IMAGINE: An Imagination-Based Automatic Evaluation Metric for Natural Language Generation

Anonymous ACL submission

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 and 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., 2015a), SMD (Clark et al., 2019a), 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.

But what happens in our minds when we read, comprehend, and evaluate text? Research (Just et al., 2004; Eviatar and Just, 2006) has found that, unlike commonly designed automatic evaluation 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. Cognitive studies show that visual imagery improves comprehension during human language processing (Gambrell and Bales, 1986; Joffe et al., 2007; Sadoski and Paivio, 2013). 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 recent 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. This enables us to explore the possibility of incorporating multi-modal information into NLG evaluation.

In this work, we propose IMAGINE, an imagination-based automatic evaluation metric for text generation. Specifically, IMAGINE first uses the pre-trained discrete variational autoencoder (dVAE) from the vision-language model DALL-E (Ramesh et al., 2021) to visualize imagination from sentences, which is to generate descriptive images for the candidate text and the references. Then it 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) for evaluation. Figure 1 shows an example.

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

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.

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?*

3. *What are the potentials and limitations of introducing imaginations with* IMAGINE *to NLG evaluation?*

Experimental results show that IMAGINE can serve as a complementary evaluation metric to text-based ones, and adding IMAGINE similarity scores to existing metrics surprisingly improves most of the popular metrics' correlations with human performance on a variety of text generation tasks. We further conduct comprehensive quantitative analyses with case studies to verify its effectiveness. 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 and 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 and 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., 2015b; Rubner et al., 1998; Clark et al., 2019b; 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, some metrics 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 metric for image captioning by directly comparing images with captions using 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.

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 and 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 image generation method 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 and Bales, 1986; Sadoski and Paivio, 1994; Nippold and Duthie, 2003; Just et al., 2004; Joffe et al., 2007; Sadoski and Paivio, 2013), 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 (van den Oord et al., 2018).

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.

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

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

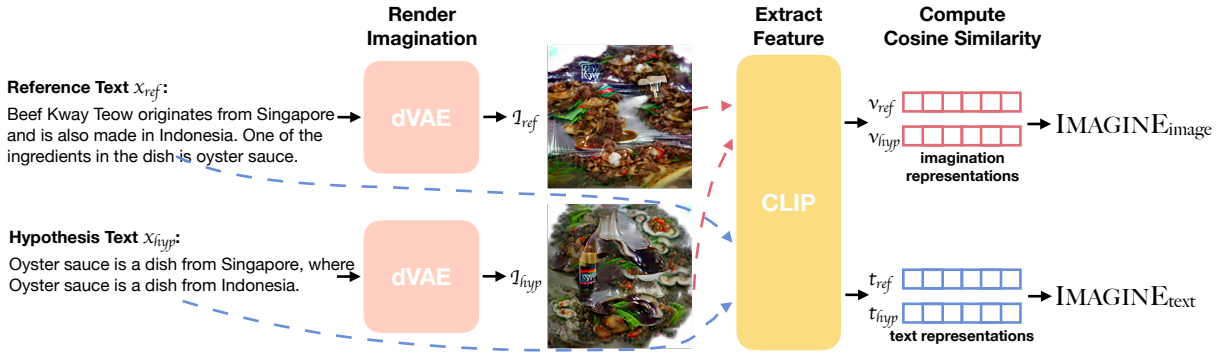


Figure 2: IMAGINE similarity score computation process. Given the reference text $\boldsymbol{x}_{ref}$ and the generated hypothesis $\boldsymbol{x}_{hyp}$, we visualize the machine imagination $\boldsymbol{I}_{ref}$ and $\boldsymbol{I}_{hyp}$ with the pre-trained dVAE. We extract features for the pair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

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.

$$\text{IMAGINE}_{text} = \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{t}_{hyp}^T\boldsymbol{t}_{ref_i}}{\|\boldsymbol{t}_{hyp}\|\|\boldsymbol{t}_{ref_i}\|}\right) \quad (2)$$

$$\text{IMAGINE}_{image} = \mathcal{L}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{v}_{hyp}^T\boldsymbol{v}_{ref_i}}{\|\boldsymbol{v}_{hyp}\|\|\boldsymbol{v}_{ref_i}\|}\right) \quad (3)$$

Here, $\mathcal{L}$ is a linear function that stretch the score distribution to the range of $[0, 1]$. More details can be found in Section 5.

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, which also naturally mimics the process of human text comprehension where text and visual imagination are both used. 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:

$$metric_score' \mathrel{+}= \text{IMAGINE}_{similarity_score} \quad (4)$$

4 Experiments

4.1 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, 2020) and WikiBioNLG (Lebret et al., 2016). We use the text generated by the KGPT (Chen et al., 2020) model in our experiments. Table 1 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

[1] https://duc.nist.gov/duc2004/
[2] https://catalog.ldc.upenn.edu/LDC2011T07

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 1: 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.

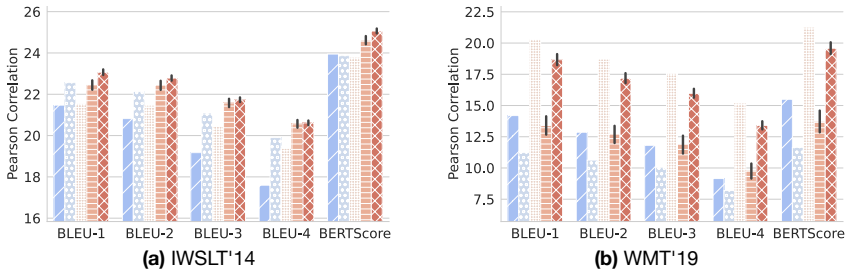


Figure 3: The effectiveness of augmenting BLEU-n (n=1,2,3,4) and BERTScore with IMAGINE similarities and BERT$_{text}$ similarity on two machine translation datasets. The y-axis shows the Pearson correlation with human judgments.

five automatic metrics for NLG, including BLEU, ROUGE-L, METEOR (Elliott and 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, containing 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.

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

4.2 Results

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. BERT$_{text}$ has mixed performance on the two machine translation task. It improves IWSLT'14 while lowers correlation on WMT'19. On the other hand, our IMAGINE$_{text\&image}$ steadily improves all the listed metrics' correlations with human scores. IMAGINE$_{image}$ and IMAGINE$_{text\&image}$ contributes the most in IWSLT'14 while IMAGINE$_{text}$ and IMAGINE$_{text\&image}$ play important roles in WMT'19.

Abstractive Text Summarization Figure 4 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. Similar to its impact on the

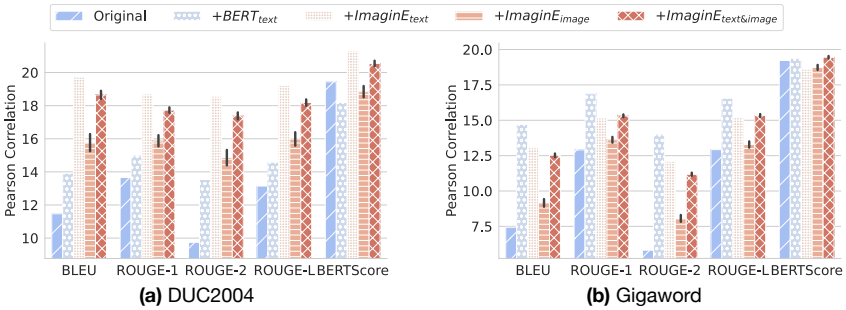


Figure 4: The effectiveness of augmenting BLEU, BERTScore and ROUGE-related metrics with IMAGINE similarities and BERT$_{text}$ similarity on two abstractive text summarization datasets. The y-axis shows the Pearson correlation with human judgments.

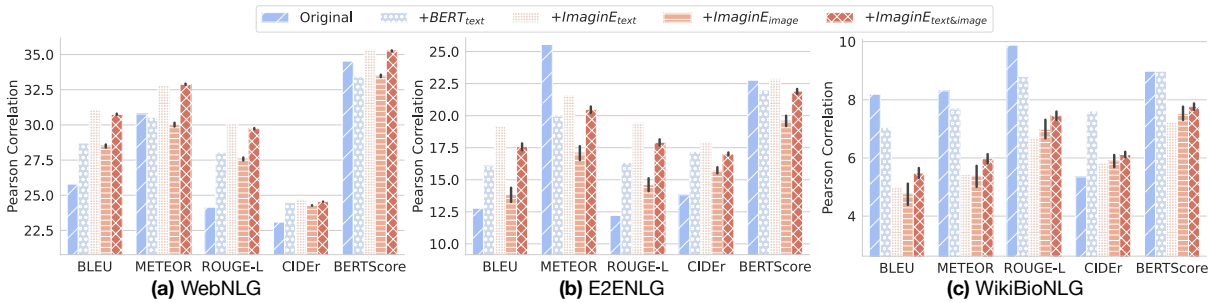


Figure 5: The effectiveness of augmenting BLEU, METEOR, ROUGE-L, CIDEr, and BERTScore with IMAGINE similarities and BERT$_{text}$ similarity on three data-to-text generation datasets. The y-axis shows the Pearson correlation with human judgments.

machine translations datasets, IMAGINE$_{text\&image}$ steadily improves the correlation with human judgments of BLEU, ROUGE-related metrics, and BERTScore on the two summarization datasets. The metric that has the most significant impact on DUC2004 and GigaWord is IMAGINE$_{text}$ and BERT$_{text}$ respectively.

Data-to-Text Generation Figure 5 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.

On WebNLG, adding IMAGINE$_{text}$ and IMAGINE$_{text\&image}$ can steadily improve all the listed metrics' correlation with human scores. IMAGINE$_{image}$ improves BLEU, ROUGE-L, and CIDEr but it only has limited impact on METEOR and BERTScore. Among the two metrics that compare textual similarity, IMAGINE$_{text}$ boosts correlations more than BERT$_{text}$.

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. IMAGINE$_{text}$ outperforms BERT$_{text}$ in all listed metrics. On the other hand, IMAGINE$_{image}$ has mixed performance, improving BLEU, ROUGE-L and CIDEr while yielding negative effects on METEOR and BERTScore. The E2ENLG dataset is built from the restaurant domain, where irrelevant high-frequency tokens such as restaurant names can misguide the visualization process of IMAGINE.

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, which contains many abstract concepts that may be hard to visualize. Figure 5(c) 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.

5 Discussion

Why is ImaginE helpful? As shown in Figures 3 to 5, 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 knowl-

Src: Also entschied ich mich eines tages den filialleiter zu besuchen, und ich fragte den leiter, "funktioniert dieses modell, dass sie den menschen all diese möglichkeiten bieten wirklich?"
Ref: So I one day decided to pay a visit to the manager, and I asked the manager, "is this model of offering people all this choice really working?"
Hyp: So I decided to visit the filialler one day, and I asked the ladder, "does this model work that you really offer to the people all these possibilities?"

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

Metric	Score
BLEU-1	69.70
BLEU-4	20.26
BERTScore	66.62
$\text{ImaginE}_{\text{image}}$	13.08
Human	3.1/5.0

Figure 6: A case study on IWSLT'14 dataset.

Metric	Original	$+\text{IE}_{i(dVAE)}$	$+\text{IE}_{i(BigGAN)}$	$+\text{IE}_{i(VQGAN)}$
ROUGE-1	13.7	**15.9** $\pm$ 0.9	15.7 $\pm$ 1.0	**15.9** $\pm$ 0.8
ROUGE-2	9.7	**14.9** $\pm$ 1.2	14.6 $\pm$ 1.3	**14.9** $\pm$ 1.0
ROUGE-L	13.1	**16.0** $\pm$ 1.0	15.8 $\pm$ 1.1	**16.0** $\pm$ 0.9
BERTScore	**19.4**	18.8 $\pm$ 0.9	18.7 $\pm$ 1.0	18.9 $\pm$ 0.8
BLEURT	**23.6**	23.0 $\pm$ 0.5	22.9 $\pm$ 0.6	23.1 $\pm$ 0.4

Table 2: The Pearson correlations with human judges when using IMAGINE_{image} (IE_i) to augment ROUGE, BERTScore, and BLEURT on DUC2004. We compute three sets of IMAGINE_{image} similarity scores (mean$\pm$std) with three different image generation backbones, namely dVAE, BigGAN, and VQGAN.

edge. However, how do these machine imaginations actually help text understanding and evaluation? In this section, we explore further on how and why IMAGINE works. We first provide a case study to show the uniqueness of IMAGINE over text-based metrics, then systematically analyze the effectiveness of our method from different perspectives.

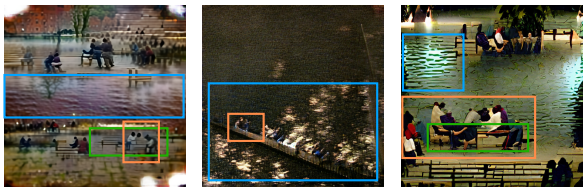
Case Study Figure 6 presents such an example in which IMAGINE captures the keyword difference between two text snippets. Regardless of the similar sentence structure between the reference and the hypothesis, the main difference is mentioning "manager" and "ladder". While other metrics score high, the quality of the generated text is questionable. IMAGINE renders distinctive images and yields a relatively low visual similarity score, which aligns with human judgment. More case studies can be found in appendix.

Sensitivity to Different Image Generation Backbones In previous sections, we implement IMAGINE with dVAE as the image generation backbone. There also appear other image generation models that use BigGAN (Brock et al., 2019) and VQGAN (Esser et al., 2021) to render images. 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 2, we find comparable IMAGINE_{image} performances for dVAE and VQGAN, both consistently surpass BigGAN on all metrics. The VQGAN leads to smaller variance on average. These models with different architectures and training data provide divergent machine imaginations and impact final text evaluation results.

	dVAE	BigGAN	VQGAN
Entity Recall	88.8%	41.2%	87.2%

Table 3: The entity recall rate on the visualizations for Flickr30k captions. We report the results for images generated by dVAE, BigGAN and VQGAN.

People sitting at a bench talking to each other by a body of water .



dVAE BigGAN VQGAN

Figure 7: An example caption from Flickr30k Entities dataset, and images rendered by dVAE, BigGAN and VQGAN. The bounding boxes point to the visualizations of the entities marked in the same color.

Reliability of Imaginations We further verify the reliability of IMAGINE's visualization on Flickr30k Entities dataset (Plummer et al., 2015). This image captioning dataset contains annotations on the entities mentioned in each caption for evaluation. We randomly sample 100 captions and use the three generative backbones to render images with 500-step learning phases. We present the caption and the rendered image to human annotators, and ask them whether the entities mentioned in the caption are visualized. Entity recall rates are reported in Table 3. The dVAE has the highest entity recall rate of around 89%, followed closely by VQGAN. BigGAN has the lowest recall rate of around 41%. Figure 7 shows a group of example of entity recall in visualization. The observations are also consistent with human correlations in Table 2, that higher quality of imaginations improves text evaluation.

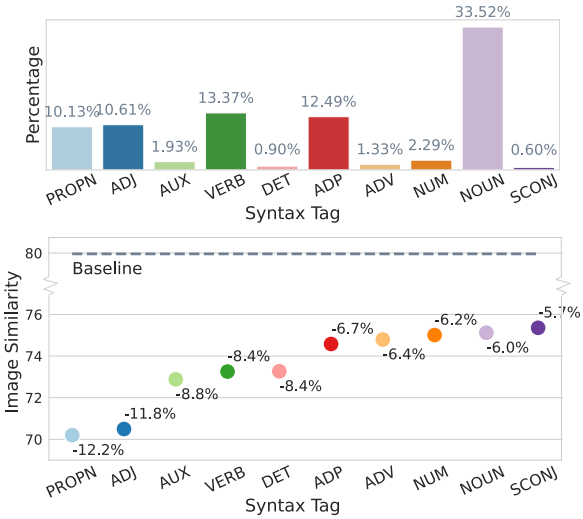


Figure 8: The influence on visualization when masking tokens of different syntax tags. Upper: The occurrence frequency of each syntax tag in DUC2004. Lower: The relative image similarity decrease after masking each syntax tag. Baseline: The average intra-group pairwise image similarity. The top-10 syntax tags that have the most significant impact on visualization are listed here.

Syntax Importance to Imaginations We assess the importance of each type of syntax token in the image generation process. We use Stanza (Qi et al., 2020) part-of-speech (POS) tagger to parse the text in DUC2004 test set. For each syntax tag[4], we create ablated test examples by masking out a token of that tag from the original text. We compare the visual similarity of the images rendered from the ablated examples to the visualization of the vanilla text. Figure 8 reports the influence of masking each syntax tag. PROPN and ADJ are two tags that have a salient impact on visualization results, and removing them leads to a relatively 12% drop in visual similarity. Surprisingly, removing NOUN tokens has a comparably smaller influence on image rendering. Table 4 lists out the most frequent NOUN, PROPN, and ADJ tokens in DUC2004, a dataset built upon news clusters. PROPN and ADJ tokens in DUC2004 cover concrete concepts such as nations, corporations, and celebrities. NOUN tokens involve more abstract concepts such as government, party, and right. For this particular dataset, our IMAGINE approach pays more attention to PROPN and ADJ tokens that are easier to visualize by nature. More analysis for other dataset domains can be found in the appendix.

[4]We report Universal POS tags in this study: https://universaldependencies.org/u/pos/

POS Tag	10 Most Frequent Tokens
NOUN	president, minister, government, space, party, station, budget, game, right, arrest
PROPN	U.S., Clinton, China, Korea, Gaza, Microsoft, Congo, Israel, Livingston, Lebanon
ADJ	new, prime, Russian, international, Asian, possible, Cambodian, first, human, economic

Table 4: The most frequent NOUN, PROPN, and ADJ tokens in DUC2004.

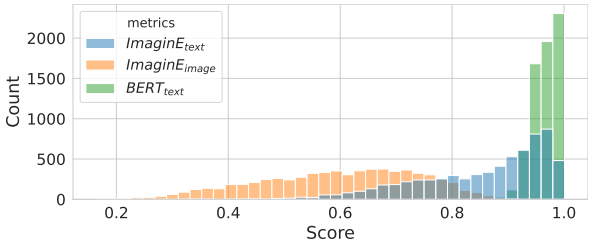


Figure 9: The score distributions of IMAGINE$_{text}$, IMAGINE$_{image}$, and BERT$_{text}$ before re-scaling.

Score Distributions We visualize the score distributions of different metrics in Figure 9. CLIP$_{image}$ mostly lies between $[0.25, 1]$ and CLIP$_{text}$ is in $[0.5, 1]$. Overall, our imagination-based methods lead to smoother distributions. CLIP$_{image}$ is more diverse than text-based metrics with the same measurement (i.e., cosine similarity). Following CLIP-Score (Hessel et al., 2021), we linearly normalize the score distributions to $[0, 1]$, which is also the score range for most of the automatic metrics covered in this study. More specifically, the similarity score s will be re-scaled by:

$$s' = \frac{s - l}{1 - l}, \quad l = \begin{cases} 0.50, & \text{for} \quad \text{IMAGINE}_{text}, \\ 0.25, & \text{for} \quad \text{IMAGINE}_{image}, \\ 0.90, & \text{for} \quad \text{BERT}_{text}. \end{cases} \tag{5}$$

6 Conclusion

In this paper, we propose IMAGINE, an imagination-based automatic evaluation metric for NLG. Experiments on seven datasets across three different tasks indicate that adding IMAGINE similarity scores as an extension to current automatic NLG metrics can improve their correlations with human judgments in many circumstances. In the future, it is interesting to explore effective ways of visualizing abstract concepts, and how to generate imaginations efficiently. 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[5], BigSleep(BigGAN+CLIP)[6] and VQGAN+CLIP[7].

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. 2019. Findings of the 2019 conference on machine translation (wmt19). In *WMT*.

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

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

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. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

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

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

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

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

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

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

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

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. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

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

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

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

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

Zohar Eviatar and Marcel Adam Just. 2006. Brain correlates of discourse processing: An fmri investigation of irony and conventional metaphor comprehension. *Neuropsychologia*, 44(12):2348–2359.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

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

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

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

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

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

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *ACL*.

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. 2021. Learning transferable visual models from natural language supervision. *ArXiv*, abs/2103.00020.

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

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

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

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

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

Mark Sadoski and Allan Paivio. 2013. *Imagery and text: A dual coding theory of reading and writing*. Routledge.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Appendix

A.1 Random Initialization

We assess the influence of random initialization by repeating the image generation process five times on the DUC2004 test set and computing pairwise visual similarities within each group of 5 images. Notice in Figure 10 that dVAE and VQGAN have similar intra-group visual similarity distributions, with VQGAN slightly higher on average. There appear extremely low values for all three backbones, which suggests that random initialization should be taken into consideration. In Figure 12, we show several groups of images generated by dVAE, BigGAN and VQGAN with random initialization.

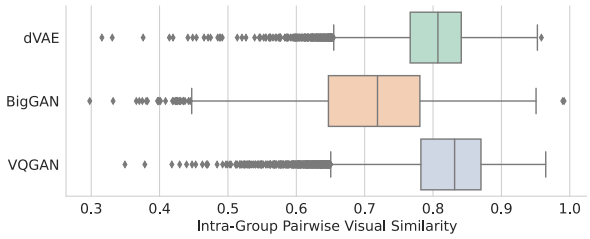


Figure 10: The intra-group pairwise visual similarity distributions for images generated by dVAE, BigGAN, and VQGAN. The plot shows the three quartile values and the extreme values.

A.2 Syntax Importance to Imaginations

In Section 5, we discuss the importance of DUC2004 POS tags on imaginations. Here we display the syntax importance to visualization on another dataset domain. Figure 11 shows the syntax importance on text examples from Flickr30k Entities (Plummer et al., 2015), which is an image captioning dataset. The ranking of the most influential POS tags is different from the results on DUC2004 in Figure 8. However, the results on Flickr30k also display a tendency that the concrete concepts are easier to be visualized, thus playing a more important role in visualization.

A.3 Rendering Iterations

In this paper, we use dVAE to render 512x512 images with a 1000-step learning phase. Figure 13 illustrate a few set of examples of images rendered with different iterations. The main contents of the figure only have after minor changes after 400 iterations. Future work may apply less iterations to reduce computation budget.

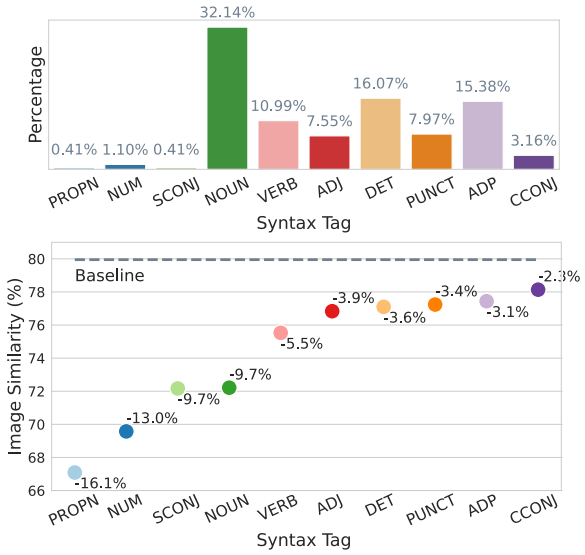


Figure 11: The influence on visualization when masking tokens of different syntax tags. Upper: The occurrence frequency of each syntax tag in DUC2004. Lower: The relative image similarity decrease after masking each syntax tag. Baseline: The average intra-group pairwise image similarity. The top-10 syntax tags that have the most significant impact on visualization are listed here.

A.4 Correlation Results

We list the numbers on Pearson correlation in Tables 8, 10 and 12 that match Figures 3 to 5 in the main paper. Table 5 lists out each metric's Pearson correlation with human judgments on each dataset. Tables 6, 7, 9 and 11 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.5 Case Study

We provide more case studies for the three NLG tasks used in our study in Figures 14 to 16.

A.6 Computation Expenses

We conduct experiments on 8 Titan RTX GPUs. It takes $\sim$ 200 hours to generate all the imagination figures used in our study. However, as discussed in Section A.3, future work may greatly cut the computational budget by reducing the rendering iterations.

A.7 Limitations and Future Work

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

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	14.19	12.86	11.81	9.15	15.50	16.14	2.54	**24.94**	3.81 $\pm$ 2.96	16.89 $\pm$ 2.70
	IWSLT14	21.47	20.82	19.17	17.60	**23.95**	22.93	18.42	14.11	15.63 $\pm$ 1.01	17.56 $\pm$ 0.75
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	13.90 $\pm$ 1.65	18.79 $\pm$ 0.85
	GigaWord	7.44	12.90	5.82	12.92	19.23	**20.20**	16.76	14.73	6.05 $\pm$ 1.00	14.70 $\pm$ 0.63
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	19.59 $\pm$ 0.53	25.99 $\pm$ 0.24
	E2ENLG	12.78	**25.55**	12.22	13.83	22.76	22.75	13.11	18.19	11.23 $\pm$ 1.70	16.12 $\pm$ 0.85
	WikiBioNLG	8.19	8.31	**9.88**	5.35	8.98	9.21	6.07	4.14	3.50 $\pm$ 1.03	4.42 $\pm$ 0.47

Table 5: The Pearson correlations with human judgement for each individual metric.

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	11.51	11.19	9.37	7.16	10.68	10.63	3.22	**17.35**	2.96 $\pm$ 2.20	11.73 $\pm$ 2.31
	IWSLT14	14.19	14.26	13.68	12.79	**16.68**	14.64	13.84	12.90	10.75 $\pm$ 0.71	12.70 $\pm$ 0.59
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	7.51 $\pm$ 1.22	9.04 $\pm$ 0.73
	GigaWord	11.56	11.08	7.77	11.77	13.92	**15.27**	12.63	11.95	3.67 $\pm$ 0.78	11.22 $\pm$ 0.58
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	12.73 $\pm$ 0.38	17.72 $\pm$ 0.23
	E2ENLG	11.53	**18.46**	8.60	10.29	14.45	14.61	10.86	10.59	6.63 $\pm$ 1.27	9.24 $\pm$ 0.51
	WikiBioNLG	3.27	3.73	4.00	2.10	5.09	**5.32**	3.07	2.08	1.45 $\pm$ 0.79	1.90 $\pm$ 0.35

Table 6: The Kendall correlations with human judgement for each individual metric.

Dataset	Kendall Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
WMT19	BLEU-1	11.51	9.29	**15.08**	10.53 $\pm$ 1.29	13.89 $\pm$ 0.92
	BLEU-2	11.19	8.86	**14.28**	9.89 $\pm$ 1.22	12.88 $\pm$ 0.88
	BLEU-3	9.37	7.64	**12.71**	9.00 $\pm$ 1.21	11.27 $\pm$ 0.86
	BLEU-4	7.16	5.55	**10.70**	7.28 $\pm$ 1.17	9.31 $\pm$ 0.76
	BERTScore	10.68	8.10	**14.83**	9.30 $\pm$ 1.38	13.36 $\pm$ 0.92
	BLEURT	10.63	9.60	**13.60**	10.91 $\pm$ 0.81	12.93 $\pm$ 0.51
IWSLT14	BLEU-1	14.19	15.43	15.42	14.92 $\pm$ 0.46	**15.75** $\pm$ 0.29
	BLEU-2	14.26	15.32	15.21	15.06 $\pm$ 0.41	**15.64** $\pm$ 0.27
	BLEU-3	13.68	14.66	14.64	14.60 $\pm$ 0.37	**15.00** $\pm$ 0.26
	BLEU-4	12.79	14.09	14.05	14.03 $\pm$ 0.34	**14.28** $\pm$ 0.22
	BERTScore	16.68	16.56	17.70	16.74 $\pm$ 0.40	**17.71** $\pm$ 0.25
	BLEURT	14.64	15.09	15.14	15.35 $\pm$ 0.20	**15.40** $\pm$ 0.11

Table 7: 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	14.19	11.23	**20.29**	13.40 $\pm$ 1.94	18.67 $\pm$ 1.11
	BLEU-2	12.86	10.63	**18.72**	12.71 $\pm$ 1.81	17.18 $\pm$ 1.01
	BLEU-3	11.81	10.06	**17.57**	11.91 $\pm$ 1.70	15.97 $\pm$ 0.94
	BLEU-4	9.15	8.19	**15.13**	9.77 $\pm$ 1.60	13.41 $\pm$ 0.88
	BERTScore	15.50	11.65	**21.21**	13.70 $\pm$ 2.11	19.58 $\pm$ 1.19
	BLEURT	16.14	14.90	**19.03**	16.72 $\pm$ 1.03	18.40 $\pm$ 0.53
IWSLT14	BLEU-1	21.47	22.56	21.56	22.45 $\pm$ 0.59	**23.07** $\pm$ 0.36
	BLEU-2	20.82	22.14	21.44	22.44 $\pm$ 0.53	**22.79** $\pm$ 0.32
	BLEU-3	19.17	21.05	20.44	21.58 $\pm$ 0.50	**21.73** $\pm$ 0.30
	BLEU-4	17.60	19.90	19.38	20.57 $\pm$ 0.49	**20.61** $\pm$ 0.30
	BERTScore	23.95	23.87	23.77	24.62 $\pm$ 0.52	**25.06** $\pm$ 0.31
	BLEURT	22.93	23.37	23.16	**24.09** $\pm$ 0.26	23.81 $\pm$ 0.14

Table 8: 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	ROUGE-1	8.71	9.69	**10.36**	9.18 $\pm$ 0.58	10.06 $\pm$ 0.28
	ROUGE-2	7.75	9.49	**9.87**	8.55 $\pm$ 0.83	9.73 $\pm$ 0.47
	ROUGE-L	7.22	9.20	**10.21**	9.02 $\pm$ 0.63	10.05 $\pm$ 0.34
	BERTScore	12.82	**12.96**	12.06	11.36 $\pm$ 0.50	12.07 $\pm$ 0.33
	BLEURT	**16.04**	15.45	15.71	15.07 $\pm$ 0.31	15.55 $\pm$ 0.18
GigaWord	ROUGE-1	11.08	**12.66**	11.52	10.34 $\pm$ 0.40	11.71 $\pm$ 0.23
	ROUGE-2	7.77	**12.55**	11.49	6.63 $\pm$ 0.58	10.66 $\pm$ 0.37
	ROUGE-L	11.77	**12.87**	12.03	10.84 $\pm$ 0.41	12.43 $\pm$ 0.24
	BERTScore	13.92	14.24	13.88	13.53 $\pm$ 0.32	**14.46** $\pm$ 0.20
	BLEURT	15.27	**15.59**	14.81	15.28 $\pm$ 0.19	15.39 $\pm$ 0.10

Table 9: 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	ROUGE-1	13.66	14.97	**18.70**	15.70 $\pm$ 0.80	17.75 $\pm$ 0.41
	ROUGE-2	9.74	13.54	**18.54**	14.60 $\pm$ 1.05	17.42 $\pm$ 0.52
	ROUGE-L	13.14	14.59	**19.22**	15.83 $\pm$ 0.93	18.23 $\pm$ 0.47
	BERTScore	19.44	18.19	**21.30**	18.79 $\pm$ 0.76	20.62 $\pm$ 0.39
	BLEURT	23.59	22.53	**24.52**	23.02 $\pm$ 0.43	24.02 $\pm$ 0.23
GigaWord	ROUGE-1	12.90	**16.88**	15.16	13.62 $\pm$ 0.53	15.31 $\pm$ 0.26
	ROUGE-2	5.82	**13.97**	12.08	8.06 $\pm$ 0.61	11.18 $\pm$ 0.30
	ROUGE-L	12.92	**16.56**	15.18	13.29 $\pm$ 0.57	15.33 $\pm$ 0.28
	BERTScore	19.23	19.39	18.61	18.73 $\pm$ 0.45	**19.46** $\pm$ 0.23
	BLEURT	20.20	**20.86**	19.93	20.43 $\pm$ 0.26	20.45 $\pm$ 0.14

Table 10: 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	18.90	**20.18**	18.44 ± 0.22	20.07 ± 0.12
	METEOR	21.30	20.12	**22.57**	19.71 ± 0.26	22.22 ± 0.14
	ROUGE-L	15.24	18.80	**19.80**	17.80 ± 0.23	19.45 ± 0.12
	CIDEr	13.40	14.93	**15.21**	14.76 ± 0.06	15.07 ± 0.03
	BERTScore	23.44	22.55	**23.95**	22.39 ± 0.18	23.75 ± 0.10
	BLEURT	24.31	23.91	**25.05**	24.43 ± 0.10	24.94 ± 0.06
E2ENLG	BLEU	11.53	**12.41**	11.39	8.59 ± 1.14	10.76 ± 0.49
	METEOR	**18.46**	14.37	12.93	10.55 ± 1.18	12.42 ± 0.45
	ROUGE-L	8.60	**12.36**	11.52	9.30 ± 0.96	10.94 ± 0.39
	CIDEr	10.29	**13.23**	11.21	9.30 ± 0.93	10.85 ± 0.44
	BERTScore	14.45	**15.04**	13.73	12.19 ± 0.96	13.31 ± 0.37
	BLEURT	14.61	**15.74**	15.27	14.81 ± 0.39	15.04 ± 0.16
WikiBioNLG	BLEU	3.27	**3.34**	2.38	1.99 ± 0.74	2.31 ± 0.31
	METEOR	**3.73**	3.43	2.59	2.29 ± 0.68	2.57 ± 0.31
	ROUGE-L	**4.00**	3.72	3.35	3.14 ± 0.62	3.58 ± 0.29
	CIDEr	2.10	**2.44**	2.21	1.48 ± 0.62	2.03 ± 0.25
	BERTScore	**5.09**	4.57	3.97	4.42 ± 0.44	4.33 ± 0.21
	BLEURT	5.32	**5.72**	4.23	4.70 ± 0.39	4.59 ± 0.17

Table 11: The Kendall correlations with human judgement on the data-to-text task.

Dataset	Pearson Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
WebNLG	BLEU	25.79	28.71	**31.07**	28.50 ± 0.29	30.75 ± 0.12
	METEOR	30.78	30.54	32.81	30.02 ± 0.36	**32.89** ± 0.15
	ROUGE-L	24.15	28.01	**30.12**	27.58 ± 0.28	29.73 ± 0.12
	CIDEr	23.09	24.48	**24.78**	24.28 ± 0.05	24.55 ± 0.02
	BERTScore	34.53	33.41	**35.38**	33.48 ± 0.23	35.25 ± 0.11
	BLEURT	35.97	35.30	**36.99**	36.08 ± 0.13	36.80 ± 0.07
E2ENLG	BLEU	12.78	16.08	**19.19**	13.84 ± 1.45	17.59 ± 0.70
	METEOR	**25.55**	19.98	21.59	17.09 ± 1.39	20.47 ± 0.66
	ROUGE-L	12.22	16.36	**19.49**	14.63 ± 1.29	17.93 ± 0.62
	CIDEr	13.83	17.21	**18.00**	15.75 ± 0.62	17.01 ± 0.30
	BERTScore	22.76	22.00	**22.88**	19.60 ± 1.04	21.92 ± 0.49
	BLEURT	22.75	23.89	**23.95**	22.45 ± 0.50	23.34 ± 0.25
WikiBioNLG	BLEU	**8.19**	7.04	4.98	4.77 ± 0.97	5.48 ± 0.43
	METEOR	**8.31**	7.68	5.43	5.38 ± 0.90	5.97 ± 0.40
	ROUGE-L	**9.88**	8.80	6.69	6.99 ± 0.85	7.45 ± 0.38
	CIDEr	5.35	**7.61**	5.85	5.90 ± 0.54	6.12 ± 0.24
	BERTScore	**8.98**	**8.98**	7.26	7.53 ± 0.61	7.76 ± 0.28
	BLEURT	9.21	**9.66**	8.19	8.42 ± 0.44	8.54 ± 0.20

Table 12: The Pearson correlations with human judgement on the data-to-text task.

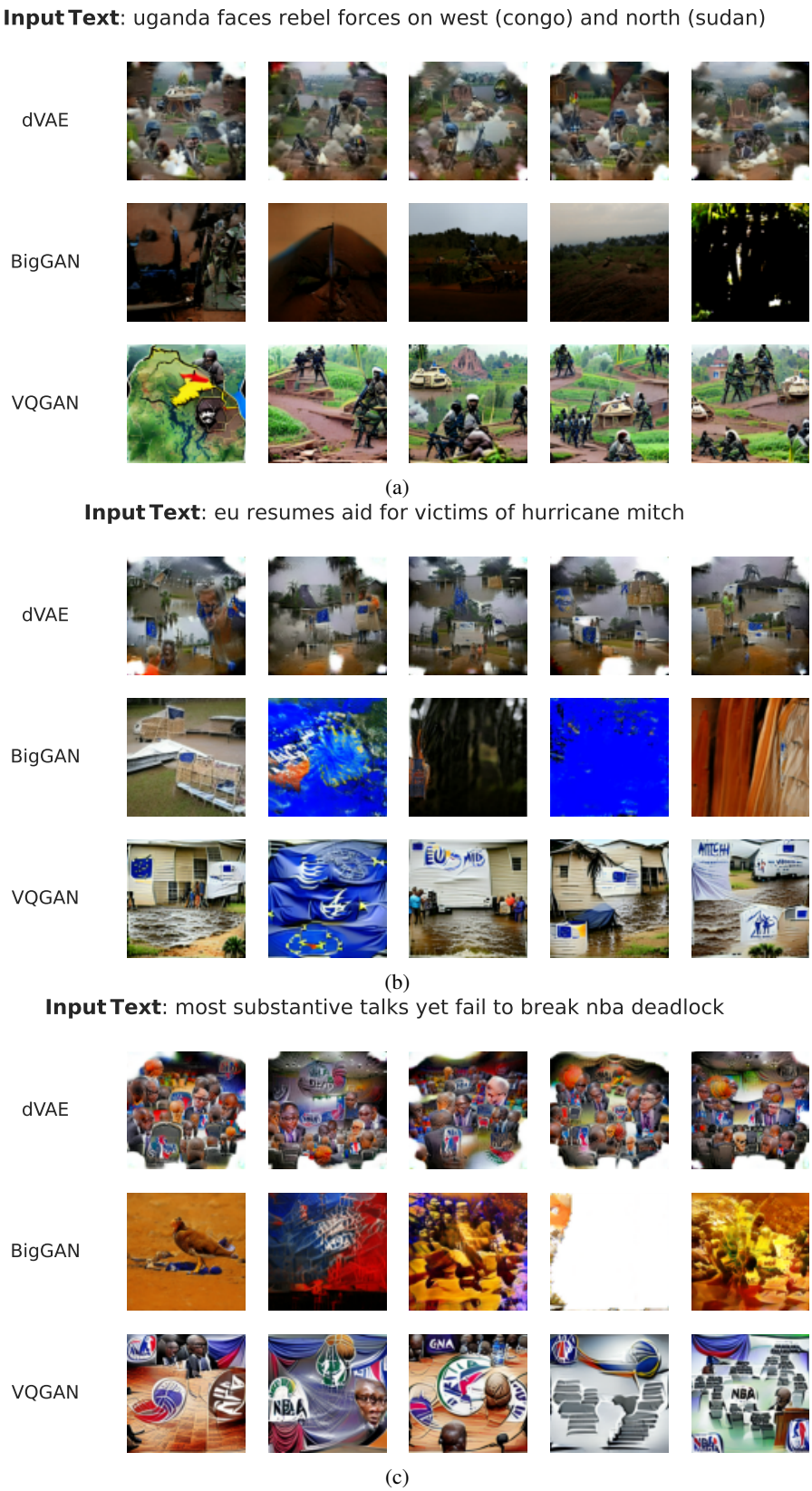


Figure 12: Groups of images generated by IMAGINE with different image genrative backbones with random initializations. The image generative backbones are dVAE, BigGAN and VQGAN.

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.

Figure 13: Three groups of images rendered with different iteration steps. Images are visualized from reference text snippet in DUC2004.

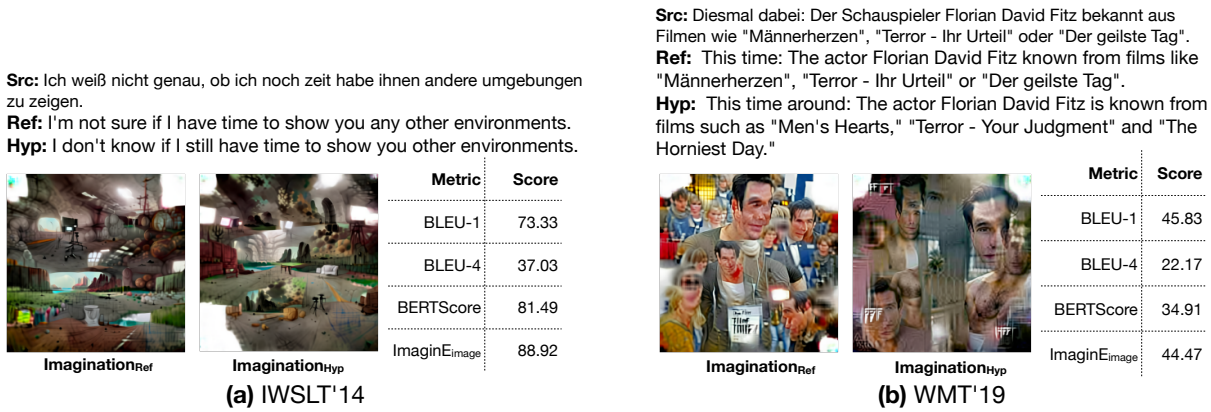


Figure 14: Case studies for machine translation. **Src**: the German text to be translated. **Ref**: the reference translation. **Hyp**: the generated translation candidate. We report the metric scores and the human score for the reported pair of (Ref, Hyp).

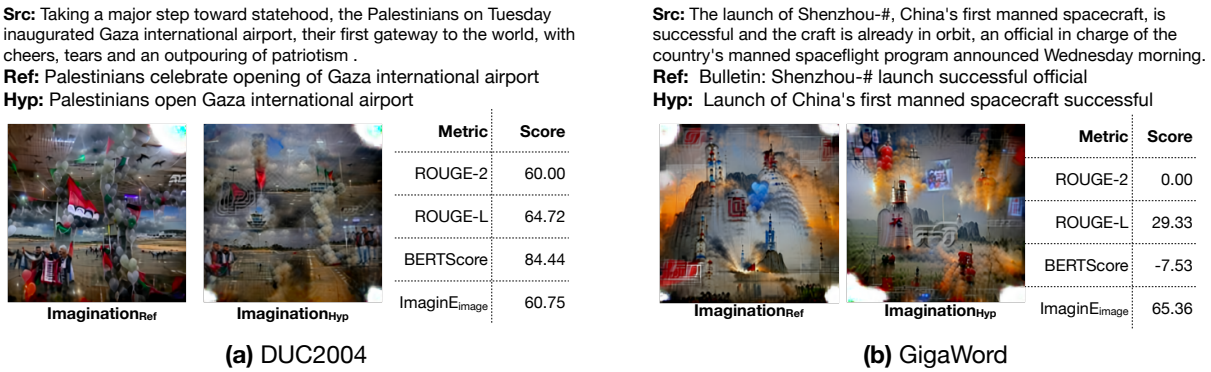


Figure 15: Case studies for abstractive text summarization. **Src**: the text to be summarized. **Ref**: the reference summary. **Hyp**: the generated summary candidate. We report the metric scores and the human score for the reported pair of (Ref, Hyp).

Ref: Julia Morgan was the architect of the grounds of Asilomar Conference.
Hyp: Julia Morgan was the architect of the Asilomar Conference Grounds.

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

Metric	Score
BLEU	65.25
METEOR	49.17
BERTScore	90.05
$\text{ImaginE}_{\text{image}}$	35.16

(a) WebNLG

Ref: Sven Leuenberger (born August 25, 1969 in Niederuzwil, Switzerland) is a retired Swiss professional ice hockey defender.
Hyp: 25 ft tall, Nieder Niederberger was a member of the club's shoots team.

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

Metric	Score
BLEU	1.92
METEOR	6.09
BERTScore	-16.43
$\text{ImaginE}_{\text{image}}$	15.29

(b) WikiBioNLG

Ref: Giraffe, in the riverside area, near the Rainbow Vegetarian Café, there is a pub with fast food, of and it is kid friendly.
Hyp: Giraffe is a dish that can be served as a dessert.

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

Metric	Score
BLEU	2.43
METEOR	6.03
BERTScore	17.79
$\text{ImaginE}_{\text{image}}$	40.17

(c) E2ENLG

Ref: There is a coffee shop Blue Spice in the riverside area.
Hyp: Blue Spice is a type of coffee shop.

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

Metric	Score
BLEU	18.00
METEOR	29.91
BERTScore	46.41
$\text{ImaginE}_{\text{image}}$	67.24

(d) E2ENLGNLG

Figure 16: 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).

A.8 Human Evaluation

Figure 17 shows an example of instructions provided to MTurk annotators. Our study is approved for IRB exempt. The estimated hourly wage paid to MTurk annotators is $12.

Instructions

A high-quality piece of text should be **fluent** , **grammatically correct**, and be **factually consistent** with the original sentence.

Please use the sliders to indicate how well each piece of generated text align with the reference text in the following three aspects.

Note: It is not necessary to align with the reference word-by-word, as long as it preserves the factual consistency.

Sample 1

- **Reference:**

  pope michael iii of alexandria (also known as khail iii) was the coptic pope of alexandria and patriarch of the see of st. mark (880 -- 907) .

- **Generated Text:**

  907 march 1607 was the date of the death of Pope michael of alexandria.

The generated text is fluent:

(1 = Stongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly Agree)

The generated text is grammartically correct:

(1 = Stongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly Agree)

The generated text is factually consistent with the original text:

(1 = Stongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly Agree)

Figure 17: The instructions for MTurk annotators to evaluate the text generated for the data-to-text generation task.