IMAGINE: AN IMAGINATION-BASED AUTOMATIC EVALUATION METRIC FOR NATURAL LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic evaluations for natural language generation (NLG) conventionally rely on token-level or embedding-level comparisons with the text references. This is different from human language processing, for which visual imaginations often improve comprehension. In this work, we propose IMAGINE, an imagination-based automatic evaluation metric for natural language generation. With the help of CLIP (Radford et al., 2021) and DALL-E (Ramesh et al., 2021), two cross-modal models pre-trained on large-scale image-text pairs, we automatically generate an image as the embodied imagination for the text snippet and compute the imagination similarity using contextual embeddings. Experiments spanning several text generation tasks demonstrate that adding imagination with our IMAGINE displays great potential in introducing multi-modal information into NLG evaluation, and improves existing automatic metrics' correlations with human similarity judgments in many circumstances.

1 INTRODUCTION

A major challenge for natural language generation (NLG) is to design an automatic evaluation metric that can align well with human judgments. To this end, many approaches have been investigated. Metrics that base on matching mechanisms such as BLEU (Papineni et al., 2002), METEOR (Elliott & Keller, 2013), CIDEr (Vedantam et al., 2015), have been widely adopted in the field. Edit-distance based metrics, such as CharacTER (Wang et al., 2016), WMD (Kusner et al., 2015b), SMD (Clark et al., 2019b), have also been explored. Recently, Zhang et al. (2020) proposed to leverage BERT (Devlin et al., 2019) embeddings for computing text similarity, which correlates better with human judgments than previous methods. These automatic evaluation metrics make use of textual information from various angles extensively.

Unlike commonly used automatic methods that compare the generated candidates with the references on the text domain only, humans, in contrast, leverage visual imagination and trigger neural activation in vision-related brain areas when reading text (Just et al., 2004). Cognitive studies show that visual imagery improves comprehension during human language processing (Sadoski & Paivio, 1994). Inspired by this imagination-based multi-modal mechanism in human text comprehension, we ask a critical research question: *can machines create a visual picture of any underlying sentence, and leverage their imaginations to improve natural language understanding?* The advances of powerful pre-trained vision-language models such as CLIP (Radford et al., 2021) provide an excellent opportunity for us to utilize the learned image-text representations and achieve high performance on image-text similarity estimation in a zero-shot fashion. This enables us to introduce multi-modal information into NLG evaluation by generating visual pictures as embodied imaginations.

In this work, we propose IMAGINE, an imagination-based automatic evaluation metric for NLG. IMAGINE first uses the pre-trained discrete variational autoencoder (dVAE) from the vision-language model DALL-E (Ramesh et al., 2021) to visualize imagination, which is to generate descriptive images for the candidate text and the references. Then IMAGINE computes the similarity of the two text snippets and the similarity of the two imaginative images with the pre-trained CLIP model (Radford et al., 2021). Figure 1 shows an evaluation example.

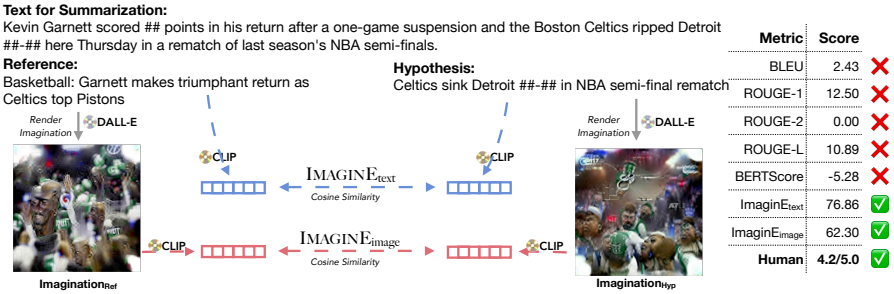


Figure 1: An evaluation example on GigaWord for text summarization. IMAGINE visualizes machine imagination with DALL-E's pre-trained dVAE and extracts textual and visual representations with CLIP. While traditional evaluation metrics for natural language generation rely on n-grams matching or textual embeddings comparison, IMAGINE introduces imagination into the evaluation process and understands the text snippet as a whole with the help of multi-modal information.

To understand the role imagination plays in NLG evaluation, we conduct a series of experiments with IMAGINE on multiple NLG tasks, including machine translation, abstractive text summarization, and data-to-text generation, aiming to answer the following questions:

1. *How influential is* IMAGINE *in NLG evaluation in terms of correlations with human judgments? Can it provide additional reference information on top of existing metrics?*
2. *What are the applicable scenarios of introducing* IMAGINE *to NLG evaluation? When and why does imagination help or not?*
3. *What are the potentials and limitations of introducing imaginations with* IMAGINE *to NLG evaluation?*

Experimental results point out that in a standalone mode for pairwise comparisons, IMAGINE cannot replace textual similarity metrics. However, adding IMAGINE similarity scores to existing metrics surprisingly improves most of the popular metrics' correlations with human performance. Analysis of case studies indicates that IMAGINE can reflect the keyword difference in the visualized imagination, even if the hypothesis and reference text have high n-grams overlaps. In addition, IMAGINE can grasp the gist of two text snippets with similar meanings and renders imaginations that are alike, even if the two pieces of text have distinct word choices. Overall, IMAGINE displays great potential in introducing multi-modal information into NLG evaluation.

2 RELATED WORK

Automatic Metrics for Natural Language Generation Common practices for NLG evaluation compare the generated hypothesis text with the annotated references. Metric performance is conventionally evaluated by its correlation with human judgments. Existing automatic evaluation metric calculations are mainly based on three mechanisms: n-grams overlap, edit distance, and embedding matching. Some typical n-gram based metrics include BLEU (Papineni et al., 2002), ROUGE-n (Lin, 2004), METEOR (Elliott & Keller, 2013) and CIDEr Vedantam et al. (2015), which are widely used for text generation tasks. Another direction is based on edit distance (Tomás et al., 2003; Snover et al., 2006; Panja & Naskar, 2018; Tillmann et al., 1997; Wang et al., 2016) , where they caculate the edit distance between the two text snippets with different optimizations. Embedding-based metrics (Kusner et al., 2015a; Rubner et al., 1998; Clark et al., 2019a; kiu Lo, 2017; 2019) evaluate text quality using word and sentence embeddings, and more recently, with the help of BERT (Zhang et al., 2020; Sellam et al., 2020).

Multi-Modal Automatic Metrics Aside from previous text-only metrics, there also appear metrics that utilize pre-trained multi-modal models and introduce visual features on top of text references for NLG evaluation. TIGEr (Jiang et al., 2019) computes the text-image grounding scores with pre-trained SCAN (Lee et al., 2018). ViLBERTScore-F (Lee et al., 2020) relies on pre-trained ViLBERT (Lu et al., 2019) to extract image-conditioned embeddings for the text. The concurrent CLIPScore (Hessel et al., 2021) proposes a text-reference-free metric for image captioning by directly comparing the image features with caption embeddings with CLIP (Radford et al., 2021). Our method differs in that we use visual picture generation as embodied imaginations and apply our metric to various text-to-text generation tasks.

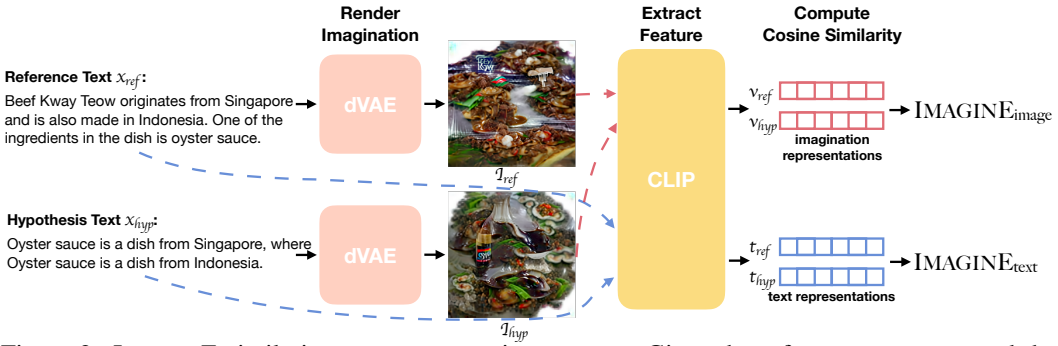


Figure 2: IMAGINE similarity score computation process. Given the reference text $\boldsymbol{x}_{ref}$ and the generated hypothesis $\boldsymbol{x}_{hyp}$, we visualize the machine imagination $\boldsymbol{I}_{ref}$ and $\boldsymbol{I}_{hyp}$ with the pre-trained dVAE. We extract features for the pair of text and corresponding pair of imagination with CLIP. IMAGINE$_{image}$ is the cosine similarity of the imagination representations, while IMAGINE$_{text}$ is the cosine similarity of the text representations.

Mental Imagery The great imagery debate is still an open question in the neuroscience and psychology community (Troscianko, 2013). The debate between pictorialists and propositionalists is about how imagery information is stored in the human brain. We follow the views from pictorialists that information can be stored in a depictive and pictorial format in addition to language-like forms (Kosslyn et al., 2001; Pearson & Kosslyn, 2015). In pictorialists' model, mental imagery is constructed in the "visual buffer" either from the retinal image in seeing or from a long-term memory store of "deep representations" in the brain. Our method of image generation is to mimic the generation of deep representations in machines, with the help of recent powerful text-to-image models. Inspired by empirical studies from cognitive science that visual imagination improves human text comprehension (Gambrell & Bales, 1986; Nippold & Duthie, 2003; Just et al., 2004; Joffe et al., 2007), we are interested in exploring if one can draw similar conclusions from automatic text evaluations by machines.

3 IMAGINE

3.1 MODEL DETAILS

CLIP CLIP (Radford et al., 2021) is a cross-modal retrieval model trained on WebImageText, which consists of 400M (image, caption) pairs gathered from the web. WebImageText was constructed by searching for 500K queries on a search engine. The base query list is all words occurring at least 100 times in the English version of Wikipedia, augmented with bi-grams with high pointwise mutual information as well as the names of all Wikipedia articles above a certain search volume. Each query includes 20K (image, text) pairs for class balance.

In this work, we use the ViT-B/32 version of CLIP, in which the Vision Transformer (Dosovitskiy et al., 2020; Vaswani et al., 2017) adopts BERT-Base configuration and uses 32×32 input patch size. The Vision Transformer takes 224×224 input image and the self-attention maps are calculated between 7×7 grid of image patches. The Text Transformer has 12-layer, 8-head and uses a hidden size of 512, and is trained over a vocab of 49K BPE token types (Radford et al., 2019; Sennrich et al., 2016). The text representation is the last hidden state of the "[EOT]" token being projected by a linear layer. The model's weights are trained to maximize the similarity of truly corresponding image/caption pairs while simultaneously minimizing the similarity of mismatched image/caption pairs using InfoNCE (Sohn, 2016; van den Oord et al., 2018).

DALL-E DALL-E (Ramesh et al., 2021) is a 12-billion parameter version of GPT-3 (Brown et al., 2020) trained to generate images from text descriptions. The model is trained on a dataset of a similar scale to JFT-300M (Sun et al., 2017) by collecting 250 million text-image pairs from the internet, which incorporates Conceptual Captions (Sharma et al., 2018), the text-image pairs from Wikipedia, and a filtered subset of YFCC100M (Thomee et al., 2016).

DALL-E trains a discrete variational autoencoder (dVAE) (Rolfe, 2017) to encode each 256×256 RGB image into a 32×32 grid of image tokens with a vocabulary size of 8192. The image tokens are concatenated with a maximum of 256 BPE-encoded (Sennrich et al., 2016; Radford et al., 2019)

tokens with a vocabulary size of 16384 that represents the paired image caption. DALL-E trains an autoregressive transformer to model the joint distribution over the text and image tokens. The pre-trained dVAE has been made public, while the pre-trained transformer is not released. Thus, we use DALL-E's pre-trained dVAE to render images in this project.

3.2 IMAGINE SIMILARITY SCORE

Construct Imagination For each image, we randomly initialize a latent matrix $\boldsymbol{H}$ and use the pre-trained dVAE to produce the RGB image $\boldsymbol{I} = dVAE\_decoder(\boldsymbol{H})$. We use the ViT-B/32 version of the CLIP model to encode the generated image $\boldsymbol{I}$ and the input text $\boldsymbol{x}$. Then we use CLIP to compute the similarity between the received image embedding $\boldsymbol{v} = CLIP(\boldsymbol{I})$ and text embedding $\boldsymbol{t} = CLIP(\boldsymbol{x})$ as the loss to optimize the hidden matrix while keeping the weights of the network unchanged. We optimize each generation process for 1000 steps, and refer to the generated image as the imagination for further computation.

$$loss_{generation} = -\frac{\boldsymbol{v}^T\boldsymbol{t}}{\|\boldsymbol{v}\|\|\boldsymbol{t}\|} \tag{1}$$

Similarity Measure For the generated text snippet $\boldsymbol{x}_{hyp}$ and all the references $\{\boldsymbol{x}_{ref_i}\}_{i=1}^n$, we generate corresponding images $\boldsymbol{I}_{hyp}$ and $\boldsymbol{I}_{ref_i}$ for $i \in [1, n]$, where n is the number of parallel references. During evaluation, we pass both the pair of text snippets and the corresponding imaginations through corresponding CLIP feature extractors to receive the textual representation $\boldsymbol{t}_{hyp}$, $\boldsymbol{t}_{ref_i}$, and the imagination representations $\boldsymbol{v}_{hyp}$, $\boldsymbol{v}_{ref_i}$. Then, we compute three types of similarity scores for IMAGINE with the received embeddings: IMAGINE$_{text}$ compares the hypothesis text $\boldsymbol{x}_{hyp}$ with the text references $\boldsymbol{x}_{ref_i}$; IMAGINE$_{image}$ compares the visualized imaginations $\boldsymbol{I}_{hyp}$ with $\boldsymbol{I}_{ref_i}$, generated by the pre-trained dVAE in previous steps; IMAGINE$_{text\&image}$ is the average of IMAGINE$_{text}$ and IMAGINE$_{image}$, which takes both the text and the imagination into consideration.

$$\text{IMAGINE}_{text} = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{t}_{hyp}^T\boldsymbol{t}_{ref_i}}{\|\boldsymbol{t}_{hyp}\|\|\boldsymbol{t}_{ref_i}\|} \tag{2}$$

$$\text{IMAGINE}_{image} = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{v}_{hyp}^T\boldsymbol{v}_{ref_i}}{\|\boldsymbol{v}_{hyp}\|\|\boldsymbol{v}_{ref_i}\|} \tag{3}$$

3.3 EXTENSION TO EXISTING METRICS

The IMAGINE similarity scores can be used as individual automatic metrics. Apart from this, IMAGINE can also act as an extension to existing metrics, as it provides multimodal references that compensate for current text-only evaluations that compare tokens or text-embeddings. Our adaptation of IMAGINE to other automatic metrics is direct, which is summing up IMAGINE similarity score with the other automatic metric score for each example:

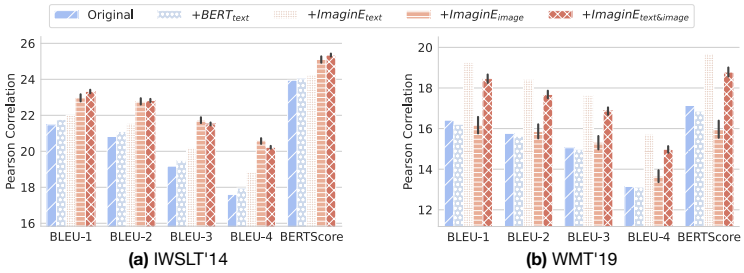
$$metric\_score' = metric\_score + \text{IMAGINE}\_similarity\_score \tag{4}$$

4 EXPERIMENTAL SETUP

Tasks, Datasets, and Models We evaluate our approach on three natural language generation tasks: machine translation, abstractive text summarization, and data-to-text generation. For machine translation, we use Fairseq (Ott et al., 2019) implementation to generate English translation from German on IWSLT'14 (Bell et al., 2014) and WMT'19 (Barrault et al., 2019) datasets. We choose these two to-English translation tasks because currently, DALL-E and CLIP only support English. For abstractive text summarization, we use the implementation of Li et al. (2017) to generate sentence summarization on DUC2004[1] and use ProphetNet (Yan et al., 2020) for generation on Gigaword[2]. We choose abstractive text summarization instead of document summarization since CLIP sets a length limit of input text of 77 BPE tokens. For data-to-text generation, we conduct experiments on three datasets, namely WebNLG (Gardent et al., 2017), E2ENLG (Dusek et al., 2019;

[1] https://duc.nist.gov/duc2004/

[2] https://catalog.ldc.upenn.edu/LDC2011T07



(a) IWSLT'14 **(b)** WMT'19

Figure 3: The effectiveness of augmenting BLEU-n (n=1,2,3,4) and BERTScore with IMAGINE similarities and BERT$_{text}$ similarity on two machine translation datasets. The y-axis shows the Pearson correlation with human judgments.

(a) IWSLT'14 **(b)** WMT'19

Figure 4: Case studies for machine translation. **Src**: the German text to be translated. **Ref**: the reference translation. **Hyp**: the generated translation candidate. We report the metric scores and the human score for the reported pair of (Ref, Hyp).

2020) and WikiBioNLG (Lebret et al., 2016). We use the text generated by the KGPT (Chen et al., 2020) model in our experiments. Table 3 lists out the statistics of the test set used for each dataset.

Automatic Metrics For machine translation, we report BLEU-n (Papineni et al., 2002) for $n = 1, 2, 3, 4$ and BERTScore (Zhang et al., 2020). For abstractive text summarization, we report results on ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004) and BERTScore. For data-to-text generation, we utilize five automatic metrics for NLG, including BLEU, ROUGE-L, METEOR (Elliott & Keller, 2013), CIDEr (Vedantam et al., 2015) and BERTScore. In comparison with IMAGINE$_{text}$, we also compute BERT$_{text}$, the text similarity score with BERT encoder. We use the last hidden state for the "[CLS]" token as the representation of the text snippet, and compute cosine similarity with the two "[CLS]" embeddings for the reference and the generated text candidate.

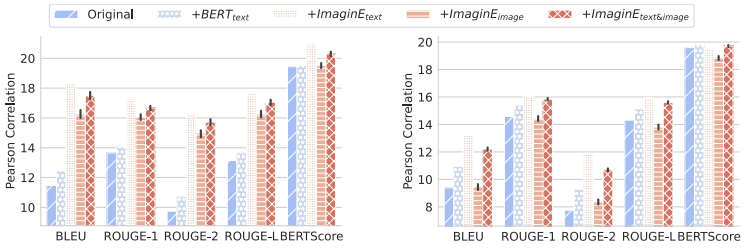
Human Evaluation We invite MTurk[3] annotators to judge the quality of the generated text. The estimated hourly wage is $12. We use the complete test set for DUC2004 and E2ENLG, which contains 500 and 630 examples, respectively. For the remaining five datasets, we randomly sample 1k pair of test examples for human evaluation due to the consideration of expenses. Each example is scored by three human judges using a 5-point Likert scale. The generated text is evaluated from three aspects, namely fluency, grammar correctness, and factual consistency with the reference text. We take the mean of human scores to compute correlations. In the following sections, we report Pearson correlation (Freedman et al., 2007) to human scores. We also record Kendall correlation (Kendall, 1938) in the Appendix.

5 RESULTS

5.1 MACHINE TRANSLATION

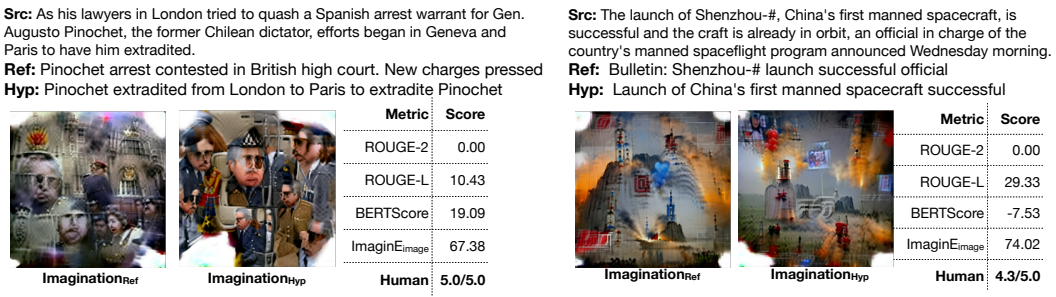
Figure 3 shows the system-level Pearson correlation to human judges when extending our IMAGINE similarity to existing automatic NLG metrics on the IWSLT'14 and WMT'19 German to English datasets. IMAGINE$_{text}$ and IMAGINE$_{text\&image}$ steadily improves all the listed metrics'

[3] https://www.mturk.com/



(a) DUC2004 **(b)** Gigaword

Figure 5: The effectiveness of augmenting BLEU, BERTScore and ROUGE-related metrics with IMAGINE similarities and BERT$_{text}$ similarity on two abstractive text summarization datasets. The y-axis shows the Pearson correlation with human judgments.



(a) DUC2004 **(b)** GigaWord

Figure 6: Case studies for abstractive text summarization. **Src**: the text to be summarized. **Ref**: the reference summary. **Hyp**: the generated summary candidate. We report the metric scores and the human score for the reported pair of (Ref, Hyp).

correlations with human scores. IMAGINE$_{image}$ and IMAGINE$_{text\&image}$ contributes the most in IWSLT'14 while IMAGINE$_{text}$ plays the most important role in WMT'19. IMAGINE$_{image}$ also enhances most of the metrics' correlations except for BERTScore in WMT'19. BERT$_{text}$ has relatively small impact on improving other metrics' correlation in the machine translation task.

Figure 4 lists out two examples for the case study. We notice that IMAGINE can capture the keyword difference between the reference and the hypothesis text, even if they have similar sentence structures and high n-grams overlaps. IMAGINE shows its sensitivity to word choice in Figure 4(a). The main difference between the reference text and the generated text is the mention of "manager" and "ladder". While other metrics score high, the quality of the generated text is questionable. In contrast, our IMAGINE renders distinct imaginations and assigns lower image similarity. In Figure 4(b), the reference text leaves the movie names in German, while the hypothesis text translates all contents to English. Aside from this, the translations are nearly identical. However, IMAGINE yields completely different imaginations. This suggests that IMAGINE's performance is greatly impaired when applied to non-English scenarios.

5.2 ABSTRACTIVE TEXT SUMMARIZATION

Figure 5 shows the system-level Pearson correlation to human judges when extending our IMAGINE similarity to existing automatic NLG metrics on the DUC2004 and Gigaword. Both datasets are built upon news articles. IMAGINE can steadily improve BLEU, ROUGE-related metrics, and BERTScore on DUC2004. IMAGINE$_{text}$ contributes to the most significant improvement on Gigaword. IMAGINE$_{text}$ surpasses BERT$_{text}$ on all metrics except for BERTScore on Gigaword.

IMAGINE can capture the gist of texts with similar meanings and renders reasonable descriptive imaginations that are alike, regardless of word choices. Figure 6 shows two sets of examples where the hypothesis summary scores high in human evaluation but scores low on existing automatic evaluation metrics. Both examples have low n-grams overlaps between the hypothesis and reference summary, but IMAGINE renders similar imagination and assigns high image similarity scores, which align with human scores.

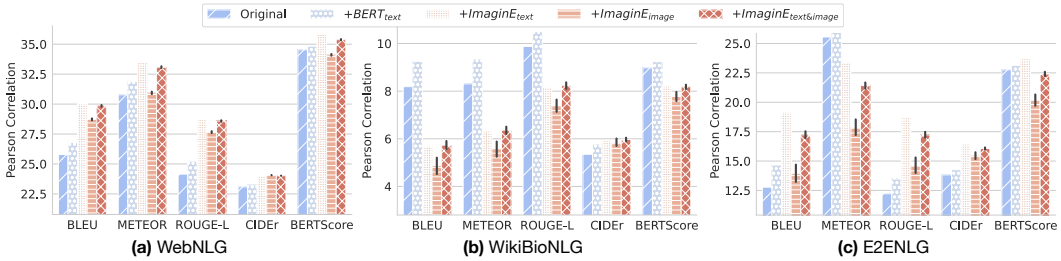


Figure 7: The effectiveness of augmenting BLEU, METEOR, ROUGE-L, CIDEr, and BERTScore with IMAGINE similarities and BERT$_{text}$ similarity on three data-to-text generation datasets. The y-axis shows the Pearson correlation with human judgments.

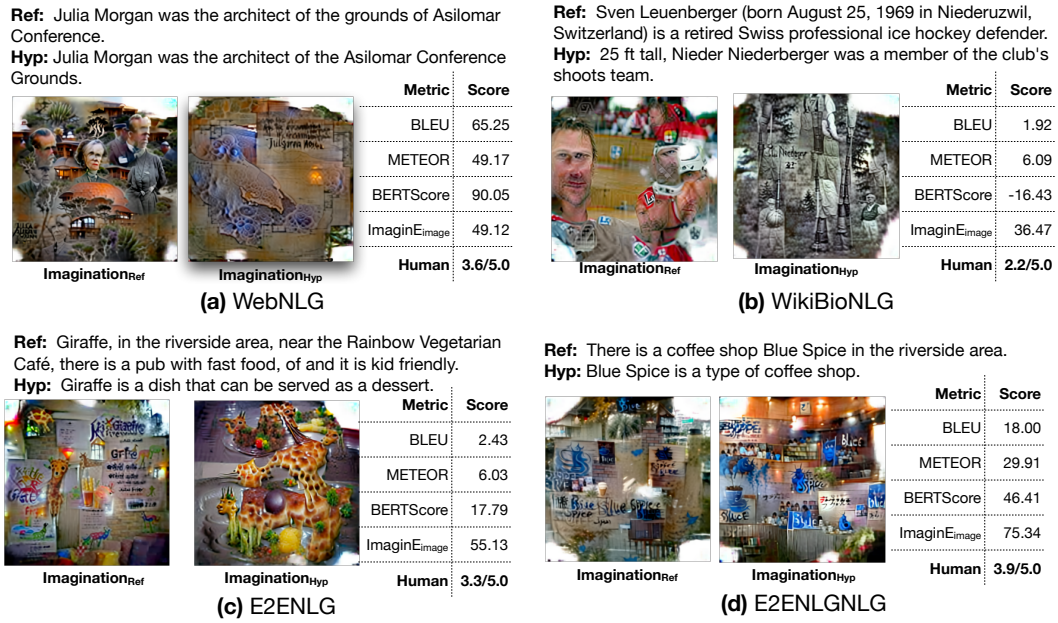


Figure 8: Case studies for data-to-text generation. **Ref**: the reference text. **Hyp**: the generated text candidate. We report the metric scores and the human score for the reported pair of (Ref, Hyp).

5.3 DATA-TO-TEXT GENERATION

Figure 7 shows the system-level Pearson correlation to human judges when extending our IMAGINE similarity to existing automatic NLG metrics on the WebNLG, WikiBioNLG, and E2ENLG datasets. Figure 8 lists out four examples for the case study.

On WebNLG, adding IMAGINE$_{text}$ and IMAGINE$_{text\&image}$ can steadily improve all the listed metrics' correlation with human scores. IMAGINE$_{image}$ improves BLEU, METEOR, ROUGE-L, and CIDEr but it only has limited impact on BERTScore. Among the two metrics that compare textual similarity, IMAGINE$_{text}$ boosts correlations more than BERT$_{text}$. As discussed in Section 5.1, IMAGINE shows its sensitivity to the input text snippet. We see this again in Figure 8(a), in which changing the relative position of "grounds" shifts the central part of the imagination from a person to the dirt ground.

We witness a drawback in most listed metrics' correlations after applying our IMAGINE approach on WikiBioNLG. This is because the WikiBioNLG dataset is built upon Wikipedia biography, and IMAGINE is not good at visualizing abstract concepts. In Figure 8(b), our IMAGINE failed to visualize the player's birth date or height. Such information may be contained in BERT pre-training data, but is not as likely to be covered by the dataset to train CLIP, which explains IMAGINE$_{text}$'s inferior performance compared to BERT$_{text}$. Figure 7(b) shows the lowest Pearson correlation among all three datasets on all metrics, which means this dataset is not only a challenge to our IMAGINE approach but also to other existing metrics as well.

On E2ENLG, textual similarity scores play a more influential role in improving correlation as it has a positive impact on all listed metrics except for METEOR. BERT$_{text}$ outperforms IMAGINE$_{text}$ in all listed metrics except for ROUGE-L. On the other hand, IMAGINE$_{image}$ has a salient negative impact on correlation. The E2ENLG dataset is built upon restaurant domain information. We found that IMAGINE is sensitive and may be misguided by irrelevant information, such as the restaurant names, which explains the poor performance of IMAGINE$_{image}$. For example, "Giraffe" and "Rainbow" in Figure 8(c) result in weird imagination that is unrelated to the main content of the generated text. "Blue Spice" leads to the appearance of blue patches in Figure 8(d).

6 DISCUSSION

Metric	Original	+BERT$_{text}$	+IE$_{text}$	+IE$_{image(dVAE)}$	+IE$_{image(BigGAN)}$	+IE$_{image(VQGAN)}$
ROUGE-1	13.66	14.05	**17.21**	16.05 $\pm$ 0.46	15.82 $\pm$ 0.72	15.93 $\pm$ 0.91
ROUGE-2	9.74	10.71	**16.29**	14.92 $\pm$ 0.61	14.62 $\pm$ 0.96	14.77 $\pm$ 1.21
ROUGE-L	13.14	13.65	**17.66**	16.25 $\pm$ 0.55	16.01 $\pm$ 0.85	16.12 $\pm$ 1.07
BERTScore	19.44	19.50	**20.97**	19.50 $\pm$ 0.43	19.29 $\pm$ 0.70	19.39 $\pm$ 0.90
BLEURT	23.59	23.53	**24.28**	23.47 $\pm$ 0.23	23.33 $\pm$ 0.39	23.39 $\pm$ 0.46

Table 1: The Pearson correlations with human judges when using BERT$_{text}$ similarity and IMAGINE similarities to augment ROUGE, BERTScore, and BLEURT on DUC2004. Here we computes three sets of IMAGINE$_{image}$ similarity scores (mean$\pm$std) with three different image generation backbones for IMAGINE, namely dVAE, BigGAN, and VQGAN. IE: IMAGINE.

Image Generation Backbones In previous sections, we implement IMAGINE with dVAE as the image generation backbone. There also appear a number of exciting and creative CLIP-based image generation repositories such as BigSleep[4] and VQGAN-CLIP[5], which use BigGAN (Brock et al., 2019) and VQGAN (Esser et al., 2021) to generate images respectively.

Here we discuss the choice of IMAGINE's image generation backbone and its effect on evaluation performance. We conduct experiments on DUC2004 for summarization, and compare dVAE with BigGAN and VQGAN. For fair comparisons, each generative backbone has a 1000-step learning phase to render a 512x512 image for each piece of input text. Examining Table 1, we find comparable IMAGINE$_{image}$ performances when using different generative backbones. The dVAE leads to slightly higher correlations and smaller variance. The variability of random initialization may cause the larger variances of the two GAN-based image generators.

To assess the influence of random initialization, we repeat the image generation process five times and compute pairwise visual similarities within each group of 5 images. Notice in Figure 9(a) that dVAE has the highest intra-group visual similarity, which suggests that compared to the two GAN-based generative backbones, dVAE is relatively more robust to the random initialization.

Applicable Scenarios As shown in Figures 3, 5 and 7, we notice that adding certain type of IMAGINE similarities improves non-embedding-based metrics' correlations with human scores in most cases. This suggests that it is helpful to extend text-only non-embedding-based metrics with multimodal knowledge. Table 2 lists out each metric's Pearson correlation with human judgments on each dataset. In standalone-mode for pairwise comparisons, IMAGINE similarity scores can not replace textual similarity metrics. In Section 5.3, we find that IMAGINE struggles to render informative images on WikiBioNLG, a dataset that contains many abstract concepts that are hard to visualize, such as specific date, length, weight, etc.

From Figures 5 and 7, it also occurs to us that IMAGINE sometimes fails to improve BERTScore's performance, while BERT$_{text}$ often has further improvements over BERTScore. One possible explanation is the domain difference between CLIP and BERT, which causes their embeddings to lie in distinct spaces. Since BERTScore is computed on top of BERT-based textual embeddings that are pre-trained on another source of data, our CLIP-based IMAGINE may not be supportive.

Score Distribution To further validate the effectiveness of our methods, we visualize the score distributions of different metrics. As shown in Figure 9(b), BERT$_{text}$ has the sharpest distribution,

[4] https://github.com/lucidrains/big-sleep

[5] https://github.com/nerdyrodent/VQGAN-CLIP

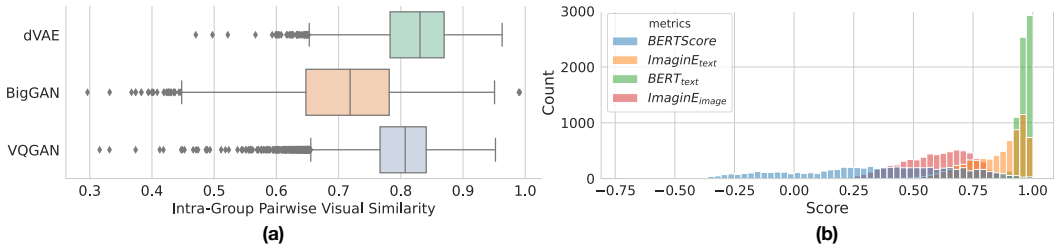


Figure 9: (a) The intra-group pairwise visual similarity distributions for images generated by dVAE, BigGAN, and VQGAN. The plot shows the three quartile values and the extreme values. (b) The score distributions histplot of IMAGINE, BERT$_{text}$ and BERTScore used in our experiments. All four metrics range between [-1, 1].

Task	Dataset	Pearson Correlation									
MT		BLEU-1	BLEU-2	BLEU-3	BLEU-4	BERTScore	BLEURT	BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$
	WMT19	16.41	15.76	15.06	13.15	17.14	17.79	4.37	**20.34**	3.80 ± 1.78	10.11 ± 1.51
	IWSLT14	21.47	20.82	19.17	17.60	**23.95**	22.93	18.42	14.11	15.92 ± 0.95	17.75 ± 0.71
TS		BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BLEURT	BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$
	DUC2004	11.47	13.66	9.74	13.14	19.44	**23.59**	12.10	19.81	15.01 ± 1.03	18.03 ± 1.08
	GigaWord	9.39	14.58	7.75	14.31	19.59	**20.23**	17.49	15.56	3.74 ± 0.98	12.27 ± 0.69
DT		BLEU	METEOR	ROUGE-L	CIDEr	BERTScore	BLEURT	BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$
	WebNLG	25.79	30.78	24.15	23.09	34.53	**35.97**	22.38	26.81	19.69 ± 0.49	24.82 ± 0.37
	E2ENLG	12.78	**25.55**	12.22	13.83	22.76	22.75	13.11	18.19	10.89 ± 2.40	15.33 ± 1.02
	WikiBioNLG	8.19	8.31	**9.88**	5.35	8.98	9.21	6.07	4.14	3.32 ± 0.89	4.10 ± 0.50

Table 2: The Pearson correlations with human judgement for each individual metric. IE: IMAGINE. MT: machine translation. TS: abstractive text summarization. DT: data-to-text generation.

while our imagination-based methods lead to smoother distributions. This indicates IMAGINE$_{image}$ is more diverse than text-based metrics with the same measurement (i.e., cosine similarity). We also observe that BERTScore, which computes maximum matching after calculating cosine similarity on token embeddings, provides a more uniform distribution compared to the other three. Currently, the value of IMAGINE$_{text}$ usually lies between [0.6, 1], and IMAGINE$_{image}$ usually lies between [0.3, 1]. It would be preferable if future work can help IMAGINE to be more distinctive.

Future Work As noted in Section 5, IMAGINE can capture the keyword difference and render distinct imaginations for two pieces of similar text. One supportive case is Figure 4(a). While this ensures IMAGINE's ability to distinguish keyword differences, it also cast doubt on IMAGINE's robustness. In Figure 8(a), merely changing the relative position of "grounds" result in two entirely different images. In Figure 8(c) and (d), the name of the restaurants also reduces the quality of the imagination. Future work may systematically examine the robustness of CLIP and DALL-E regarding textual variance.

Furthermore, even though we have access to DALL-E's pre-trained dVAE decoder, we still need to generate the imagination from scratch for each example, which can be compute-intensive. We are interested in exploring more efficient ways to speed up the image generation process.

Aside from the above points listed, we also find the following topics worth exploring. Currently, the CLIP text encoder has a length constraint of 77 BPE tokens, [BOS] and [EOS] included. This limits our attempt on longer text generation tasks, such as story generation, document summarization, etc. Also, CLIP and DALL-E only support English for now. With a multilingual CLIP and DALL-E, we may cross verify the similarity with text and imagination in other source languages.

7 CONCLUSION

In this paper, we propose IMAGINE, an imagination-based automatic evaluation metric for NLG. Experiments on three tasks and seven datasets find out that adding IMAGINE similarity scores as an extension to current non-embedding-based metrics can improve their correlations with human judgments in many circumstances. We hope our work can contribute to the construction of multi-modal representations and the discussion of multi-modal studies.

ETHICAL STATEMENT

Our study is approved for IRB exempt. The estimated hourly wage paid to MTurk annotators is $12. Speaking of potential ethical concerns, our "imagination" approach may face an issue of fairness if there exists any bias in the training dataset for CLIP or DALL-E. In such circumstances, IMAGINE might display a tendency to render specific types of images that it has seen in the training data. Even though we did not witness such issues in our study, we should keep in mind that this unfair behavior would impair IMAGINE's effectiveness as an evaluation tool.

REPRODUCIBILITY STATEMENT

All of the datasets used in our study on machine translation, data-to-text generation and abstractive text summarization tasks are publicly available. We use the public repositories to implement IMAGINE. The implementations of CLIP-based image generators used in our study are dVAE+CLIP[6], Big-Sleep(BigGAN+CLIP)[7] and VQGA+CLIP[8].

REFERENCES

Loïc Barrault, Ondrej Bojar, M. Costa-jussà, C. Federmann, M. Fishel, Yvette Graham, B. Haddow, M. Huck, Philipp Koehn, S. Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (wmt19). In *WMT*, 2019.

P. Bell, P. Swietojanski, J. Driesen, M. Sinclair, F. McInnes, and S. Renals. 11th international workshop on spoken language translation (iwslt 2014). 2014.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019.

T. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Wenhu Chen, Yu Su, X. Yan, and W. Wang. Kgpt: Knowledge-grounded pre-training for data-to-text generation. In *EMNLP*, 2020.

Elizabeth Clark, A. Çelikyilmaz, and Noah A. Smith. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *ACL*, 2019a.

Elizabeth Clark, A. Çelikyilmaz, and Noah A. Smith. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *ACL*, 2019b.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

A. Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, M. Dehghani, Matthias Minderer, G. Heigold, S. Gelly, Jakob Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

Ondrej Dusek, David M. Howcroft, and Verena Rieser. Semantic noise matters for neural natural language generation. In *INLG*, 2019.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Comput. Speech Lang.*, 59:123–156, 2020.

[6] https://github.com/openai/DALL-E

[7] https://github.com/lucidrains/big-sleep

[8] https://github.com/nerdyrodent/VQGAN-CLIP

Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1292–1302, 2013.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

David Freedman, Robert Pisani, and Roger Purves. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

Linda B Gambrell and Ruby J Bales. Mental imagery and the comprehension-monitoring performance of fourth-and fifth-grade poor readers. *Reading Research Quarterly*, pp. 454–464, 1986.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for nlg micro-planners. In *ACL*, 2017.

Jack Hessel, Ariel Holtzman, Maxwell Forbes, R. L. Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718, 2021.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: Text-to-image grounding for image caption evaluation. In *EMNLP*, 2019.

Victoria L Joffe, Kate Cain, and Nataša Marić. Comprehension problems in children with specific language impairment: does mental imagery training help? *International Journal of Language & Communication Disorders*, 42(6):648–664, 2007.

M. Just, S. Newman, T. Keller, A. McEleney, and P. Carpenter. Imagery in sentence comprehension: an fmri study. *NeuroImage*, 21:112–124, 2004.

M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

Chi kiu Lo. Meant 2.0: Accurate semantic mt evaluation for any output language. In *WMT*, 2017.

Chi kiu Lo. Yisi - a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *WMT*, 2019.

Stephen M Kosslyn, Giorgio Ganis, and William L Thompson. Neural foundations of imagery. *Nature reviews neuroscience*, 2(9):635–642, 2001.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *ICML*, 2015a.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *ICML*, 2015b.

Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *EMNLP*, 2016.

H. Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and K. Jung. Vilbertscore: Evaluating image caption using vision-and-language bert. In *EVAL4NLP*, 2020.

Kuang-Huei Lee, X. Chen, G. Hua, H. Hu, and Xiaodong He. Stacked cross attention for image-text matching. *ArXiv*, abs/1803.08024, 2018.

Piji Li, Wai Lam, Lidong Bing, and Z. Wang. Deep recurrent generative decoder for abstractive text summarization. *ArXiv*, abs/1708.00625, 2017.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

Marilyn A Nippold and Jill K Duthie. Mental imagery and idiom comprehension. 2003.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

J. Panja and S. Naskar. Iter: Improving translation edit rate through optimizable edit costs. In *WMT*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Joel Pearson and Stephen M Kosslyn. The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, 112(33):10089–10092, 2015.

Alec Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, J. W. Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, J. Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ArXiv*, abs/2103.00020, 2021.

A. Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021.

J. Rolfe. Discrete variational autoencoders. *ArXiv*, abs/1609.02200, 2017.

Y. Rubner, Carlo Tomasi, and L. Guibas. A metric for distributions with applications to image databases. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 59–66, 1998.

Mark Sadoski and A. Paivio. A dual coding view of imagery and verbal processes in reading comprehension. 1994.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. Bleurt: Learning robust metrics for text generation. In *ACL*, 2020.

Rico Sennrich, B. Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909, 2016.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

Matthew G. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *AMTA*, 2006.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

C. Sun, Abhinav Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 843–852, 2017.

B. Thomee, D. Shamma, G. Friedland, Benjamin Elizalde, Karl S. Ni, Douglas N. Poland, Damian Borth, and L. Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59:64–73, 2016.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated dp based search for statistical translation. In *EUROSPEECH*, 1997.

Jesús Tomás, J. Mas, and F. Casacuberta. A quantitative method for machine translation evaluation. 2003.

Emily T Troscianko. Reading imaginatively: the imagination in cognitive science and cognitive literary studies. *Journal of Literary Semantics*, 42(2):181–198, 2013.

Aäron van den Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Weiyue Wang, J. Peter, Hendrik Rosendahl, and H. Ney. Character: Translation edit rate on character level. In *WMT*, 2016.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, J. Chen, R. Zhang, and M. Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *ArXiv*, abs/2001.04063, 2020.

Tianyi Zhang, V. Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2020.

A Appendix

A.1 Dataset Details

Table 3 lists out the statistical details of the datasets' test sets used in our study.

Task	Dataset	$\#sample$	$\#ref$	$\#len_{ref}$	$\#len_{hyp}$
Machine Translation	WMT'19	2,000	1.0	22.4	22.4
	IWSLT'14	6,750	1.0	20.3	19.1
Abstractive Text Summarization	DUC2004	500	4.0	14.0	10.0
	GigaWord	1,950	1.0	9.9	11.9
Data-to-Text Generation	WebNLG	1,600	2.6	28.3	26.9
	E2ENLG	630	7.4	28.0	11.6
	WikiBioNLG	2,000	1.0	34.8	19.0

Table 3: Dataset statistics. *#sample* is the number of samples in the test set; *#ref* is the number of parallel references per visual instance; *#len* is the average reference length.

A.2 Random Initialization

We discussed the influence of random initialization for different image generative backbones in Section 6. In Figure 10, we show several groups of images generated by dVAE, BigGAN and VQGAN with random initialization.

A.3 Correlation Results

We list the numbers on Pearson correlation in Tables 6, 8 and 10 that match Figures 3, 5 and 7 in the main paper. Tables 4, 5, 7 and 9 display results on Kendall correlation for the three NLG tasks used in our study. The Kendall correlations with human judgement show similar trends as those on Pearson correlation.

A.4 Case Study

We provide more case studies for the three NLG tasks used in our study in Figures 11 to 17. For each dataset in each task, we list 4 groups of examples together with the imagination rendered by ImaginE and the automatic evaluation scores.

Task	Dataset	Kendall Correlation									
MT		BLEU-1	BLEU-2	BLEU-3	BLEU-4	BERTScore	BLEURT	BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$
	WMT19	13.22	12.98	12.07	10.74	12.23	13.06	7.28	**15.90**	2.83 ± 1.42	7.15 ± 1.12
	IWSLT14	14.19	14.26	13.68	12.79	**16.68**	14.64	13.84	12.90	10.87 ± 0.73	12.58 ± 0.61
TS		BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BLEURT	BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$
	DUC2004	8.96	8.71	7.75	7.22	12.82	**16.04**	8.31	9.49	8.23 ± 1.02	8.94 ± 0.95
	GigaWord	12.26	12.15	9.21	12.40	14.10	**15.16**	13.11	12.83	2.18 ± 0.85	9.04 ± 0.62
DT		BLEU	METEOR	ROUGE-L	CIDEr	BERTScore	BLEURT	BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$
	WebNLG	15.94	21.30	15.24	13.40	23.44	**24.31**	15.41	19.55	12.84 ± 0.44	16.73 ± 0.30
	E2ENLG	11.53	**18.46**	8.60	10.29	14.45	14.61	10.86	10.59	6.37 ± 1.49	8.86 ± 0.81
	WikiBioNLG	3.27	3.73	4.00	2.10	5.09	**5.32**	3.07	2.08	1.36 ± 0.66	1.68 ± 0.35

Table 4: The Kendall correlations with human judgement for each individual metric. IE: IMAGINE. MT: machine translation. TS: abstractive text summarization. DT: data-to-text generation.

Dataset	Kendall Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
WMT19	BLEU-1	13.22	13.02	**14.98**	12.46 ± 0.67	14.12 ± 0.43
	BLEU-2	12.98	12.74	**14.35**	12.40 ± 0.64	13.77 ± 0.40
	BLEU-3	12.07	11.91	**13.52**	11.98 ± 0.61	12.97 ± 0.35
	BLEU-4	10.74	10.63	**12.42**	10.67 ± 0.60	11.63 ± 0.30
	BERTScore	12.23	11.98	**13.96**	11.21 ± 0.75	13.02 ± 0.51
	BLEURT	13.06	13.05	**14.31**	13.02 ± 0.46	13.82 ± 0.25
IWSLT14	BLEU-1	14.19	14.42	15.07	15.15 ± 0.38	**15.67** ± 0.23
	BLEU-2	14.26	14.48	14.95	15.31 ± 0.35	**15.52** ± 0.19
	BLEU-3	13.68	13.82	14.25	14.69 ± 0.31	**14.81** ± 0.17
	BLEU-4	12.79	13.02	13.39	**14.00** ± 0.28	13.99 ± 0.16
	BERTScore	16.68	16.70	17.38	17.10 ± 0.34	**17.70** ± 0.17
	BLEURT	14.64	14.68	14.93	**15.36** ± 0.17	15.28 ± 0.08

Table 5: The Kendall correlations with human judgement on the machine translation task.

Dataset	Pearson Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
WMT19	BLEU-1	16.41	16.21	**19.25**	16.17 ± 0.99	18.46 ± 0.54
	BLEU-2	15.76	15.62	**18.41**	15.86 ± 0.89	17.68 ± 0.49
	BLEU-3	15.06	14.96	**17.61**	15.30 ± 0.81	16.87 ± 0.45
	BLEU-4	13.15	13.12	**15.72**	13.66 ± 0.78	14.98 ± 0.42
	BERTScore	17.14	16.86	**19.70**	15.95 ± 1.07	18.78 ± 0.59
	BLEURT	17.79	17.73	**18.86**	18.40 ± 0.53	18.77 ± 0.25
IWSLT14	BLEU-1	21.47	21.77	22.01	22.97 ± 0.50	**23.33** ± 0.26
	BLEU-2	20.82	21.10	21.56	22.77 ± 0.45	**22.82** ± 0.22
	BLEU-3	19.17	19.50	20.21	**21.73** ± 0.42	21.52 ± 0.21
	BLEU-4	17.60	17.96	18.88	**20.58** ± 0.41	20.22 ± 0.20
	BERTScore	23.95	24.02	24.24	25.10 ± 0.43	**25.34** ± 0.21
	BLEURT	22.93	23.00	23.12	**24.06** ± 0.20	23.74 ± 0.09

Table 6: The Pearson correlations with human judgement on the machine translation task.

Dataset	Kendall Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
DUC2004	BLEU	8.96	9.42	**10.03**	9.10 $\pm$ 0.71	9.59 $\pm$ 0.57
	ROUGE-1	8.71	8.86	**9.96**	9.49 $\pm$ 0.39	9.77 $\pm$ 0.30
	ROUGE-2	7.75	9.49	**9.89**	9.10 $\pm$ 0.61	9.62 $\pm$ 0.55
	ROUGE-L	7.22	8.09	**9.91**	9.40 $\pm$ 0.51	9.61 $\pm$ 0.37
	BERTScore	12.82	**13.15**	12.63	11.93 $\pm$ 0.43	12.35 $\pm$ 0.32
	BLEURT	16.04	**16.11**	16.00	15.52 $\pm$ 0.22	15.74 $\pm$ 0.20
GigaWord	BLEU	12.26	**12.50**	12.37	7.49 $\pm$ 0.68	11.47 $\pm$ 0.33
	ROUGE-1	12.15	12.14	**12.18**	11.04 $\pm$ 0.39	12.13 $\pm$ 0.19
	ROUGE-2	9.21	**12.04**	11.79	6.74 $\pm$ 0.63	10.10 $\pm$ 0.34
	ROUGE-L	12.40	12.59	12.55	11.26 $\pm$ 0.45	**12.69** $\pm$ 0.21
	BERTScore	14.10	14.24	14.32	13.56 $\pm$ 0.32	**14.39** $\pm$ 0.16
	BLEURT	15.16	**15.24**	14.96	14.91 $\pm$ 0.19	15.09 $\pm$ 0.09

Table 7: The Kendall correlations with human judgement on the abstractive text summarization task.

Dataset	Pearson Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
DUC2004	BLEU	11.47	12.47	**18.31**	16.25 $\pm$ 0.71	17.50 $\pm$ 0.62
	ROUGE-1	13.66	14.05	**17.21**	16.05 $\pm$ 0.46	16.67 $\pm$ 0.39
	ROUGE-2	9.74	10.71	**16.29**	14.92 $\pm$ 0.61	15.72 $\pm$ 0.53
	ROUGE-L	13.14	13.65	**17.66**	16.25 $\pm$ 0.55	17.05 $\pm$ 0.46
	BERTScore	19.44	19.50	**20.97**	19.50 $\pm$ 0.43	20.30 $\pm$ 0.37
	BLEURT	23.59	23.53	**24.28**	23.47 $\pm$ 0.23	23.86 $\pm$ 0.19
GigaWord	BLEU	9.39	10.95	**13.21**	9.44 $\pm$ 0.58	12.21 $\pm$ 0.29
	ROUGE-1	14.58	15.40	**16.06**	14.44 $\pm$ 0.45	15.85 $\pm$ 0.22
	ROUGE-2	7.75	9.27	**11.84**	8.35 $\pm$ 0.51	10.71 $\pm$ 0.25
	ROUGE-L	14.31	15.13	**15.93**	13.81 $\pm$ 0.48	15.60 $\pm$ 0.24
	BERTScore	19.59	**19.81**	19.51	18.84 $\pm$ 0.39	19.71 $\pm$ 0.18
	BLEURT	20.23	**20.41**	20.28	20.19 $\pm$ 0.21	20.40 $\pm$ 0.11

Table 8: The Pearson correlations with human judgement on the abstractive text summarization task.

Dataset	Kendall Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
WebNLG	BLEU	15.94	16.76	19.25	18.65 $\pm$ 0.20	**19.49** $\pm$ 0.13
	METEOR	21.30	22.03	**23.08**	20.43 $\pm$ 0.25	22.42 $\pm$ 0.17
	ROUGE-L	15.24	16.16	**18.74**	17.92 $\pm$ 0.20	**18.74** $\pm$ 0.13
	CIDEr	13.40	13.59	14.43	**14.56** $\pm$ 0.05	14.52 $\pm$ 0.03
	BERTScore	23.44	23.68	**24.19**	22.93 $\pm$ 0.19	23.90 $\pm$ 0.11
	BLEURT	24.31	24.38	**24.92**	24.54 $\pm$ 0.09	24.84 $\pm$ 0.05
E2ENLG	BLEU	11.53	**13.32**	11.99	8.80 $\pm$ 1.29	11.04 $\pm$ 0.65
	METEOR	18.46	**18.86**	14.37	11.08 $\pm$ 1.20	13.41 $\pm$ 0.67
	ROUGE-L	8.60	9.59	**11.60**	9.49 $\pm$ 1.03	10.82 $\pm$ 0.46
	CIDEr	10.29	**12.45**	11.56	9.41 $\pm$ 1.06	11.03 $\pm$ 0.56
	BERTScore	14.45	**14.85**	14.47	12.72 $\pm$ 0.95	13.90 $\pm$ 0.45
	BLEURT	14.61	14.80	**15.26**	14.86 $\pm$ 0.40	15.08 $\pm$ 0.19
WikiBioNLG	BLEU	3.27	**3.61**	2.61	2.01 $\pm$ 0.58	2.31 $\pm$ 0.33
	METEOR	3.73	**4.30**	3.03	2.43 $\pm$ 0.53	2.65 $\pm$ 0.30
	ROUGE-L	4.00	**4.27**	4.17	3.39 $\pm$ 0.46	3.89 $\pm$ 0.30
	CIDEr	2.10	**2.60**	2.22	1.39 $\pm$ 0.44	1.67 $\pm$ 0.28
	BERTScore	5.09	**5.18**	4.58	4.62 $\pm$ 0.35	4.73 $\pm$ 0.18
	BLEURT	5.32	**5.43**	4.87	4.85 $\pm$ 0.29	4.90 $\pm$ 0.13

Table 9: The Kendall correlations with human judgement on the data-to-text task.

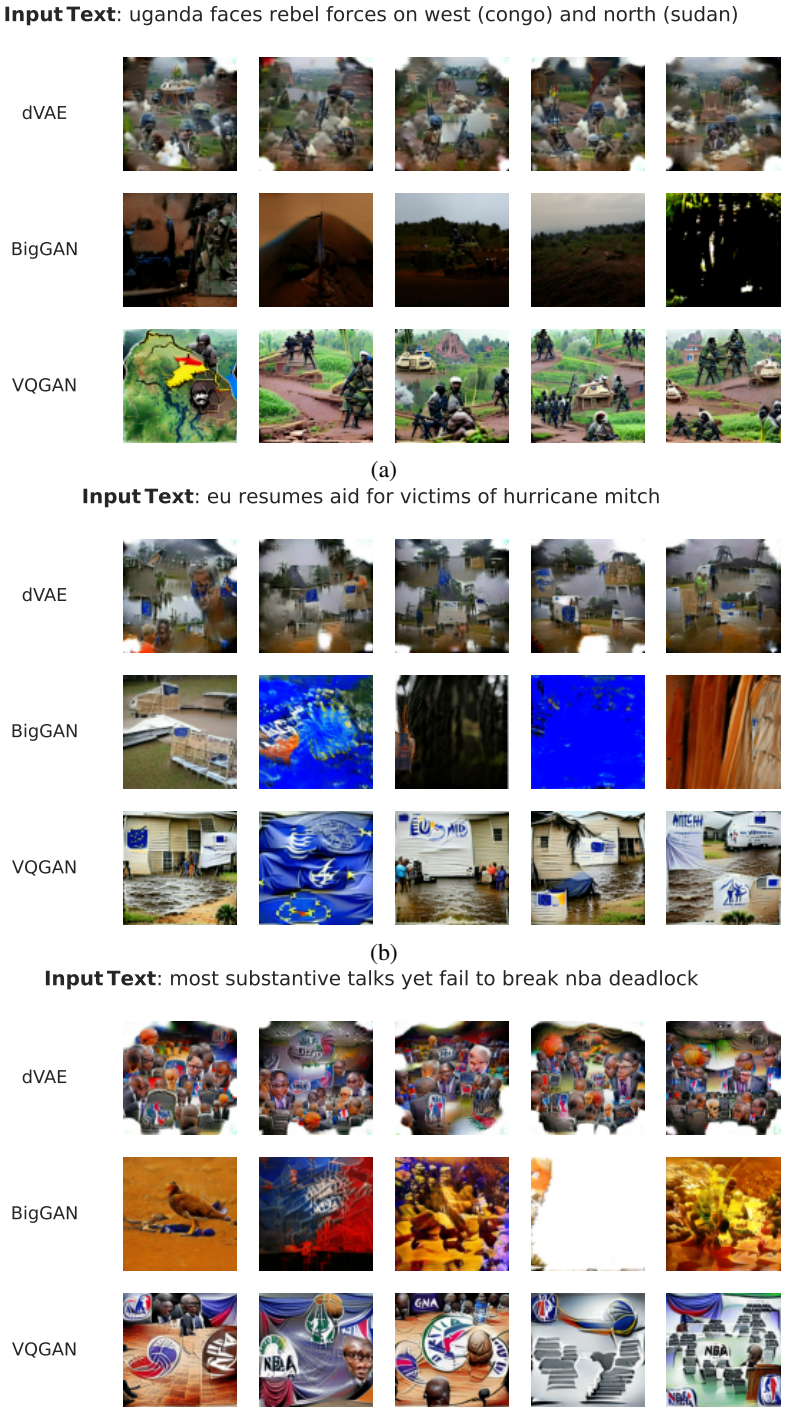


Figure 10: Groups of images generated by IMAGINE with different image genrative backbones with random initializations. The image generative backbones are dVAE, BigGAN and VQGAN.

Dataset	Pearson Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
WebNLG	BLEU	25.79	26.80	**30.04**	28.72 ± 0.22	29.86 ± 0.16
	METEOR	30.78	31.88	**33.50**	30.94 ± 0.27	33.08 ± 0.20
	ROUGE-L	24.15	25.23	**28.70**	27.66 ± 0.21	28.59 ± 0.15
	CIDEr	23.09	23.25	23.98	**24.07** ± 0.04	24.02 ± 0.02
	BERTScore	34.53	34.84	**35.82**	34.11 ± 0.19	35.38 ± 0.12
	BLEURT	35.97	36.00	**36.80**	36.26 ± 0.10	36.62 ± 0.06
E2ENLG	BLEU	12.78	14.66	**19.11**	13.93 ± 1.94	17.23 ± 0.76
	METEOR	25.55	**25.93**	23.37	17.85 ± 1.80	21.44 ± 0.69
	ROUGE-L	12.22	13.48	**18.69**	14.62 ± 1.66	17.22 ± 0.62
	CIDEr	13.83	14.27	**16.46**	15.45 ± 0.73	16.06 ± 0.26
	BERTScore	22.76	23.15	**23.71**	20.18 ± 1.27	22.40 ± 0.47
	BLEURT	22.75	22.96	**23.68**	22.55 ± 0.59	23.22 ± 0.21
WikiBioNLG	BLEU	8.19	**9.25**	5.67	4.88 ± 0.82	5.73 ± 0.45
	METEOR	8.31	**9.35**	6.33	5.58 ± 0.75	6.36 ± 0.40
	ROUGE-L	9.88	**10.51**	8.16	7.40 ± 0.69	8.23 ± 0.36
	CIDEr	5.35	5.78	5.92	5.85 ± 0.39	**5.97** ± 0.19
	BERTScore	8.98	**9.24**	8.22	7.78 ± 0.47	8.19 ± 0.23
	BLEURT	9.21	**9.39**	8.84	8.58 ± 0.32	8.80 ± 0.15

Table 10: The Pearson correlations with human judgement on the data-to-text task.

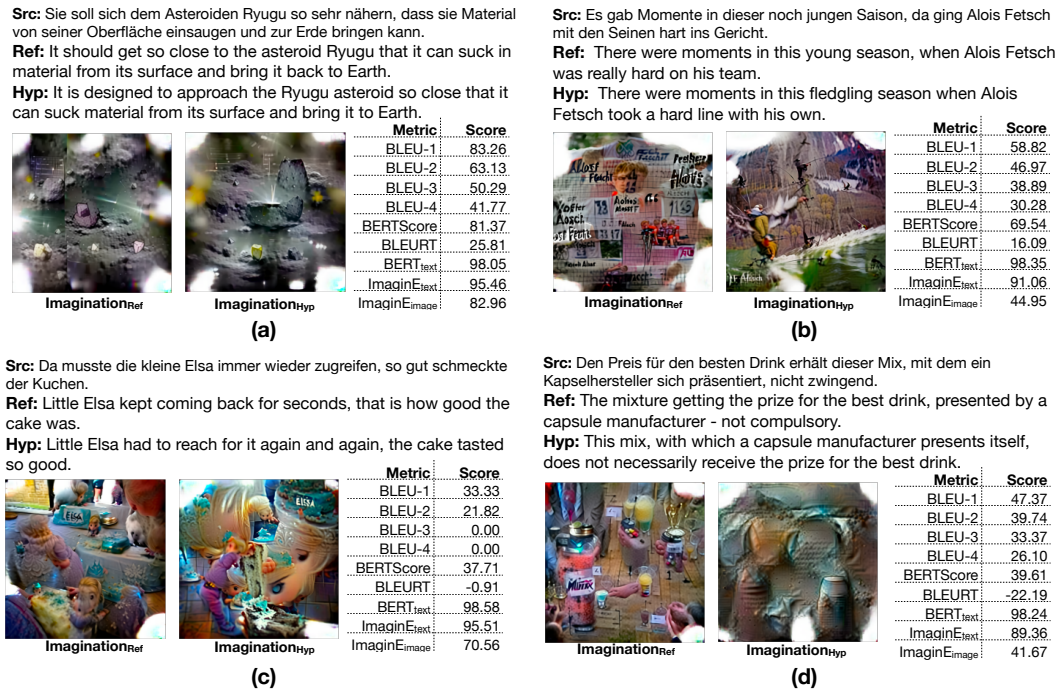


Figure 11: More examples for the machine translation task on WMT'19. **Src**: the German text to be translated. **Ref**: the reference translation. **Hyp**: the generated translation candidate.

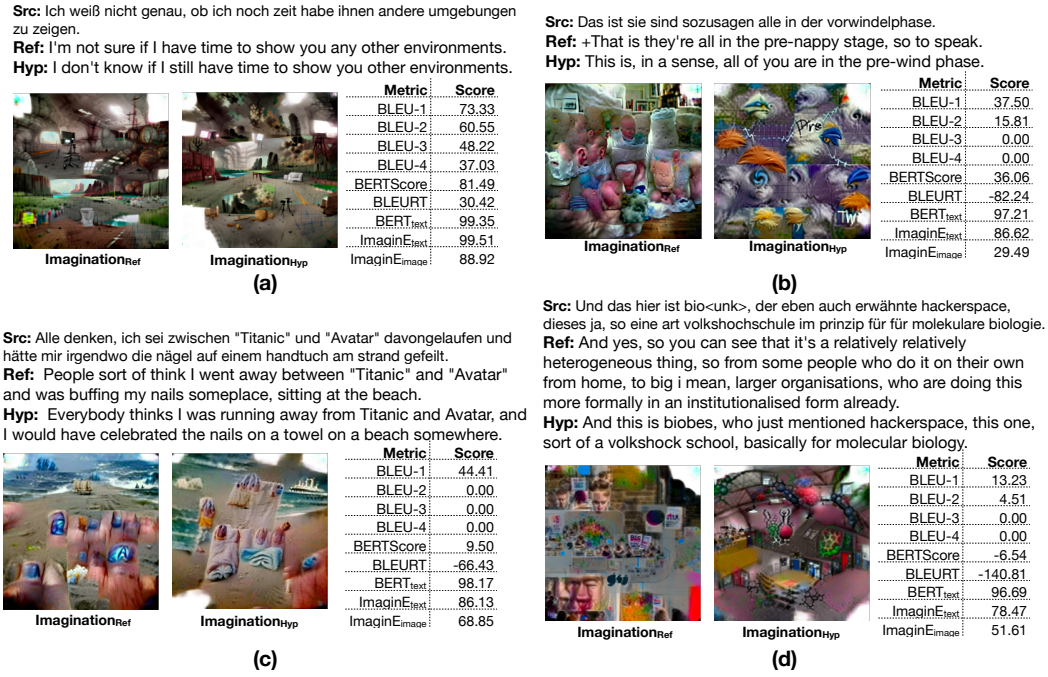


Figure 12: More examples for the machine translation task on IWSLT'14. **Src**: the German text to be translated. **Ref**: the reference translation. **Hyp**: the generated translation candidate.

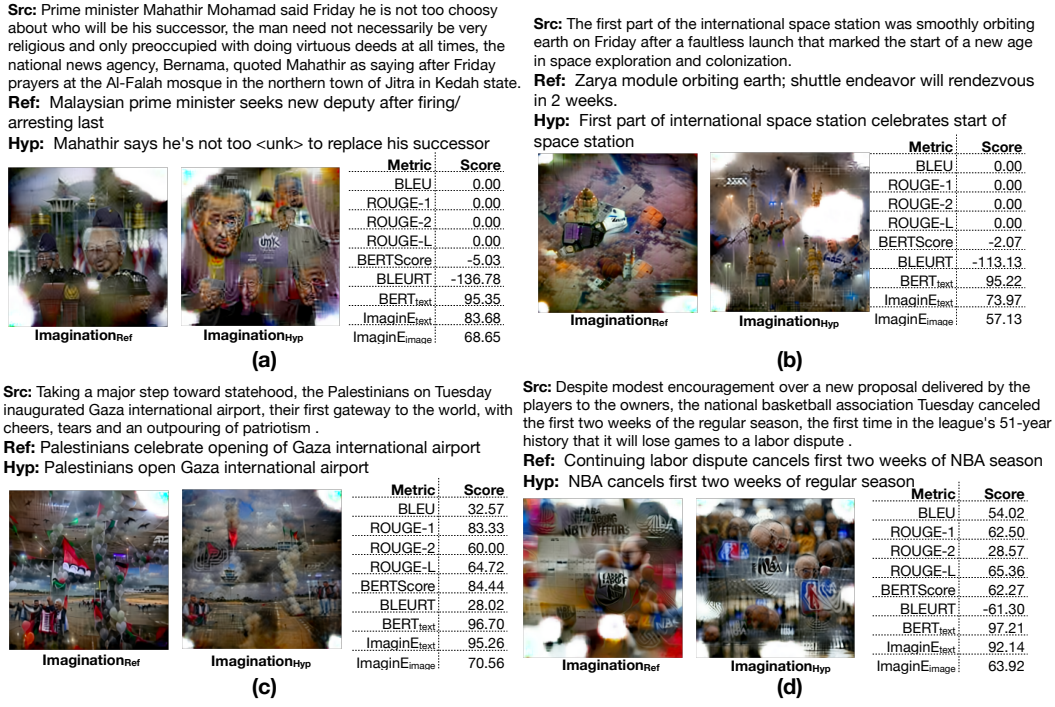


Figure 13: More examples for the abstractive text summarization task on DUC2004. **Src**: the text to be summarized. **Ref**: the reference summary. **Hyp**: the generated summary candidate.

Src: Opec's president Ammar UNK arrived late Friday in Qatar on the third stage of a tour of the oil states of the gulf.
Ref: Opec president arrives in Qatar on next stage of gulf tour
Hyp: Opec president arrives in Qatar on third stage of gulf tour

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	81.23
ROUGE-1	100.00
ROUGE-2	100.00
ROUGE-L	90.91
BERTScore	93.32
BLEURT	70.14
$\text{BERT}_{\text{text}}$	99.82
$\text{ImaginE}_{\text{text}}$	99.46
$\text{ImaginE}_{\text{image}}$	82.28

(a)

Src: Five French trekkers and mountaineers were among six foreigners killed during the autumn climbing season in Nepal, the French embassy and Nepalese officials said Friday.
Ref: Nepal climbing season claims Five french lives
Hyp: Five french climber Canadians among six killed in Nepal

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	13.44
ROUGE-1	33.33
ROUGE-2	0.00
ROUGE-L	25.58
BERTScore	28.73
BLEURT	-59.89
$\text{BERT}_{\text{text}}$	95.10
$\text{ImaginE}_{\text{text}}$	91.89
$\text{ImaginE}_{\text{image}}$	75.93

(b)

Src: Around ### clandestine immigrants Wednesday staged a peaceful breakout from a detention center in Malta and demonstrated on a road shouting ``we want freedom.
Ref: Mass breakout of immigrants from Malta center
Hyp: Malta's immigrants stage peaceful breakout from detention center

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	13.89
ROUGE-1	80.00
ROUGE-2	0.00
ROUGE-L	38.36
BERTScore	36.27
BLEURT	-25.00
$\text{BERT}_{\text{text}}$	96.36
$\text{ImaginE}_{\text{text}}$	78.08
$\text{ImaginE}_{\text{image}}$	54.25

(c)

Src: Billionaire basketball team owner mark Cuban was a no show, but the head of UNICEF made it and pop star prince rounded off the evening by throwing a guitar over his head.
Ref: UNK awards crown their prince by Giles Hewitt
Hyp: Billionaire donates guitar to Cuban billionaire

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	0.00
ROUGE-1	0.00
ROUGE-2	0.00
ROUGE-L	0.00
BERTScore	-12.14
BLEURT	-153.43
$\text{BERT}_{\text{text}}$	92.12
$\text{ImaginE}_{\text{text}}$	60.79
$\text{ImaginE}_{\text{image}}$	28.49

(d)

Figure 14: More examples for the abstractive text summarization task on GigaWord. **Src**: the text to be summarized. **Ref**: the reference summary. **Hyp**: the generated summary candidate.

Ref: Beef kway teow originates from Singapore and is also made in Indonesia. One of the ingredients in the dish is oyster sauce.
Hyp: Oyster sauce is a dish from Singapore, where Oyster sauce is a dish from Indonesia.

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	3.91
METEOR	19.14
ROUGE-L	15.21
CIDEr	5.38
BERTScore	39.04
BLEURT	-34.61
$\text{BERT}_{\text{text}}$	93.13
$\text{ImaginE}_{\text{text}}$	73.73
$\text{ImaginE}_{\text{image}}$	68.85

(a)

Ref: The capital of Turkey is Ankara and the leader is Ahmet Davutoglu. Turkey is the location of the bronze Ataturk monument.
Hyp: Atatürk Monument, located in Ankara, Turkey, is the capital of Turkey.

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	5.82
METEOR	21.55
ROUGE-L	17.15
CIDEr	3.37
BERTScore	34.12
BLEURT	-6.13
$\text{BERT}_{\text{text}}$	98.35
$\text{ImaginE}_{\text{text}}$	86.13
$\text{ImaginE}_{\text{image}}$	64.45

(b)

Ref: Athens International Airport serves the city of Athens, in Greek speaking Greece. Two of the leaders of which are Alexis Tsipras and Prokopis Pavlopoulos.
Hyp: Alexis Tsipras (Greek: Αντανίς Φεμπελος) is a commander of the city of Athens, Greece, where Alexis Tsipras is the leader.

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	23.76
METEOR	20.60
ROUGE-L	24.81
CIDEr	116.75
BERTScore	-5.23
BLEURT	-52.37
$\text{BERT}_{\text{text}}$	96.12
$\text{ImaginE}_{\text{text}}$	80.08
$\text{ImaginE}_{\text{image}}$	48.61

(c)

Ref: Batagor, which is found in Indonesia and Siomay are both variations of the same dish. Shumai is a variation of the dish Siomay while Shiitake is an ingredient of Shumai.
Hyp: The main ingredient of binagor is ``shumai'', a dish from Indonesia where the main ingredient is ``shitake''.

$\text{Imagination}_{\text{Ref}}$ $\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	8.06
METEOR	13.43
ROUGE-L	18.59
CIDEr	8.62
BERTScore	18.59
BLEURT	-46.34
$\text{BERT}_{\text{text}}$	98.96
$\text{ImaginE}_{\text{text}}$	89.50
$\text{ImaginE}_{\text{image}}$	79.15

(d)

Figure 15: More examples for the data-to-text task on WebNLG. **Ref**: the reference text. **Hyp**: the generated text candidate.

Ref: Wildwood is a pub located in riverside area near Raja Indian Cuisine. It serves Italian food and It is not family-friendly.
Hyp: Wildwood is a variation of Raja Cuisine.

$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	3.94
METEOR	12.30
ROUGE-L	25.23
CIDEr	4.10
BERTScore	34.53
BLEURT	-84.00
$\text{BERT}_{\text{text}}$	95.47
$\text{ImaginE}_{\text{text}}$	81.01
$\text{ImaginE}_{\text{image}}$	46.17

(a)

Ref: A restaurant that is kid friendly near Raja Indian Cuisine named The Wrestlers in the riverside area has a price range of more than £30 that serves Italian food.
Hyp: Raja Cuisine is a variation of The Wrestlers.

$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	1.36
METEOR	9.78
ROUGE-L	19.61
CIDEr	0.02
BERTScore	19.43
BLEURT	-118.30
$\text{BERT}_{\text{text}}$	96.65
$\text{ImaginE}_{\text{text}}$	72.56
$\text{ImaginE}_{\text{image}}$	50.73

(b)

Ref: Located near Rainbow Vegetarian Café on the river, The Vaults is a low cost, family friendly pub.
Hyp: The main ingredients of a riverside riverside riverside riverside riverside riverside riverside rivers.

$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	2.38
METEOR	2.60
ROUGE-L	10.73
CIDEr	0.07
BERTScore	-17.89
BLEURT	-140.35
$\text{BERT}_{\text{text}}$	94.61
$\text{ImaginE}_{\text{text}}$	68.02
$\text{ImaginE}_{\text{image}}$	36.30

(c)

Ref: The Punter near Rainbow Vegetarian Café in the riverside as a restaurant with a high price range is not children friendly. They provide Italian food with a customer rating 1 out of 5.
Hyp: The Punter is a variation of the Rainbow Vegetable.

$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	1.69
METEOR	7.71
ROUGE-L	21.05
CIDEr	0.01
BERTScore	13.96
BLEURT	-135.02
$\text{BERT}_{\text{text}}$	95.08
$\text{ImaginE}_{\text{text}}$	75.05
$\text{ImaginE}_{\text{image}}$	48.29

(d)

Figure 16: More examples for the data-to-text task on E2ENLG. **Ref**: the reference text. **Hyp**: the generated text candidate.

Ref: Eden Ants was a Canadian indie rock band from Toronto, founded in 2000 by the Montreal-born ender brothers.
Hyp: Ants Eden is a synthpop guitar player.

$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	1.67
METEOR	8.19
ROUGE-L	13.63
CIDEr	7.31
BERTScore	10.40
BLEURT	-119.97
$\text{BERT}_{\text{text}}$	98.47
$\text{ImaginE}_{\text{text}}$	87.79
$\text{ImaginE}_{\text{image}}$	57.91

(a)

Ref: Rose mortem is the fashion label and nom de guerre of American fashion designer Rose Hemlock.
Hyp: Mortem is a fashion design type used in the comics.

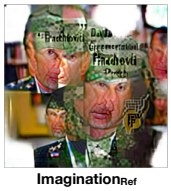
$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	7.58
METEOR	12.60
ROUGE-L	21.23
CIDEr	50.60
BERTScore	11.32
BLEURT	-96.62
$\text{BERT}_{\text{text}}$	96.89
$\text{ImaginE}_{\text{text}}$	82.42
$\text{ImaginE}_{\text{image}}$	52.10

(b)

Ref: David P. Fridovich is a retired lieutenant general and green beret in the United States army.
Hyp: David P. Fridovich is the general manager of the United States united Force.



$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	27.95
METEOR	27.96
ROUGE-L	45.57
CIDEr	205.81
BERTScore	47.20
BLEURT	-39.11
$\text{BERT}_{\text{text}}$	98.47
$\text{ImaginE}_{\text{text}}$	87.79
$\text{ImaginE}_{\text{image}}$	57.91

(c)

Ref: Byzantine is a heavy metal band from Charleston, West Virginia that formed in 2000.
Hyp: The band Cerzantine Trombony skip is the independent name of the band.

$\text{Imagination}_{\text{Ref}}$

$\text{Imagination}_{\text{Hyp}}$

Metric	Score
BLEU	2.99
METEOR	6.07
ROUGE-L	6.96
CIDEr	9.14
BERTScore	-0.37
BLEURT	-122.91
$\text{BERT}_{\text{text}}$	96.47
$\text{ImaginE}_{\text{text}}$	77.24
$\text{ImaginE}_{\text{image}}$	42.70

(d)

Figure 17: More examples for the data-to-text task on WikiBioNLG. **Ref**: the reference text. **Hyp**: the generated text candidate.