better rate-semantic distortion performance than traditional methods. We look forward to specialized semantic loss functions to optimize end-to-end image captioning and reconstruction models.**

I. Introduction

Image compression succinctly represents the information contained in images for efficient storage and transmission. Lossless compression devises invertible encodings preserving all of this information, while lossy compression strives to retain the information needed by the user of the decompressed image.

We consider the information in an image to consist of "visual" and "semantic" information. Visual information, such as the position, hue, or intensity of each pixel, is *denotative*. Semantic information is *connotative*: it emerges from the relative positions of the pixels, by which they accumulate into identifiable elements.

The Joint Photographic Experts Group (JPEG) standard [1] is a lossy image compression algorithm which excises a subset of the visual information (such as color frequencies and sharp hue or intensity transitions) through frequency-domain quantization. At higher resolutions, JPEGs mainly excise visually imperceptible information and are practically indistinguishable from much larger raw images.

Image captioning is a *textual transform* which preserves different details of image content and affect at different resolutions [2]. Bhown et al. have demonstrated that humans can recreate the content of images by following another person's written instructions [3], [4]. We regard these instructions as an image caption in exhaustive detail, affirming the hypothesis that a sufficiently advanced textual transform of an image can store a basis set of its semantic content and affective qualities, from which a semantically equivalent image can be reconstructed.

Our investigation seeks to demonstrate the usefulness and feasibility of automated and optimized textual transform compression schemes for extremely low-rate use and transmission of semantically equivalent images. To motivate their automation and optimization, we address the following questions:

1) Can images recreated from textual transforms be viable alternatives to JPEGs in terms of their semantic similarity to the original image?
2) Can images recreated from AI-generated captions provide an alternative to those based on human captions in terms of their semantic similarity to the original image?

By answering these questions, we wish to verify that textual transforms can truly provide similar levels of semantic distortion to the current lossy standard, and at massively lower rates. We also wish to determine whether current AI captioning capabilities can supplant human effort for practical use of textual transforms in compression. In the process, we aim to identify those elements of content and affect which contribute to varying degrees of human satisfaction regarding the semantic similarity of original and reconstructed images. This is intended to inform the training of end-to-end AI image captioning and reconstruction pipelines for efficient textual transform compression.

II. Related Work

In the context of image compression, textual transforms require an image-to-text (captioning) and text-to-image (reconstruction) pipeline. Since we are interested in the potential for high-performing textual transforms using AI image captioning and reconstruction, we review the related frameworks:

A. Image captioning.

Novel image-to-text methods are often grounded in deep learning, which helps manage the complexity of harmonizing image content with text syntax and semantics. Unsupervised attempts [5], [6] reduce the cost of sourcing image-sentence datasets and make the caption length and content more flexible.

However, much research continues in supervised deep learning captioners, including those explicitly trained to consider semantic content [7], [8].

Instead of these more semantically specialized models, we choose Bootstrapping Language-Image Pre-training (BLIP) [9] and Generative Pre-Trained Transformer 4 (GPT-4) [10] as our captioners in this paper. A simple justification is that the two methods work at very different resolutions; BLIP captions are much less detailed (see Fig. 2 as a reconstruction of Fig. 1), while GPT-4 captions are often as complete as human ones and permit reconstructions that reflect this completeness (see Fig. 3). Section IV (Methods) provides further explanation of this choice, and of the functioning principles in BLIP and GPT-4.

B. Image reconstruction.

Many recent text-to-image tools are founded on Contrastive Language-Image Pre-training (CLIP) [11], which learns text and image embeddings in a joint space. These embeddings can be leveraged by various methods to transfer textual into visual semantics. Such methods include generative adversarial networks (GAN) [12], [13] and diffusion models [14], [15]. DALL-E 2 [16] is a prime example of a successful CLIP-diffusion pairing.

Transformers have opened up new possibilities for image generation [17], sometimes in combination with diffusion techniques [18] and, at times, also with CLIP [19]. Although some of these advances report greater success than DALL-E 2, we prefer to use DALL-E as our reconstruction method because of its accessibility and popularity, and because of its continuity with GPT-4 (another OpenAI product).

C. Semantic information.

Definitions of semantic information have ranged over the past decades from logical-foundational [20], [21], [22] to practical or task-oriented [23], [24], [25]. However, none of these definitions have been accepted as canonical. In the absence of such a unified measure of semantic information, we find sufficient a general description of how semantic content emerges from technical content, separating semantic from visual information without quantifying semantic information along the compression pipeline.

Our definition distinguishes the technical content of the image (visual information) from the conceptual content (semantic information). Conceptual content is emergent from technical content; a picture of a table is made up of its pixels, and does not exist without them. In this paper, the word *content* will exclusively reference conceptual content. Technical content will henceforth be referred to only as *visual information* for clarity. The effect on the viewer of these concepts and their interactions will be referred to as *affect*.

D. Semantic similarity.

Semantic similarity measures can operate inter- or intramodally [26], [27], comparing concept or token embeddings [28], taxonomies [29], or image contexts [30], among other methods. Semantic similarity (or distortion) measures therefore suffer from the same multiplicity of definition as semantic information, leaving room for fundamentalist interpretations such as ours. We use human perceptions of the closeness of two images in their content and affect as our main indication of semantic similarity, as well as our basis for a finer decomposition of the elements that constitute content and affect. This is based on the concept of perceptual distortion [31] which in a non-semantic context is the foundation of many lossy compression schemes [32], [33] - including JPEG [34], [35], [36].

E. Semantic compression.

Finally, we consider alternative semantic image compression methods. Image captions can be combined with useful side information, such as the "compressed spatial conditioning map" in [37], to more exactly preserve semantic features like object or element arrangement and orientation. It has been proposed that images can be preserved as cascading hierarchical representations of extracted semantic features [38]. Deep learning has also been integrated into semantic compression, facilitating end-to-end training of semantic compression pipelines where the intermediate representation is not textual, but rather a semantic feature coding or embedding [39], [40], [41].

III. METHODS

We choose three images from five different categories to investigate the following content questions:

1) **One central object.** Can captions preserve the perceived *centrality* of an object (its obvious and superior importance to the viewer's understanding of the image compared to any other object in its surroundings)?
2) **Multiple scattered objects.** Does the caption create false centrality for any object?
3) **One single person.** What features of a human face and body does the caption communicate?
4) **Group of people.** Can the caption convey tone and geometry in the relations between people?
5) **Landscape.** Can the caption represent a distant and general environment without any central objects?

All chosen images are presented in our Github repository (https://github.com/lara-arikan/textualtransform24), sorted by category. We compress each image to .jpegs at lower (reducing filesize by 90%) and higher (reducing filesize by 80%) operating points, and caption it using three methods:

1) **Human captioning.** The human writes a sentence including these ordered elements if relevant:
    a) central or foreground objects with attributes (color, shape, appearance, expressions, disposition)
    b) geometry of central or foreground objects (side by side, in a group...)
    c) action with attributes of central or foreground objects (standing stiffly, staring sweetly...)
    d) immediate context (by a table, on a road, on a bench, at a counter...)
    e) background context (in a forest, in a schoolyard, in a train station...)

f) temporal context (on a bright summer morning, in red evening light...)

An example caption for the dog in Fig. 1a would be **A fluffy German shepherd, mouth open and tongue hanging, prancing merrily on a road in a forest in the afternoon light.**

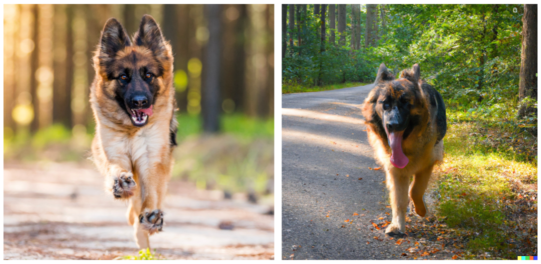


Fig. 1. **(a) Left:** Prancing dog. Image credit to Charlotte Reeves Photography. **(b) Right:** DALL-E 2 reconstruction of human caption of (a).

2) **Bootstrapping Language-Image Pre-training (BLIP).** Devised by Li et al [9], BLIP encodes image features using a visual transformer and pre-trains the model using a multimodal mixture of encoder-decoder. We chose BLIP because it is appropriate for "understanding-based, generation-based tasks"; because its relation to a well-studied language model such as BERT enhances its interpretability; and because it is readily available for use as part of the Library for Language-Vision Intelligence maintained by Salesforce.

   BLIP captions are often much shorter and less detailed than human captions in our syntax. For instance, Fig. 1a is described as **a dog running down a dirt road in the woods.**

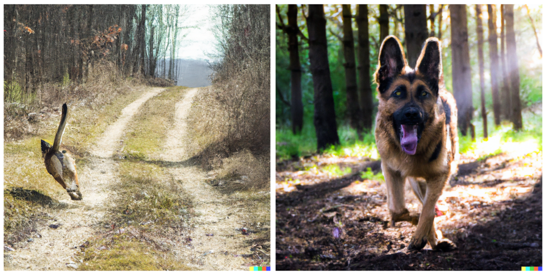


Fig. 2. **BLIP** (left) and **GPT-4** (right) captions of Fig. 1a, reconstructed by DALL-E 2.

3) **Generative Pre-Trained Transformer (GPT-4)**. OpenAI's GPT-4 [10] can both caption and generate images, and so is an excellent candidate for further experiments in textual transform compression- decompression pipelines. It also includes a near-human level of detail in its captions, which are replete with lavish adjectives. To curb its poetic inclinations, we prompted GPT-4 to caption each image "in a way DALL-E 2 would understand." Fig. 1a received the GPT-4 caption **A German Shepherd dog running directly towards the camera with a joyful expression, its tongue out, in a sunlit forest setting with light filtering through the trees creating a bokeh effect in the background.**

These 45 captions were reconstructed using DALL-E 2, and their performance was evaluated using three surveys:

1) Survey A: **Semantic Satisfaction Survey.** Quantifies the semantic similarity of semantically or classically decompressed images (i.e. reconstructions or JPEGs) to the original on a scale of 1-10 (*totally dissimilar* to *identical*).
2) Survey B: **Comparative Semantic Content Survey.** Quantifies on a scale of 1-5 the similarity of original and reconstructed images in terms of **(a)** their objects, elements and people, **(b)** the appearance thereof, **(c)** the positioning and orientation thereof, **(d)** the surroundings thereof, **(e)** their color scheme.
3) Survey C: **Detailed Semantic Content Survey.** Decomposes the content and affect of each reconstruction, asking after **(a)** the objects, elements and people (free response), **(b)** the relationship between the two "most important" elements as determined by the viewer (free response), **(c)** three adjectives that best describe the image (drop-down selection).

These three surveys were launched on Mechanical Turk and received 20 responses per question, totaling 1500 question responses for Survey A, 7500 for Survey B, and 2700 for Survey C. The questions as presented to survey takers, with exact wording and sample responses, can be found on our Github repository.

To compare our experimental results with a well-known similarity metric, we used the cosine similarity between CLIP embeddings of original and reconstructed or compressed images, scaled to lie between 0 and 1.

IV. Results and Discussion

A. *Viability of textual transforms.*

Fig. 3 plots human estimation of semantic similarity between the reconstructed and original images, against their compressed size. Because Mechanical Turk data is inherently noisy, the semantic similarity ratings in Survey A vary greatly (with almost every reconstruction receiving a score close to 1, and a score close to 10). We therefore use the median semantic similarity rating for every original-reconstructed or original-decompressed image pair as our semantic satisfaction index. The compressed image sizes for textual transforms are simply the 8-bit Unicode encoding length of the captions, on the order of tens or hundreds of bytes. For the .jpegs, we use their local storage filesizes, which constitute some tens or hundreds of kilobytes.

We observe that the textual transforms of no image achieve a median semantic similarity higher than 8.5 out of 10. However, the magenta iso-satisfaction curve in Fig. 3 shows that in certain cases, a textual transform using GPT-4 captioning is two orders of magnitude smaller than the smallest .jpeg which yields a median satisfaction of 8.5. Furthermore, .jpegs which

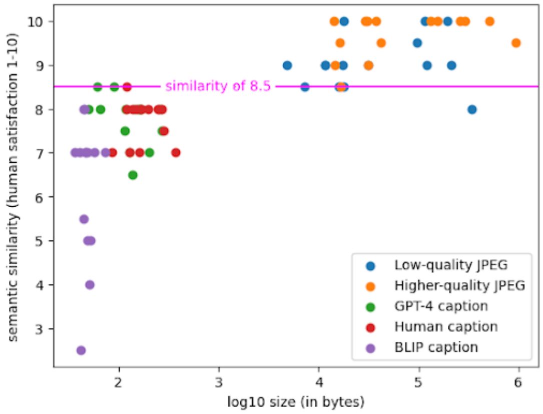


Fig. 3. Median semantic similarity rating (1-10) against compressed size, by compression method.

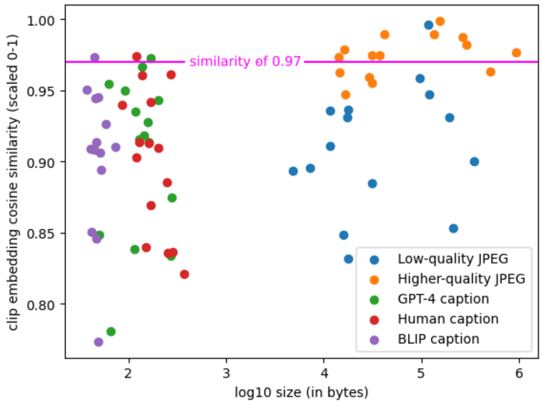


Fig. 4. CLIP embedding cosine similarity between original and reconstructed or compressed images, scaled to fall between 0-1.

surpass this value range up to four orders of magnitude larger than the highest-performing textual transforms.

So far, our analysis has cross-compared transforms and compressions on different images, aiming for a general evaluation of their relative performance. We now compare them in the case of a single image. The human caption of "Object 3" is, at 120 bytes, the largest textual transform achieving a median semantic similarity of 8.5 (with standard deviation 1.94). At 14,134 bytes, the higher-quality .jpeg of "Object 3" is the smallest .jpeg that achieves the highest possible median semantic similarity of 10 (with standard deviation 1.6). We therefore choose this image as an eminent example of the promise of textual transforms: that they can stand at 99% smaller than the highest-performing .jpegs with only a 15% loss of semantic satisfaction, as measured by human raters.

The importance of this result is difficult to determine absolutely, as a 15% loss in similarity may far outweigh the 99% size-savings for certain applications. For instance, a mother might value a slightly blurry JPEG of her own child's face far above some clear reconstruction of its description, such as "a baby playing and laughing on a red bedspread." Further, a different similarity metric might show greater distortion.

To address this issue, we show that CLIP embedding similarity between the original and reconstructed or compressed images follows the same pattern. Fig. 4 shows that textual transforms regularly perform as well as JPEGs, often achieving over 90% similarity. As expected, the highest textual transform similarity (97.4%) is for a human caption of a single object - the simplest semantic setting - while the lowest, at 77.3%, is a BLIP caption of a group of people. Textual transforms are competitive with traditional lossy compression methods in rate-semantic distortion performance for CLIP embedding similarity, indicating our results may well be robust to other semantic distortion metrics.

This conclusion is generalizable across different types of images according to Fig. 5, which colors by category the median similarity score of each image. There is no obvious clustering of similarity scores to identify images of single objects as more suitable than any other type of image for meaning-preserving textual transforms. We can therefore extend our observations for Object 3 to other candidate images, and expect semantic compression methods to perform at relatively low loss of semantic satisfaction at significantly smaller file sizes.

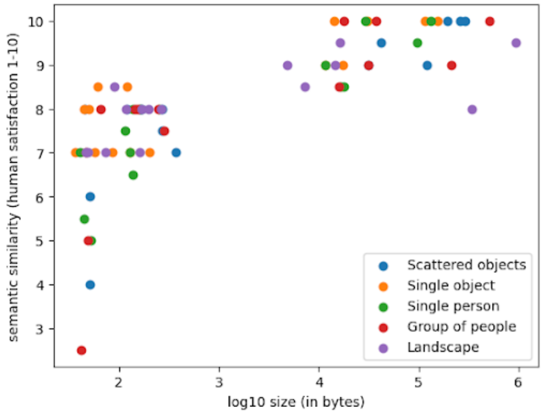


Fig. 5. Median Semantic Similarity-Compressed Size by Image Category.

Survey B provides more specific insight into the semantic dimensions along which different reconstructions and .jpegs may diverge from the original. Figures 1G-4G on our Github repository display the mean similarity on a scale of 1-5 between each original image and its reconstructions or .jpegs in terms of **(a)** their objects, elements and people, **(b)** the appearance thereof, **(c)** the positioning and orientation thereof, **(d)** the surroundings thereof, **(e)** their color scheme (as described in Methods).

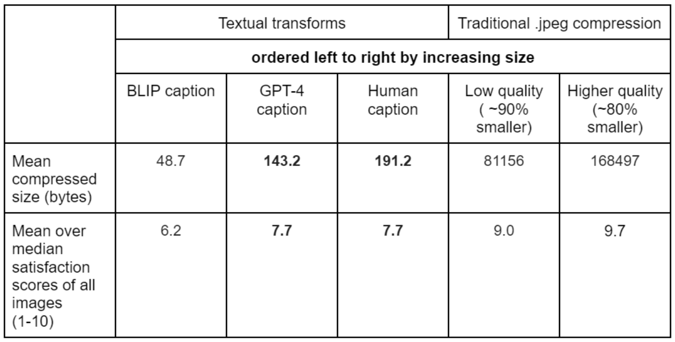
Along all the listed dimensions, the similarity between "Single object" images and their .jpegs and reconstructions is marginally higher than the similarity for other image categories. This is likely due to the simplicity of content in images of single objects, as well as a lack of complex affect associations. However, the score distributions generally overlap between categories, and the large standard deviations preclude strict judgments about category-specific captioning capacities.

The lowest similarity scores for any category correspond to colors in the original and reconstructed images, as captions often do not incorporate detailed information about the colors of each element.

B. Viability of AI-based textual transformers.

To address our second question, we compare the DALL-E reconstructions of captions generated with GPT-4; with BLIP; and with human effort in terms of how similar in feeling and meaning they are to the original image.

Table 1 shows that human captions are the longest on average, but they do not as a rule yield reconstructions more similar to the original image than those based on GPT-4 captions. In fact, the average of all median semantic similarity scores received by human caption reconstructions is 7.7, the same as that of those received by GPT-4 caption reconstructions, even while GPT-4 captions are shorter. This means that textual transforms using AI captions perform just as well as those based on human captions, and at an even lower rate.



	Textual transforms			Traditional .jpeg compression	
	ordered left to right by increasing size				
	BLIP caption	GPT-4 caption	Human caption	Low quality (~90% smaller)	Higher quality (~80% smaller)
Mean compressed size (bytes)	48.7	**143.2**	**191.2**	81156	168497
Mean over median satisfaction scores of all images (1-10)	6.2	**7.7**	**7.7**	9.0	9.7

Table 1: Mean size and semantic satisfaction for five compression modes.

To inform future selection of reconstruction methods, we wish to gather more information about the transference of semantic information from caption to reconstruction. BLIP captions frequently yield reconstructions with the lowest semantic similarity to the original image. Since they are the shortest, we assume they excise perceptually significant semantic information. To identify what is excised, we turn to Survey C. Table 5G on our Github repository selects the three most frequent answers to each question in Survey C for each captioning method. We collapse conceptually identical answers to a single phrase, convert plurals to singular concepts, and examine the overlap between answers for the three methods.

As a percentage of distinct answers, we see that the most central objects in reconstructions from GPT-4 and human captions overlap 42% of the time, identical to the 42% of those from GPT-4 and BLIP and barely lower than the 45% of those from BLIP and human captions. Identified relations between these objects are the same for all methods, while adjectives describing the tone of each reconstruction overlap even more frequently for BLIP and GPT-4 reconstructions (89%) than GPT-4 and human caption reconstructions (73%). This suggests that semantic similarity is significantly determined by details whose presence is not queried by Survey C, which are nevertheless omitted from BLIP captions. Of the semantic components in our human captions, these determinants of similarity may be the attributes of the central objects and their relations, or their immediate, background or temporal context.

Finally, we acknowledge the possibility of biased or skewed results in our study, particularly along the GPT-4 to DALL-E pipeline. Our images were sourced from internet searches using category labels as query terms. Either GPT-4 or DALL-E may therefore have been trained on the same images as they were asked to caption or reconstruct in our paper. Future work might verify these conclusions with privately sourced images.

V. Conclusion

Our results suggest that textual transform based reconstructions can provide comparable semantic satisfaction at far lower rates than lossy compression standards like JPEG. This encourages the development of semantic compression pipelines using carefully selected textual transforms.

In this vein, we demonstrate that AI-based textual transforms can perform comparably to human captions in the preservation of semantic information, as in the case of GPT-4. They are able to include the same semantic components (the same elements, in similar orientations, with similar appearance) and to have the same effect on the viewer's perception (a similar tone and feeling).

We further find that GPT-4 captions are preferable to human captions because they are shorter on average. This may be because AI usually omits the level of detail a human captioner feels compelled to include, though the inclusion of those details does not measurably increase semantic satisfaction after reconstruction.

AI permits unique improvements in the realm of textual transform coding. It opens the possibility for end-to-end training of an encoder-decoder network that could optimize for the most concise textual representation of an image that would contain all the elements its own decoder would need to create a faithful reconstruction, and would remain interpretable to human readers. If interpretability were not a concern, the intermediate semantic representations could turn out to be even more compressible.

The next steps of this study might aim to integrate more formal definitions of "semantic information" and "semantic fidelity" into cost functions for optimized textual transform generation. One might also explore different architectures that make the best use of these functions, balancing the computational complexity introduced by high-performing AI captioners, and opening textual transform coding up for practical use.

Acknowledgments

We thank the participants of the 2023 SHTEM program at Stanford University, Sitara Mitragotri, Rini Khandelwal, Jessica Kang, Kayla Le and Adit Chintamaneni, for their work in image selection and captioning. We further thank Dr. Pulkit Tandon for his assistance with our Mechanical Turk experiments.

References

[1] G. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[2] T. Weissman, "Toward textual transform coding," *arXiv preprint arXiv:2305.01857*, 2023.

[3] A. Bhown, S. Mukherjee, S. Yang, S. Chandak, I. Fischer-Hwang, K. Tatwawadi, and T. Weissman, "Humans are still the best lossy image compressors," in *2019 Data Compression Conference (DCC)*, 2019, pp. 558–558.

[4] A. Bhown, S. Mukherjee, S. Yang, S. Chandak, I. Fischer-Hwang, K. Tatwawadi, J. Fan, and T. Weissman, "Towards improved lossy image compression: Human image reconstruction with public-domain images," 2019.

[5] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] R. Yang, X. Cui, Q. Qin, Z. Deng, R. Lan, and X. Luo, "Fast rf-uic: A fast unsupervised image captioning model," *Displays*, vol. 79, p. 102490, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141938223001233

[7] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] T. Xian, Z. Li, Z. Tang, and H. Ma, "Adaptive path selection for dynamic image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5762–5775, 2022.

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[10] OpenAI, :, J. Achiam, and others., "Gpt-4 technical report," 2023.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[12] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "Vqgan-clip: Open domain image generation and editing with natural language guidance," 2022.

[13] M. Tao, B.-K. Bao, H. Tang, and C. Xu, "Galip: Generative adversarial clips for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 214–14 223.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021. [Online]. Available: https://arxiv.org/abs/2112.10752

[15] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative ai: A survey," 2023.

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

[17] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *CoRR*, vol. abs/2102.12092, 2021. [Online]. Available: https://arxiv.org/abs/2102.12092

[18] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022.

[19] Z. Wang, W. Liu, Q. He, X. Wu, and Z. Yi, "Clip-gen: Language-free training of a text-to-image generator with clip," 2022.

[20] Y. Bar-Hillel and R. Carnap, "Semantic information," *The British Journal for the Philosophy of Science*, vol. 4, no. 14, pp. 147–157, 1953.

[21] J. Hintikka, "On semantic information," in *Information and inference*. Springer, 1970, pp. 3–27.

[22] L. Floridi, "Is semantic information meaningful data?" *Philosophy and phenomenological research*, vol. 70, no. 2, pp. 351–370, 2005.

[23] X. Li and D. Roth, "Learning question classifiers: the role of semantic information," *Natural Language Engineering*, vol. 12, no. 3, pp. 229–249, 2006.

[24] H. Zhang, H. Wang, Y. Li, K. Long, and V. C. Leung, "Toward intelligent resource allocation on task-oriented semantic communication," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 70–77, 2023.

[25] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.

[26] Z. Parekh, J. Baldridge, D. Cer, A. Waters, and Y. Yang, "Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco," *arXiv preprint arXiv:2004.15020*, 2020.

[27] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5585–5599, 2018.

[28] J. Choi, M. Cho, S. H. Park, and P. Kim, "Concept-based image retrieval using the new semantic similarity measurement," in *Computational Science and Its Applications—ICCSA 2003: International Conference Montreal, Canada, May 18–21, 2003 Proceedings, Part I 3*. Springer, 2003, pp. 79–88.

[29] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[30] V. Franzoni, A. Milani, S. Pallottelli, C. H. Leung, and Y. Li, "Context-based image semantic similarity," in *2015 12th international conference on fuzzy systems and knowledge discovery (fskd)*. IEEE, 2015, pp. 1280–1284.

[31] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proceedings of 1st International Conference on Image Processing*, vol. 2. IEEE, 1994, pp. 982–986.

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

[33] Y. Patel, S. Appalaraju, and R. Manmatha, "Deep perceptual compression," *arXiv preprint arXiv:1907.08310*, 2019.

[34] Z. Liu, L. J. Karam, and A. B. Watson, "Jpeg2000 encoding with perceptual distortion control," *IEEE transactions on image processing*, vol. 15, no. 7, pp. 1763–1778, 2006.

[35] C.-C. Chen and O.-C. Chen, "Region of interest determined by perceptual-quality and rate-distortion optimization in jpeg 2000," in *2004 IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 3. IEEE, 2004, pp. III–869.

[36] K. J. Han, M. A. Robertson, and B. W. Suter, "Image recompression and perceptual distortion analysis," in *Ultrahigh-and High-Speed Photography, Photonics, and Videography*, vol. 5210. SPIE, 2004, pp. 157–168.

[37] E. Lei, Y. B. Uslu, H. Hassani, and S. S. Bidokhti, "Text+ sketch: Image compression at ultra low rates," *arXiv preprint arXiv:2307.01944*, 2023.

[38] K. Liu, D. Liu, L. Li, N. Yan, and H. Li, "Semantics-to-signal scalable image compression with learned revertible representations," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2605–2621, 2021.

[39] S. Luo, Y. Yang, Y. Yin, C. Shen, Y. Zhao, and M. Song, "Deepsic: Deep semantic image compression," in *Neural Information Processing*, L. Cheng, A. C. S. Leung, and S. Ozawa, Eds. Cham: Springer International Publishing, 2018, pp. 96–106.

[40] M. Akbari, J. Liang, and J. Han, "Dsslic: Deep semantic segmentation-based layered image compression," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2042–2046.

[41] S. Luo, G. Fang, and M. Song, "Deep semantic image compression via cooperative network pruning," *Journal of Visual Communication and Image Representation*, vol. 95, p. 103897, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320323001475