# Rate-Distortion Optimization for Cross Modal Compression

Junlong Gao*, Chuanmin Jia*, Shanshe Wang*, Siwei Ma*†, and Wen Gao*†

*Peking University, Beijing, China †Peng Cheng Laboratory, Shenzhen, China.
{jlgao,cmjia,sswang,swma,wgao}@pku.edu.cn

## Abstract

Recently, cross modal compression (CMC) is proposed to compress highly redundant visual data into a compact, common, human-comprehensible domain (such as text) to preserve semantic fidelity for semantic-related applications. However, CMC only achieves a certain level of semantic fidelity at a constant rate, and the model aims to optimize the probability of the ground truth text but not directly semantic fidelity. To tackle the problems, we propose a novel scheme named rate-distortion optimized CMC (RDO-CMC). Specifically, we model the text generation process as a Markov decision process and propose rate-distortion reward which is used in reinforcement learning to optimize text generation. In rate-distortion reward, the distortion measures both the semantic fidelity and naturalness of the encoded text. The rate for the text is estimated by the sum of the amount of information of all the tokens in the text since the amount of information of each token is a lower bound of coding bits. Experimentally, RDO-CMC effectively controls the rate in the CMC framework and achieves competitive performance on MSCOCO dataset.

## Introduction

With the tremendous increase of visual data and the development of visual data analysis and understanding, a variety of compression frameworks are proposed to preserve semantic fidelity to reduce the bitrate as much as possible, rather than signal fidelity. These frameworks compress the visual signal to semantic information to achieve a high compression ratio for semantic-related applications (such as machine analysis, semantic monitoring, and human-centered applications), including video coding for machine (VCM) [1], cross modal compression (CMC) [2], etc. VCM [1] reduces the bitrate as much as possible to ensure semantic fidelity for machine analysis. However, these frameworks mainly adopt the feature of the neural network as the compression domain, which is not human-comprehensible and further analysis is necessary for semantic-related applications. Moreover, the feature is mostly task-specific and thus difficult for multi-task analysis. Recently, CMC [2] focuses on compressing visual data to a compact, common, and human-comprehensible domain (such as text, sketch, semantic map, attributions, etc.) to preserve semantic fidelity.

CMC adopts text representation to preserve semantic fidelity, and the framework is composed of a CMC encoder, lossless coding, and CMC decoder, where the encoder encodes the image to text, the lossless coding interconverts the text with a bitstream, and the decoder decodes the text to the image. However, CMC is with a constant rate since the encoder can only represent the data with the text of a fixed grain, which is converted to a bitstream of a constant rate. But in practice, variable rate is necessary

Chuanmin Jia is the corresponding author.

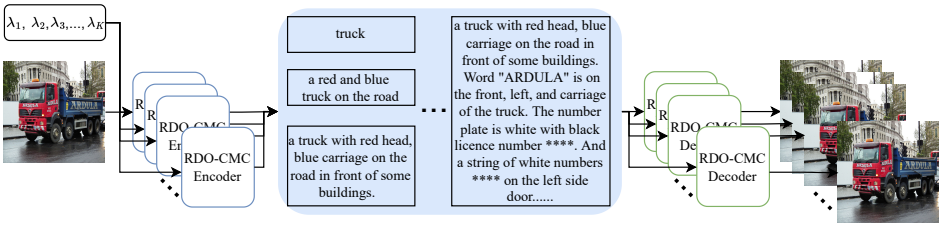


Figure 1: Illustration of RDO-CMC. RDO-CMC aims to encode the image to the text and reconstruct the images of different levels of semantic fidelity with different $\lambda$s.

due to the different demands of various transmission bandwidths, storage mediums, and levels of application requirements. For example, as shown in Fig.1, an image can be encoded into thousands of words to elaborate the scene with higher semantic fidelity or one word to illustrate the main object class with lower semantic fidelity. Moreover, since the objective for the encoder is cross entropy loss that maximizes the probability of the ground truth (GT) text, the text generation is not directly optimized for semantic fidelity, and the generated text will be less effective for CMC.

Rate-distortion optimization (RDO) is introduced into compression frameworks to achieve variable rate, where a hyper-parameter $\lambda$ is adopted to govern the trade-off between the rate and distortion. Such a paradigm lowers the distortion with the increase of the rate and vice versa. Incorporating RDO with CMC can achieve variable rate, but is non-trivial, since CMC regarding text representation as a compression domain is different from the previous frameworks and has some characteristics. (1) CMC is proposed for preserving semantic fidelity for compression, rather than signal fidelity, and thus the distortion should consider the semantic fidelity. (2) The encoded text should be grammatically correct enough as natural language. (3) The rate of text representation is required to be estimated.

In this paper, we propose the first work for rate-distortion optimized CMC (RDO-CMC) and achieve the characteristics, as shown in Fig. 1. Specifically, we model the text generation process as a Markov decision process (MDP) and propose rate-distortion reward which is used in reinforcement learning (RL) to optimize text generation. To combine semantic fidelity and naturalness, we propose text-based and image-based distortion. To estimate the rate of the encoded text, we calculate the amount of information of the tokens in the text which is the lower bound of coding bits for each token, such that the optimization can be independent of various arithmetic codings. Training with different $\lambda$s, RDO-CMC can control the rate of text representation and represent the data with different grains and satisfy the characteristics discussed above. Experimental results demonstrate our proposed model achieves variable rate and competitive compression performance on MSCOCO dataset.

## Related Work

**Compression frameworks:** Traditional block-wise image/video compression frameworks are widely used with a series of industry standards, such as JPEG [3], JPEG2000 [4] and BPG [5]. Recently, learning-based frameworks have been proposed to compress the images via end-to-end optimization [6, 7]. These frameworks introduce RDO

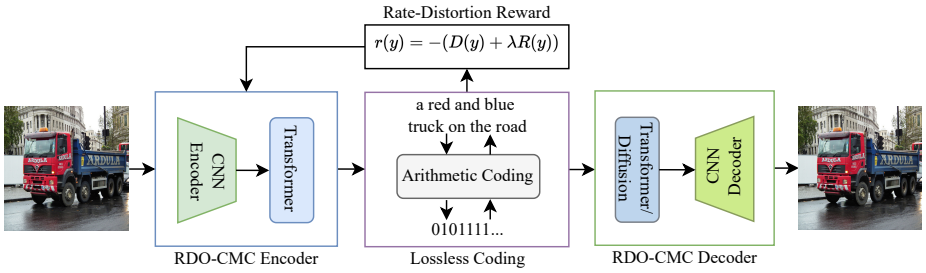


Figure 2: Architecture of our proposed RDO-CMC. The rate-distortion reward for a sampling text is calculated to train the RDO-CMC encoder.

to support variable rate and aim to improve signal fidelity and can be extended to optimize semantic fidelity for machine vision [1]. However, the compression domain of these frameworks is the feature of the neural network or bitstream and thus is not human-comprehensible. CMC [2] is proposed for the human-comprehensible compression domain since the text is formed by human language. However, CMC cannot achieve variable rate, which makes them less practical.

**Image and text transfer:** Image and text transfer is composed of image-to-text generation and text-to-image generation. Image-to-text generation aims to generate natural language for images, including image captioning, which generates a sentence to describe the scene in the image. In this field, some works use RL to address the exposure bias and optimize the non-differentiable evaluation metrics [8]. These works are based on REINFORCE algorithm [9] and introduce different kinds of baselines to stabilize the training. Text-to-image generation is developed by leaps and bounds recently. Many previous works trained the GAN network to produce text-conditional image samples [10]. Some works adopted VQ-GAN [11] to tokenize the image to code sequence, trained the model with transformers via the sequence-to-sequence manner, and produced the image codes in an autoregressive way [12]. Here, the code sequence of the image is far less than the sequence length of raw pixels and thus more effective to model text-to-image transfer. Recently, some works also applied diffusion models for this task using image code sequence [13].

## Rate-Distortion Optimization for Cross Modal Compression

The CMC framework compresses the image $x$ to a sequence of tokens $y = \{y_1, ..., y_N\}$ that is compact, common, and human-comprehensible, where $N$ is the length of text, $y_i \in A$, and $A$ is the vocabulary. Such a framework can neither achieve variable rate nor optimize the text representation directly for semantic fidelity. To tackle the problems, we optimize the rate-distortion trade-off via RL. In the following, we describe the model architecture and RDO framework with RL.

**Model architecture:** Li et al. [2] employ Show-and-tell [14] as the encoder and AttnGan [10] as the decoder. However, both of them are limited due to small model capacities and training datasets. Recently, large-scale multimodal models achieve spectacular development, including image-to-text and text-to-image generation. Such development makes the effective encoding and decoding process of CMC possible. In this paper, we employ multimodal models as the RDO-CMC encoder and decoder

to generate high-quality text and images. The model architecture of RDO-CMC is detailed as follows and shown in Fig.2.

First, in the RDO-CMC encoder, the feature map of $x$ is extracted from a CNN encoder and flattened as the input of the transformer to generate text $y$. The learning objective of the encoder is cross entropy loss as follows,

$$\mathcal{L}_{enc} = -\sum_{i=1}^{N} \log p_\theta \left( y_i \mid y_{<i}, x \right), \tag{1}$$

where $\theta$ refers to the parameters of the RDO-CMC encoder. Such an objective optimizes the sum of the log probabilities of the ground truth text and facilitates the encoder to generate text similar to the ground truth during inference.

Second, the lossless coding interconverts the text with the bitstream. We adopt Huffman coding [15] following [2]. A major difference lies in that they regard the letter as Huffman code and transform each letter in the text to coding bits, while we regard the token as Huffman code, where a token contains one or many letters. As such, the total coding bits of text will decrease since the token as Huffman code lowers the correlation between codes in the code sequence. Finally, the RDO-CMC decoder reconstructs the raw image from the text. Motivated by recent text-to-image works [11–13], our RDO-CMC decoder first predicts the latent representation using diffusion-based methods [13] or the image code sequence using transformer-based methods [12] with the input text, then generate the image raw pixels using a CNN decoder. The RDO-CMC encoder and decoder are pretrained separately, then we perform RDO for the encoder with RL to achieve variable rate and directly optimize semantic fidelity. A pretrained model will stabilize the training of RL since the pretrained model will reduce the action searching space and converge faster [8].

**Rate-distortion optimization with reinforcement learning:** In the text generation process, tokens are generated sequentially. Regarding generating a token as taking an action, text generation can be modeled as a Markov decision process (MDP), which can be optimized by RL. In this paper, we adopt RL to optimize the text generation with rate-distortion reward during the encoding stage, as shown in Fig. 2. An RL agent interacts with an external "environment" (words and image features) to determine the best text.

Formally, the text generation process can be viewed as an MDP process, including five elements $\{S, A, P, Re, \gamma\}$, where $S$ is a state space, $A$ is an action space as well as the token vocabulary, $P\left(s_{i+1}|s_i, y_i\right)$ is state transition probability, $Re(s_i, y_i)$ is reward function and $\gamma \in (0, 1]$ is the discounted factor. The whole RDO-CMC encoder with parameter $\theta$ can be viewed as the RL agent. The agent selects an action from a probability distribution $\pi\left(y\,|s\right)$ called policy, that corresponds to generating a token. The state $s_i \in S$ is considered as a list composing of the image feature $X$ and the tokens/actions $\{y_1, ..., y_{i-1}\}$ generated so far, i.e. $s_i = \{X, y_1, ..., y_{i-1}\}$. The agent interacts with the environment as follows: the agent observes the current state $s_i$ and selects an action, the environment returns a reward $r_i$ to the agent, and the agent receives the reward and the state is transferred to the next state $s_{i+1}$. However, in text generation, a reward $r(y) = Re(s_N, y_N) = Re\left(y_{1:N}\right)$ is not obtain until EOS token is

generated, and $r(y)$ is determined by the optimization goal of MDP. Therefore, we define the reward for each action as follows:

$$r_i = \begin{cases} 0, i < N \\ r(y), i = N. \end{cases} \tag{2}$$

In RL, the agent aims to maximize the cumulative reward $L(\theta) = E_\pi \left[\sum_{i=1}^{N} \gamma^{i-1} r_i\right]$, estimate the gradient $\nabla_\theta L(\theta)$, and update its parameters. In REINFORCE algorithm [9] incorporating with a baseline, the gradient $\nabla_\theta L(\theta)$ is approximated with a sample:

$$\nabla_\theta L(\theta) \approx (r(y) - b) \nabla_\theta \log \pi_\theta(y), \tag{3}$$

where $\gamma = 1$ and $b$ is the baseline to stabilize the training. The optimization goal of MDP is to maximize the cumulative reward. Since we aim to minimize the rate-distortion trade-off, we set the reward as the negative rate-distortion trade-off,

$$r(y) = -(D(y) + \lambda R(y)), \tag{4}$$

where $D(y)$ is the distortion resulting from the text representation $y$ as the encoded information, and $R(y)$ is the rate of $y$.

**Distortion estimation:** Since CMC focuses on semantic fidelity, the distortion should consider semantic fidelity, i.e. the distortion should be able to measure the amount of semantic information being encoded to text representation and decoded to the reconstructed images. Moreover, since the compression domain of CMC is natural language, the distortion should also consider the "distortion" of the naturalness of the text representation. Therefore, considering both the amount of semantic information and the naturalness, we propose two ways of estimating the distortion, namely text-based and image-based semantic distortion. As for text-based distortion, we estimate the distortion based on the generated text using SPICE [16]. SPICE is an evaluation metric to measure the similarity of semantic information between the generated text $y$ and GT text $y_{GT}$. Moreover, SPICE correlates with human judgments in terms of text quality [16]. Therefore, we calculate the distortion $D(y)$ of $y$ as $D(y) = -SPICE(y, y_{GT})$. As for image-based distortion, we estimate the distortion based on the reconstructed image. Since we focus on semantic distortion, we calculate the cos similarity of features of the decoded image $Dec(y)$ and GT image $x$, i.e. $D(y) = -SIM(Dec(y), x)$. In this paper, we adopt CLIP [17] to extract the image feature. Therefore, a higher SIM means that the generated text contains more semantic information to decode the image, and is natural and correct enough for a pretrained decoder to transfer knowledge. Note that since RDO-CMC is optimized directly for semantic fidelity, some distortion metrics measuring signal fidelity, such as mean squared error (MSE), are not appropriate for the distortion in this paper.

**Rate estimation:** Traditional block-wise compression frameworks estimate the rate using coding bits. Since coding bits of each token are lower-bounded by the amount of information, in this paper, we estimate the rate of each token by the amount of information. As such, the optimization can be independent of various types of lossless codings, such as Huffman coding [15], since the rate of text is measured by the amount of information during optimization, rather than the coding bits

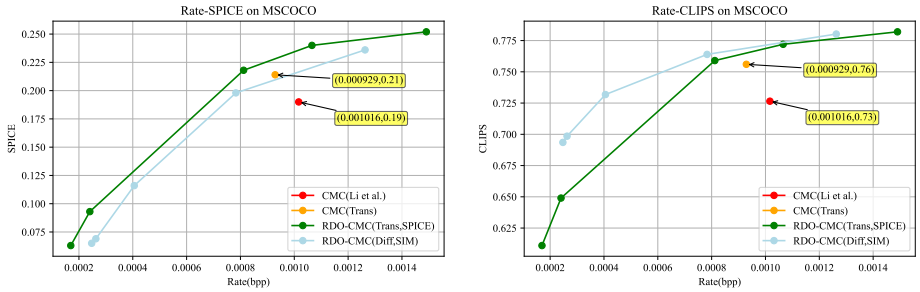


Figure 3: Rate-SPICE↑ and CLIPS↑ curves of the encoded text on MSCOCO dataset.

obtained by lossless coding. To calculate the amount of information of token $y_i$ in vocabulary $A$, we calculate the negative log base 2 of the probability of the token, i.e. $-\log_2(\Pr(y_i))$, where the probability is measured by the frequency $\Pr(y_i)$ in the training corpus. The amount of information of text $y$ should consider the amount of information of tokens in $y$ and their correlations. Here we simply calculate the sum of the amount of information of all tokens without considering their correlations, i.e. $R(y) = -\sum_{i=1}^{N} \log_2(\Pr(y_i))$. Note that lossless coding also adopts the frequencies of all tokens in vocabulary $A$ to convert tokens to a bitstream. Therefore, the rate definition during training is highly correlated with that during inference.

## Experiment

**Dataset:** We use MSCOCO [18] to evaluate our method. For MSCOCO, 82783 images/ 414113 text are used for training, and the Karparthy test split (5000 images) [19] is used to evaluate the method. The images are resized with the resolution of $256 \times 256$ following [2]. We find that short text is too rare. To better explore the rate-distortion trade-off, we truncate the text with 50% probability to broaden the text length range during training as a way of dataset augmentation. If the text needs to be truncated, we truncate the text after a randomly selected noun phrase.

**Metrics:** To evaluate the quality of the generated text, we use standard metrics, including SPICE [16] and CLIPS (short for CLIPScore) [20]. SPICE considers the similarity of the scene graph of the generated and GT text. CLIPS is the similarity of the CLIP feature of the generated text and GT images and focuses on global image-text alignment. To evaluate the quality of the reconstructed images, following CMC [2], we use two semantic fidelity metrics, namely IS [21] and FID [22], where IS measures the naturalness and the diversity of the generated images, and FID estimates the distribution distance between the input images and the generated images. We use the implementation of IS and FID using the released code in [12].

**Implementation details:** To boost the performance of RDO-CMC, the pretrained OFA [12] is base setting and adopted as the RDO-CMC encoder. The pretrained OFA and LDM [13] are both large settings and adopted as the RDO-CMC decoder, where the former is the transformer-based method and the latter is the diffusion-based method. We train the encoder with Eq. 1 from the pretrained checkpoint using the augmented dataset for 5 epochs and finetune the encoder with Eq. 3. Training with different $\lambda$s correspond to different RDO-CMC encoders, and we adopt

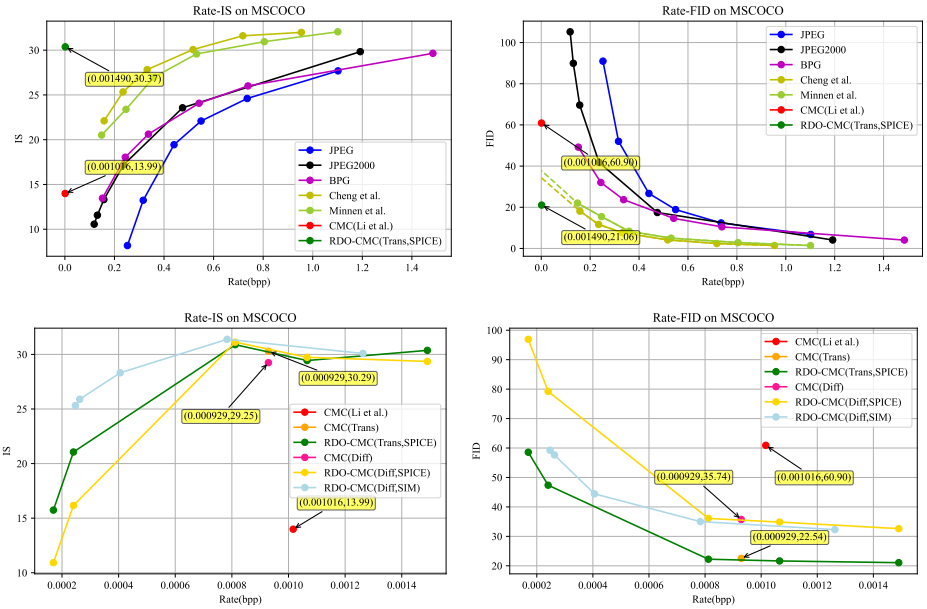


Figure 4: Rate-IS↑ and FID↓ curves of reconstructed images on MSCOCO dataset.

the same decoder for different $\lambda$s since the decoder is more general to the text of variable rate. The learning rate is initially set as $1e-5$, and the schedule is polynomial decay. The batch size is 40. We find that 500 training iterations are enough to find the trade-off. We compute $b$ as the average reward of another four sampling texts.

**Experiment settings:** The settings are listed as follows. **(1) Traditional codecs:** JPEG [3], JPEG2000 [4], BPG [5]. The rates are $Q = [1, 5, 10, 15, 25, 50]$ for JPEG, $Q = [100, 90, 75, 50, 25, 10]$ for JPEG2000, $Q = [50, 47, 45, 42, 40, 35]$ for BPG. **(2) Learning based codecs**: Minnen et al.[6] and Cheng et al.[7]. The rates are both $Q = [1, 2, 3, 4, 5, 6]$. All the traditional and learning-based codecs use the implementation[1]. **(3) CMC-based codecs**: We use the CMC framework with Show-and-tell [14] as the encoder and AttnGan [10] as the decoder, and train the model with the released code, which is denoted as CMC(Li et al.). Moreover, we also implement CMC with the architecture of RDO-CMC, which is denoted as CMC(Trans) for the transformer-based decoder and CMC(Diff) for the diffusion-based decoder. Both of them only have one rate. **(4) RDO-CMC based codecs(ours)**: RDO-CMC also has some variants, including RDO-CMC(Trans, SPICE), RDO-CMC(Diff, SPICE) and RDO-CMC(Diff,SIM), and is further trained with the rate-distortion reward. RDO-CMC(Trans,SPICE) and RDO-CMC(Diff, SPICE) utilize the text-based distortion SPICE, and RDO-CMC(Diff,SIM) utilizes the image-based distortion SIM. We do not perform RDO-CMC(Trans,SIM) since the transformer-based decoder generates images autoregressively and slowly, and is hard to optimize the decoder with SIM. The rates of the variants are $\lambda = [0.005, 0.004, 0.002, 0.001, 0]$.

**RDO behavior of encoded text:** We analyze the rate-distortion behavior of encoded text, compare the rate-distortion curves of RDO-CMC with CMC on SPICE and CLIPS in Fig. 3, and reach some conclusions. (1) Though RDO-CMC(Trans,SPICE) and RDO-CMC(Diff,SIM) use different distortions, the curves of them are incre-

[1]https://github.com/InterDigitalInc/CompressAI

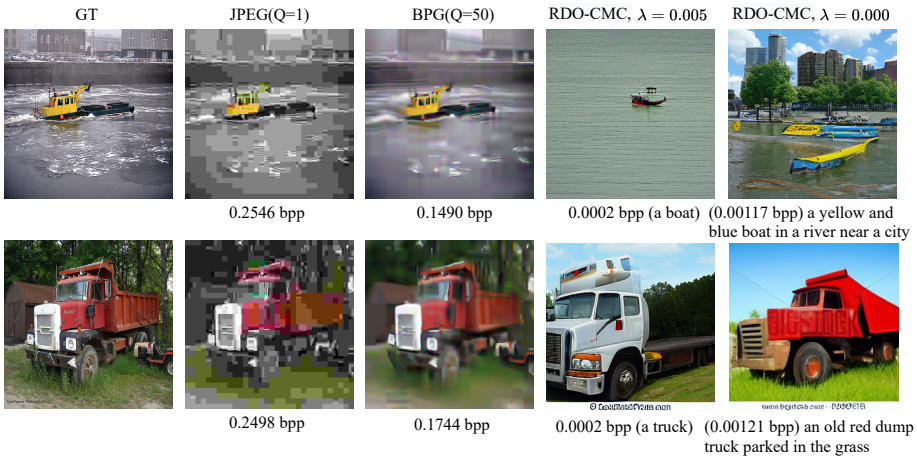


Figure 5: Qualitative results of Ground Truth image (GT), JPEG [3], BPG, and our RDO-CMC with different $\lambda$s on MSCOCO. We show the generated text and reconstructed images.

mental as the rate increases. Such a result indicates that optimizing the rate-distortion reward achieves the capability of variable rate of the encoded text. (2) RDO-CMC (Trans,SPICE) is better than CMC of both CMC [2] and CMC(Trans). The curve of RDO-CMC(Trans,SPICE) is above the results of CMC(Trans), indicating that RDO directly improves the semantic fidelity of the encoded text. Though RDO-CMC (Diff,SIM) is not optimized for SPICE, the result is also comparable to CMC(Trans). (3) Optimizing a distortion metric will improve the metric more than other metrics. Since SIM is calculated using the similarity of the CLIP feature of the generated and GT images, optimizing SIM is to implicitly optimize CLIPS. Thus RDO-CMC(Diff,SIM) is better than RDO-CMC(Trans,SPICE) on CLIPS, and RDO-CMC(Trans,SPICE) is better than RDO-CMC(Diff,SIM) on SPICE.

**Semantic fidelity of reconstructed image:** We compare the semantic fidelity of reconstructed images with IS and FID in Fig. 4, and reach some conclusions as follows. (1) The upper two figures in Fig. 4 show that the proposed RDO-CMC(Trans,SPICE) outperforms CMC[2] and all the traditional and learning-based codecs at the closing rate on MSCOCO dataset. Such a result also shows the great potential of the proposed method on semantic-level reconstruction with an ultra-high compression ratio. (2) The curves in the lower two figures of Fig. 4 are incremental, indicating that optimizing the rate-distortion reward achieves the capability of variable rate of the reconstructed images. Note that some fluctuations may exist in some metrics, such as IS, since these metrics are not directly optimized as distortion. (3) RDO-CMC(Diff,SIM) is better than CMC(Diff), CMC(Trans) is comparable to RDO-CMC(Trans,SPICE), and RDO-CMC(Diff,SIM) is better than RDO-CMC(Diff,SPICE). These results show that optimizing SIM, an image-based distortion, will improve the quality of the generated images. However, optimizing SPICE, a text-based distortion, may not enable RDO-CMC to generate better images than CMC. (4) RDO-CMC(Trans,SPICE) is better than RDO-CMC(Diff,SPICE). It may be because the transformer decoder is finetuned on MSCOCO and the diffusion is not, resulting in a better distribution distance of the former than the latter.

**Qualitative results:** We visualize some qualitative results in Fig. 5, where the upper two figures are generated by RDO-CMC(Trans,SPICE) and the lower two are generated by RDO-CMC(Diff,SIM). (1) We find that the images compressed by JPEG with Q=1 and BPG with Q=50 are of high bpp but still suffer from block artifact, and the semantic information in the image is hard to be recognized. Differently, RDO-CMC reconstructs the semantic information of the GT image, which shows the potential of semantic-level reconstruction and an ultra-high compression ratio of the CMC framework. (2) Different $\lambda$s of RDO-CMC show different levels of semantic fidelity in the encoded text and reconstructed images. The text can contain the main object as short as possible and the attribute, object, and scene as long as possible. The incremental semantic information can be reflected in the image.

## Conclusion and Discussion

In this paper, we propose RDO-CMC to achieve variable rate and directly optimize semantic fidelity. We propose rate-distortion reward with RL and achieve competitive performance on MSCOCO. Though some encouraging results are achieved, some effort can make CMC more practical. For example, a more efficient text encoder will realize high semantic fidelity in text, which benefits some semantic-related applications. A more powerful image generator to draw all semantic information in the text to the image will make RDO-CMC more realistic.

## Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under grant 62101007, 62025101, and in part by the High Performance Computing Platform of Peking University, which are gratefully acknowledged. Assistance for the development of this research work provided by Dr. Jiguo Li was greatly appreciated.

## References

[1] Lingyu Duan, Jiaying Liu, Wenhan Yang, Tiejun Huang, and Wen Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Transactions on Image Processing*, vol. 29, pp. 8680–8695, 2020.

[2] Jiguo Li, Chuanmin Jia, Xinfeng Zhang, Siwei Ma, and Wen Gao, "Cross modal compression: Towards human-comprehensible semantic compression," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4230–4238.

[3] Gregory K Wallace, "Overview of the jpeg (iso/ccitt) still image compression standard," in *Image Processing Algorithms and Techniques*. SPIE, 1990.

[4] Michael W Marcellin, Michael J Gormish, Ali Bilgin, and Martin P Boliek, "An overview of jpeg-2000," in *Proceedings of the Conference on Data Compression*, 2000, pp. 523–541.

[5] Fabrice Bellard, "Bpg image format," 2015.

[6] David Minnen, Johannes Ballé, and George D Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

[8] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.

[9] Ronald J Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[10] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1316–1324.

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12873–12883.

[12] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23318–23340.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[15] David A Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

[16] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[18] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[19] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[21] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[22] 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," *Advances in Neural Information Processing Systems*, vol. 30, 2017.