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

Wanrong Zhu[¶], Xin Eric Wang[§], An Yan[†], Miguel Eckstein[¶], William Yang Wang[¶]
[¶]UC Santa Barbara, [§]UC Santa Cruz, [†]UC Santa Diego
{wanrongzhu,william}@cs.ucsb.edu, xwang366@ucsc.edu
ayan@ucsd.edu, miguel.eckstein@psych.ucsb.edu

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 [31] and DALL-E [32], 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.

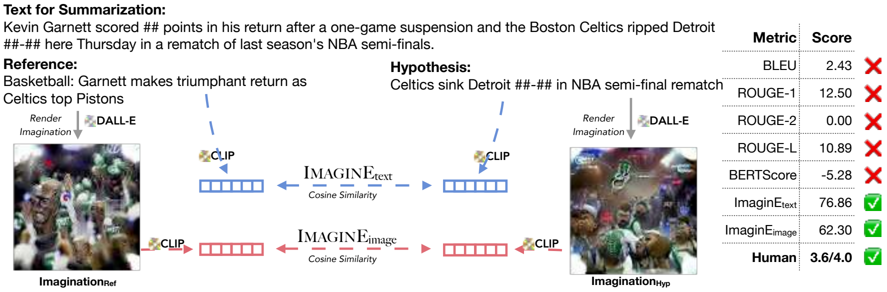
1 Introduction

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 [29], METEOR [8], CIDEr [47], have been widely adopted in the field. Edit-distance based metrics, such as CharacTER [48], WMD [18], SMD [6], have also been explored. Recently, Zhang et al. [50] proposed to leverage BERT 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 which focus on comparing 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 [13]. Cognitive studies also show that visual imagery improves comprehension during human language processing [35]. 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 [31] 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 a powerful pre-trained vision-language model DALL-E [32] 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 [31]. Figure 1 shows an example.

Preprint. Under review.



Text for Summarization:
Kevin Garnett scored ## points in his return after a one-game suspension and the Boston Celtics ripped Detroit ##-## here Thursday in a rematch of last season's NBA semi-finals.

Reference:
Basketball: Garnett makes triumphant return as Celtics top Pistons

Hypothesis:
Celtics sink Detroit ##-## in NBA semi-final rematch

Metric	Score	
BLEU	2.43	✗
ROUGE-1	12.50	✗
ROUGE-2	0.00	✗
ROUGE-L	10.89	✗
BERTScore	-5.28	✗
ImaginE$_{text}$	76.86	✓
ImaginE$_{image}$	62.30	✓
Human	**3.6/4.0**	✓

Figure 1: An evaluation example on GigaWord for abstractive text summarization. IMAGINE visualizes machine imagination with DALL-E 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

Common practices for NLG evaluation compare the generated hypothesis text with the annotated references. An evaluation metric is a function $f(\boldsymbol{x}_{ref}, \boldsymbol{x}_{hyp}) \in R$. Metric performance is conventionally evaluated by its correlation with human judgments. Existing automatic evaluation metric calculations are mainly based on the following three mechanisms: n-grams overlap, edit distance, and embedding matching.

Metrics Based on N-Gram Overlap BLEU is the geometrically weighted mean of its n-grams precision scores. While BLEU treats measure n-grams equally, NIST [25] assigns heavier weights to n-grams that appear less. ROUGE-n [22] computes n-grams overlaps, while ROUGE-L measures the longest matching sequence of words using the longest common sub-sequence. Compared to BLEU that only measures precision, ROUGE focuses on recall while METEOR [8] is based on the harmonic mean of the unigram precision and recall, in which recall is weighted higher than precision. CIDEr [47] uses a consensus-based protocal that computes cosine similarity between TF-IDF [35] weighted n-grams.

Metrics Based on Edit-Distance WER [44] calculates the edit distance between the two text snippets. TER [39] normalizes edit distance with the average length of the references. ITER [28] allows stem matching, optimizable edit costs, and better normalization. PER [43] computes position-independent word error rate. CharacTER [48] calculates translation edit rate on character level.

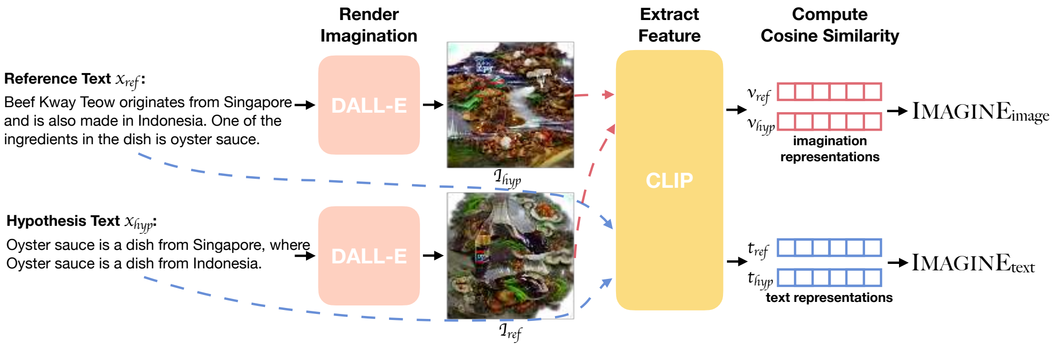

Render Imagination
Extract Feature
Compute Cosine Similarity
Reference Text x_{ref}:
Beef Kway Teow originates from Singapore and is also made in Indonesia. One of the ingredients in the dish is oyster sauce.
DALL-E
$\mathcal{I}_{hyp}$
Hypothesis Text x_{hyp}:
Oyster sauce is a dish from Singapore, where Oyster sauce is a dish from Indonesia.
DALL-E
$\mathcal{I}_{ref}$
CLIP
v_{ref}
v_{hyp}
imagination representations
IMAGINE$_{\text{image}}$
t_{ref}
t_{hyp}
text representations
IMAGINE$_{\text{text}}$

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 DALL-E. 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.

Metrics Based on Embedding WMD [17] calculates the distance of two sequences on top of EMD [34]. SMD [5] combines word and sentence embeddings when calculating similarity. MEANT 2.0 [15] and YiSi [16] calculate the phrasal similarity of the semantic role filler with the help of TF-IDF. BERTScore [50] computes pairwise cosine similarity between tokens in the hypothesis and the reference text and conducts greedy matching to maximize the similarity.

Aside from previous text-only metrics, there also appear metrics that utilize pre-trained multi-modal models and introduces visual features on top of text references for NLG evaluation. TIGEr [12] computes the text-image grounding scores with pre-trained SCAN [20]. ViLBERTScore-F [19] relies on pre-trained ViLBERT [24] to extract image-conditioned embeddings for the text. The concurrent CLIPScore [10] proposes a text-reference-free metric for image captioning by directly comparing the image features with caption embeddings with CLIP [31]. Our method differs in that we use visual picture generation as embodied imaginations and apply our metric to a variety of text-to-text generation tasks.

3 IMAGINE

In this section, we describe the computation process of IMAGINE. To evaluate the similarity of the generated text with the reference, we render machine imagination figures with DALL-E [32] for the reference and the hypothesis text. Then we extract text and imagination features with CLIP [31] and compute the visual and textual cosine similarity. Figure 2 illustrates the process.

3.1 Model Details

CLIP CLIP [31] 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 [7, 46] 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 [30, 36]. 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 [40, 45].

DALL-E DALL-E [32] is a 12-billion parameter version of GPT-3 [3] trained to generate images from text descriptions. The model is trained on a dataset of a similar scale to JFT-300M [41] by collecting 250 million text-image pairs from the internet, which incorporates Conceptual Captions [37], the text-image pairs from Wikipedia, and a filtered subset of YFCC100M [42].

DALL-E trains a discrete variational autoencoder(dVAE) [33] 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 [30, 36] 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.

3.2 IMAGINE Similarity Score

Construct Imagination For each image, we randomly initialize a latent matrix $\boldsymbol{H}$ and use DALL-E's pre-trained decoder of the 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.

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

We optimize each generation process for 1K steps, and refer to the generated image as the imagination for further computation.

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 DALL-E in previous steps; $\text{IMAGINE}_{text\&image}$ is the average of IMAGINE_{text} and IMAGINE_{image}, which takes both the text and the imagination into consideration.

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

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

$$\text{IMAGINE}_{text\&image} = \frac{1}{2}(\text{IMAGINE}_{text} + \text{IMAGINE}_{image}) \tag{4}$$

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:

$$metric\_score' = metric\_score + \text{IMAGINE}\_similarity\_score \tag{5}$$

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.

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 [27] implementation to generate English translation from German on IWSLT'14 [2] and WMT'19 [1] datasets. We choose these two to-Enligh translation tasks because currently DALL-E and CLIP only support English. For abstractive text summarization, we use the implementation of Li et al. [21] to generate sentence summarization on DUC2004[1] and use ProphetNet [49] 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 [38], E2ENLG [26] and WikiBioNLG [23]. We use the text generated by the KGPT [4] 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 [29] for $n = 1, 2, 3, 4$ and BERTScore [50]. For abstractive text summarization, we report results on ROUGE-1, ROUGE-2, ROUGE-L [22] and BERTScore. For data-to-text generation, we utilize five automatic metrics for NLG, including BLEU, ROUGE-L, METEOR [8], CIDEr [47] 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 Following the setup of Hodosh et al. [11], we invite MTurk[3] annotators to judge the quality of the generated text on a graded scale from 1 to 4, in which 1 stands for poorly generated text while 4 represents high-quality generation aligned with references. Each example is scored by 5 human judges, and we take the mean of human scores to compute correlations. In the following sections, we report Kendall correlation [14] to human scores. We also record Pearson correlation [9] in the Appendix.

5 Results

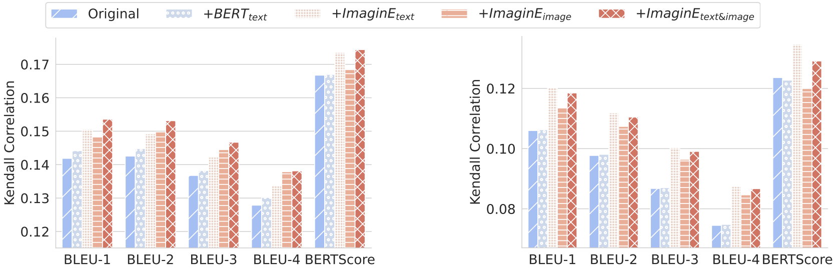
5.1 Machine Translation

Figure 3 shows the system-level Kendall 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' correlations with human scores, in which 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.

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

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

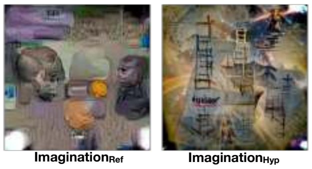
[3] https://www.mturk.com/



(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 Kendall correlation with human judgments.

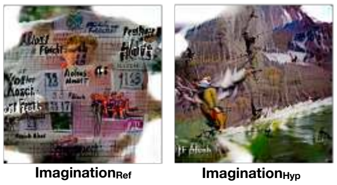
Src: Also entschied ich mich eines tages den filialleiter zu besuchen, und ich fragte den leiter, "funktioniert dieses modell, dass sie den menschen all diese möglichkeiten bieten wirklich?"
Ref: So I one day decided to pay a visit to the manager, and I asked the manager, "is this model of offering people 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?"



Metric	Score
BLEU-1	69.70
BLEU-4	20.26
BERTScore	66.62
ImaginE$_{image}$	34.81
Human	**3.2/4.0**

(a) IWSLT'14

Src: Es gab Momente in dieser noch jungen Saison, da ging Alois Fetsch mit den Seinen hart ins Gericht.
Ref: There were moments in this young season, when Alois Fetsch was really hard on his team.
Hyp: There were moments in this fledgling season when Alois Fetsch took a hard line with his own.



Metric	Score
BLEU-1	58.82
BLEU-4	30.28
BERTScore	69.54
ImaginE$_{image}$	44.95
Human	**3.6/4.0**

(b) WMT'19

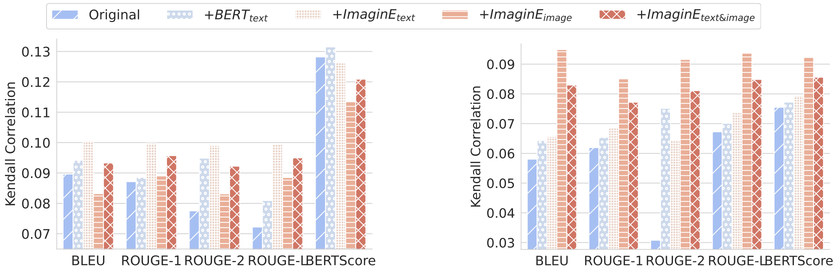
Figure 4: Case studies for machine translation. **Src**: the German text to be translated. **Ref**: the reference translation. **Hyp**: the generated translation candidate. We report the metric scores and the human score for the reported pair of (Ref, Hyp).

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), ImaginE also yields completely different imaginations in contrast with the minor difference between the reference text and the generated candidate. The translation deviates from its original text with only minor differences in this case ("be hard on his team" vs "be hard on himself"). While other metric scores are high, ImaginE gives a relatively low image similarity score in accordance with the real quality of the translation.

5.2 Abstractive Text Summarization

Figure 5 shows the system-level Kendall 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$_{text}$ and ImaginE$_{text\&image}$ can steadily improve BLEU and ROUGE-related metrics' correlations with human scores on both datasets. ImaginE$_{image}$ contributes to the biggest improvement on Gigaword, but only enhances ROUGE-1 and ROUGE-L on DUC2004. ImaginE$_{text}$ surpasses BERT$_{text}$ on all metrics. BERTScore can be improved by ImaginE on GigaWord, but not on DUC2004.

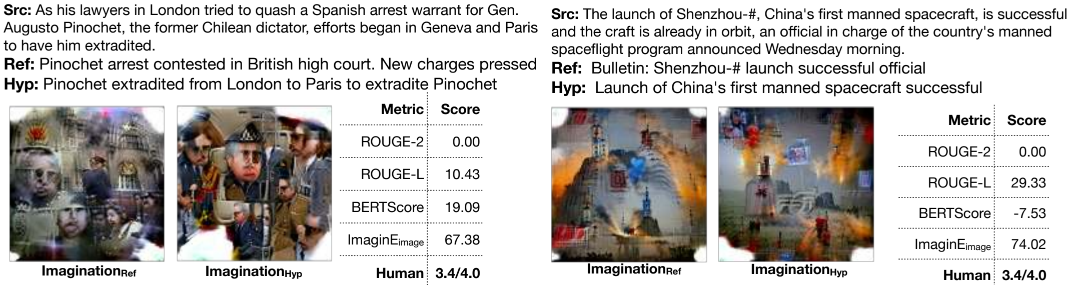
ImaginE can capture the gist of texts with similar meanings and renders reasonable descriptive imaginations that are alike, regardless of word choices. Figure 6 shows two sets of examples where the hypothesis summary scores high in human evaluation but scores low on existing automatic evaluation metrics. 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.



(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 Kendall correlation with human judgments.



Src: As his lawyers in London tried to quash a Spanish arrest warrant for Gen. Augusto Pinochet, the former Chilean dictator, efforts began in Geneva and Paris to have him extradited.
Ref: Pinochet arrest contested in British high court. New charges pressed
Hyp: Pinochet extradited from London to Paris to extradite Pinochet

Metric	Score
ROUGE-2	0.00
ROUGE-L	10.43
BERTScore	19.09
ImaginE$_{image}$	67.38
Human	3.4/4.0

Src: The launch of Shenzhou-#, China's first manned spacecraft, is successful and the craft is already in orbit, an official in charge of the country's manned spaceflight program announced Wednesday morning.
Ref: Bulletin: Shenzhou-# launch successful official
Hyp: Launch of China's first manned spacecraft successful

Metric	Score
ROUGE-2	0.00
ROUGE-L	29.33
BERTScore	-7.53
ImaginE$_{image}$	74.02
Human	3.4/4.0

(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).

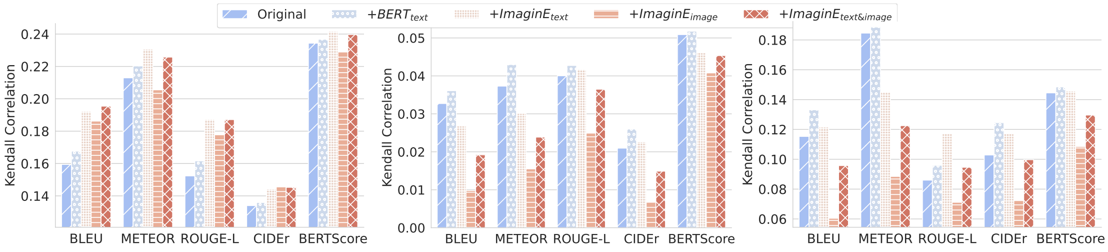
5.3 Data-to-Text Generation

Figure 7 shows the system-level Kendall 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, ROUGE-L, and CIDEr, but it does not apply to METEOR or 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 in Figure 8(a), in which changing the relative position of "grounds" shifts the main part of the imagination from a person to the dirt ground.

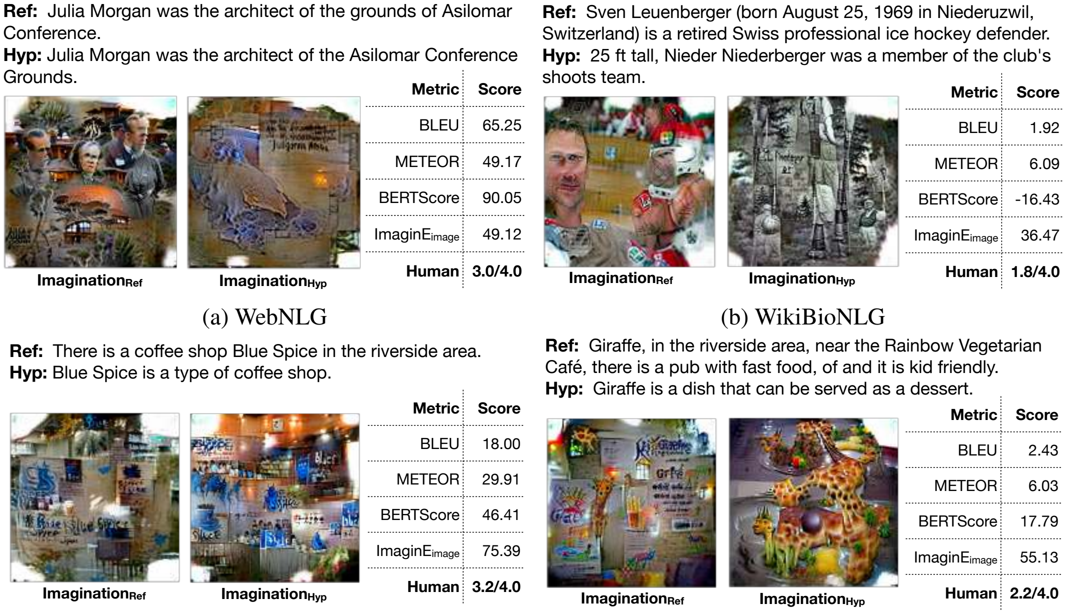
We witness a drawback in most listed metric's correlations after applying our IMAGINE approach on WikiBioNLG. This is because the WikiBioNLG dataset is built upon Wikipedia biography, and IMAGINE is not good at visualizing abstract concepts. In Figure 8(b), our IMAGINE failed to visualize the player's 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 Kendall 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, "Blue Spice" leads to the appearance of blue patches in Figure 8(c). "Giraffe" and "Rainbow" in Figure 8(d) also result in weird imagination that is unrelated to the main content of the generated text.



(a) WebNLG (b) WikiBioNLG (c) E2ENLG

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 Kendall correlation with human judgments.



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

Metric	Score
BLEU	65.25
METEOR	49.17
BERTScore	90.05
ImaginE$_{image}$	49.12
Human	**3.0/4.0**

(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.

Metric	Score
BLEU	1.92
METEOR	6.09
BERTScore	-16.43
ImaginE$_{image}$	36.47
Human	**1.8/4.0**

(b) WikiBioNLG

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

Metric	Score
BLEU	18.00
METEOR	29.91
BERTScore	46.41
ImaginE$_{image}$	75.39
Human	**3.2/4.0**

(c) E2ENLG

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.

Metric	Score
BLEU	2.43
METEOR	6.03
BERTScore	17.79
ImaginE$_{image}$	55.13
Human	**2.2/4.0**

(d) E2ENLG

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).

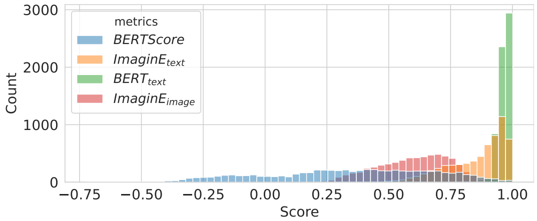


Figure 9: The score distributions histplot of IMAGINE, BERT$_{text}$ and BERTScore used in our experiments. All four metrics range between [-1, 1].

6 Discussion

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 Kendall 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

Task	Dataset	Kendall Correlation								
MT		BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$	BERTScore	BLEU-1	BLEU-2	BLEU-3	BLEU-4
	WMT19	8.713	**13.270**	7.402	9.475	12.360	10.601	9.773	8.675	7.450
	IWSLT14	13.842	12.802	10.519	11.983	**16.676**	14.188	14.255	13.677	12.787
TS		BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
	DUC2004	8.313	9.528	7.269	8.240	**12.820**	8.965	8.714	7.754	7.219
	GigaWord	8.452	8.461	**11.522**	10.619	7.549	5.804	6.193	3.078	6.725
DT		BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$	BERTScore	BLEU	METEOR	ROUGE-L	CIDEr
	WebNLG	15.408	19.520	12.866	17.001	**23.439**	15.943	21.295	15.235	13.400
	E2ENLG	10.863	10.834	3.244	7.127	14.451	11.534	**18.460**	8.603	10.289
	WikiBioNLG	3.073	2.186	0.299	1.238	**5.089**	3.273	3.731	4.003	2.097

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

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 that causes their embeddings to lie in distinct space. 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, BERT$_{text}$ has the sharpest distribution, 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]. Corner cases exist but are rare. 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. Two supportive cases would be Figure 4(a) and Figure 4(b). 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 imagination. In Figures 8(c) and 8(d), the name of the restaurants also reduce the quality of the imagination. Future work may systematically examine the robustness of CLIP and DALL-E.

Furthermore, even though we have access to a pre-trained DALL-E 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 be able to 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. We hope our work can contribute to the construction of multi-modal representations and the discussion of multi-modal studies.

References

[1] 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.

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

[3] 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.

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

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

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

[7] 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.

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

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

[10] 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.

[11] M. Hodosh, Peter Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics (extended abstract). *J. Artif. Intell. Res.*, 47:853–899, 2013.

[12] 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.

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

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

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

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

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

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

[19] 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.

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

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

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

[23] T. Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. Table-to-text generation by structure-aware seq2seq learning. *ArXiv*, abs/1711.09724, 2018.

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

[25] A. Martin and Mark A. Przybocki. The nist 1999 speaker recognition evaluation - an overview. *Digit. Signal Process.*, 10:1–18, 2000.

[26] K. Mathewson, P. S. Castro, Colin Cherry, George F. Foster, and Marc G. Bellemare. Shaping the narrative arc: An information-theoretic approach to collaborative dialogue. *ArXiv*, abs/1901.11528, 2019.

[27] 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.

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

[29] 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*, pages 311–318. Association for Computational Linguistics, 2002.

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

[31] 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.

[32] 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.

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

[34] 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)*, pages 59–66, 1998.

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

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

[37] 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.

[38] Anastasia Shimorina and Claire Gardent. Handling rare items in data-to-text generation. In *INLG*, 2018.

[39] 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.

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

[41] 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)*, pages 843–852, 2017.

[42] 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.

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

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

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

[46] 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.

[47] 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*, pages 4566–4575, 2015.

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

[49] 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.

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Guidelines
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]

A Appendix

A.1 Correlation Results

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

Task	Dataset	Pearson Correlation								
MT		BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$	BERTScore	BLEU-1	BLEU-2	BLEU-3	BLEU-4
	WMT19	6.990	**17.350**	9.931	13.378	17.015	13.743	12.496	11.306	9.099
	IWSLT14	18.417	14.052	16.126	17.615	**23.947**	21.473	20.822	19.173	17.596
TS		BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
	DUC2004	12.105	**19.980**	12.959	17.122	19.440	11.469	13.657	9.743	13.145
	GigaWord	12.271	11.360	**17.479**	15.585	11.977	2.722	7.239	1.290	7.824
DT		BERT$_{text}$	IE$_{text}$	IE$_{image}$	IE$_{text\&image}$	BERTScore	BLEU	METEOR	ROUGE-L	CIDEr
	WebNLG	22.385	26.784	19.625	24.963	**34.534**	25.792	30.777	24.152	23.091
	E2ENLG	13.113	18.409	7.526	13.181	22.759	12.776	**25.551**	12.218	13.826
	WikiBioNLG	6.066	4.217	1.879	3.332	8.982	8.189	8.306	**9.878**	5.348

Table 3: The Pearson 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	10.601	10.630	**12.028**	11.352	11.851
	BLEU-2	9.773	9.810	**11.187**	10.737	11.048
	BLEU-3	8.675	8.703	**10.035**	9.644	9.908
	BLEU-4	7.450	7.481	**8.743**	8.464	8.670
	BERTScore	12.360	12.274	**13.468**	11.995	12.912
IWSLT14	BLEU-1	14.188	14.417	15.056	14.831	**15.359**
	BLEU-2	14.255	14.479	14.922	14.978	**15.315**
	BLEU-3	13.677	13.824	14.225	14.450	**14.670**
	BLEU-4	12.787	13.019	13.365	13.789	**13.814**
	BERTScore	16.676	16.698	17.356	16.841	**17.441**

Table 4: 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	13.743	13.776	**15.997**	14.815	15.706
	BLEU-2	12.496	12.570	**14.671**	13.993	14.580
	BLEU-3	11.306	11.417	13.432	13.136	**13.490**
	BLEU-4	9.099	9.261	11.309	11.376	**11.511**
	BERTScore	17.015	16.900	**18.684**	16.696	18.040
IWSLT14	BLEU-1	21.473	21.772	21.989	22.906	**23.183**
	BLEU-2	20.822	21.100	21.540	22.675	**22.683**
	BLEU-3	19.173	19.496	20.199	**21.633**	21.404
	BLEU-4	17.596	17.957	18.872	**20.497**	20.122
	BERTScore	23.947	24.023	24.218	25.048	**25.215**

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

A.2 Case Study

We provide more case studies for the three NLG tasks used in our study in Figures 10 to 16. For each dataset in each task, we list 4 groups of examples together with the automatic evaluation scores and human judgments.

Dataset	Kendall Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
DUC2004	ROUGE-1	8.714	8.856	**9.982**	8.902	9.572
	ROUGE-2	7.754	9.487	**9.907**	8.313	9.224
	ROUGE-L	7.219	8.093	**9.961**	8.851	9.498
	BERTScore	12.820	**13.150**	12.630	11.344	12.087
GigaWord	ROUGE-1	6.193	6.536	6.852	**8.505**	7.719
	ROUGE-2	3.078	7.513	6.455	**9.157**	8.110
	ROUGE-L	6.725	7.009	7.372	**9.367**	8.486
	BERTScore	7.549	7.721	7.941	**9.224**	8.560

Table 6: 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	13.657	14.054	**17.273**	15.252	16.396
	ROUGE-2	9.743	10.709	**16.371**	13.798	15.314
	ROUGE-L	13.145	13.650	**17.729**	15.289	16.698
	BERTScore	19.440	19.502	**21.023**	18.664	19.994
GigaWord	ROUGE-1	7.239	8.024	9.249	**12.245**	10.892
	ROUGE-2	1.290	2.579	5.425	**9.386**	7.579
	ROUGE-L	7.824	8.575	9.797	**12.997**	11.570
	BERTScore	11.977	12.239	12.416	**14.757**	13.701

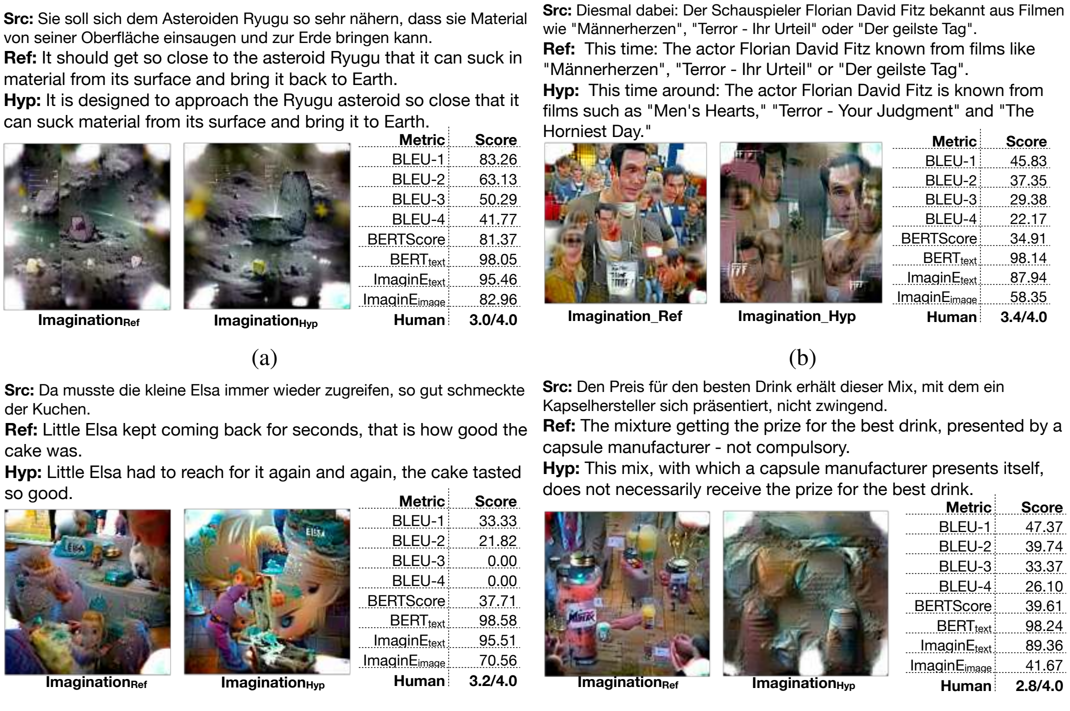
Table 7: The Pearson correlations with human judgement on the abstractive text summarization task.

Dataset	Kendall Correlation					
	Metrics	Original	+BERT$_{text}$	+IMAGINE$_{text}$	+IMAGINE$_{image}$	+IMAGINE$_{text\&image}$
WebNLG	BLEU	15.943	16.764	19.219	18.622	**19.546**
	METEOR	21.295	22.031	**23.066**	20.553	22.577
	ROUGE-L	15.235	16.162	**18.728**	17.784	18.716
	CIDEr	13.400	13.591	14.422	**14.554**	14.528
	BERTScore	23.439	23.679	**24.168**	22.937	23.970
E2ENLG	BLEU	11.534	**13.319**	12.162	6.047	9.579
	METEOR	18.460	**18.864**	14.506	8.853	12.256
	ROUGE-L	8.603	9.595	**11.741**	7.117	9.463
	CIDEr	10.289	**12.449**	11.731	7.229	9.974
	BERTScore	14.451	**14.853**	14.575	10.855	12.966
WikiBioNLG	BLEU	3.273	**3.611**	2.682	0.981	1.922
	METEOR	3.731	**4.301**	3.013	1.551	2.388
	ROUGE-L	4.003	**4.275**	4.174	2.493	3.650
	CIDEr	2.097	**2.600**	2.264	0.692	1.491
	BERTScore	5.089	**5.177**	4.609	4.079	4.536

Table 8: 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	25.792	26.795	**30.029**	28.659	29.909
	METEOR	30.777	31.878	**33.491**	30.917	33.180
	ROUGE-L	24.152	25.229	**28.691**	27.592	28.651
	CIDEr	23.091	23.246	23.975	**24.061**	24.028
	BERTScore	34.534	34.839	**35.804**	34.124	35.447
E2ENLG	BLEU	12.776	14.659	**19.259**	11.114	15.535
	METEOR	25.551	**25.933**	23.521	15.266	19.891
	ROUGE-L	12.218	13.479	**18.820**	12.212	15.835
	CIDEr	13.826	14.266	**16.505**	14.373	15.517
	BERTScore	22.759	23.145	**23.810**	18.297	21.355
WikiBioNLG	BLEU	8.189	**9.254**	5.742	3.528	5.035
	METEOR	8.306	**9.349**	6.399	4.348	5.746
	ROUGE-L	9.878	**10.510**	8.214	6.273	7.680
	CIDEr	5.348	5.776	**5.947**	5.200	5.683
	BERTScore	8.982	**9.239**	8.253	6.987	7.814

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



Src: Sie soll sich dem Asteroiden Ryugu so sehr nähern, dass sie Material von seiner Oberfläche einsaugen und zur Erde bringen kann.
Ref: It should get so close to the asteroid Ryugu that it can suck in material from its surface and bring it back to Earth.
Hyp: It is designed to approach the Ryugu asteroid so close that it can suck material from its surface and bring it to Earth.

Metric	Score
BLEU-1	83.26
BLEU-2	63.13
BLEU-3	50.29
BLEU-4	41.77
BERTScore	81.37
BERT$_{text}$	98.05
ImaginE$_{text}$	95.46
ImaginE$_{image}$	82.96
Human	**3.0/4.0**

(a)

Src: Diesmal dabei: Der Schauspieler Florian David Fitz bekannt aus Filmen wie "Männerherzen", "Terror - Ihr Urteil" oder "Der geilste Tag".
Ref: This time: The actor Florian David Fitz known from films like "Männerherzen", "Terror - Ihr Urteil" or "Der geilste Tag".
Hyp: This time around: The actor Florian David Fitz is known from films such as "Men's Hearts," "Terror - Your Judgment" and "The Horniest Day."

Metric	Score
BLEU-1	45.83
BLEU-2	37.35
BLEU-3	29.38
BLEU-4	22.17
BERTScore	34.91
BERT$_{text}$	98.14
ImaginE$_{text}$	87.94
ImaginE$_{image}$	58.35
Human	**3.4/4.0**

(b)

Src: Da musste die kleine Elsa immer wieder zugreifen, so gut schmeckte der Kuchen.
Ref: Little Elsa kept coming back for seconds, that is how good the cake was.
Hyp: Little Elsa had to reach for it again and again, the cake tasted so good.

Metric	Score
BLEU-1	33.33
BLEU-2	21.82
BLEU-3	0.00
BLEU-4	0.00
BERTScore	37.71
BERT$_{text}$	98.58
ImaginE$_{text}$	95.51
ImaginE$_{image}$	70.56
Human	**3.2/4.0**

(c)

Src: Den Preis für den besten Drink erhält dieser Mix, mit dem ein Kapselhersteller sich präsentiert, nicht zwingend.
Ref: The mixture getting the prize for the best drink, presented by a capsule manufacturer - not compulsory.
Hyp: This mix, with which a capsule manufacturer presents itself, does not necessarily receive the prize for the best drink.

Metric	Score
BLEU-1	47.37
BLEU-2	39.74
BLEU-3	33.37
BLEU-4	26.10
BERTScore	39.61
BERT$_{text}$	98.24
ImaginE$_{text}$	89.36
ImaginE$_{image}$	41.67
Human	**2.8/4.0**

(d)

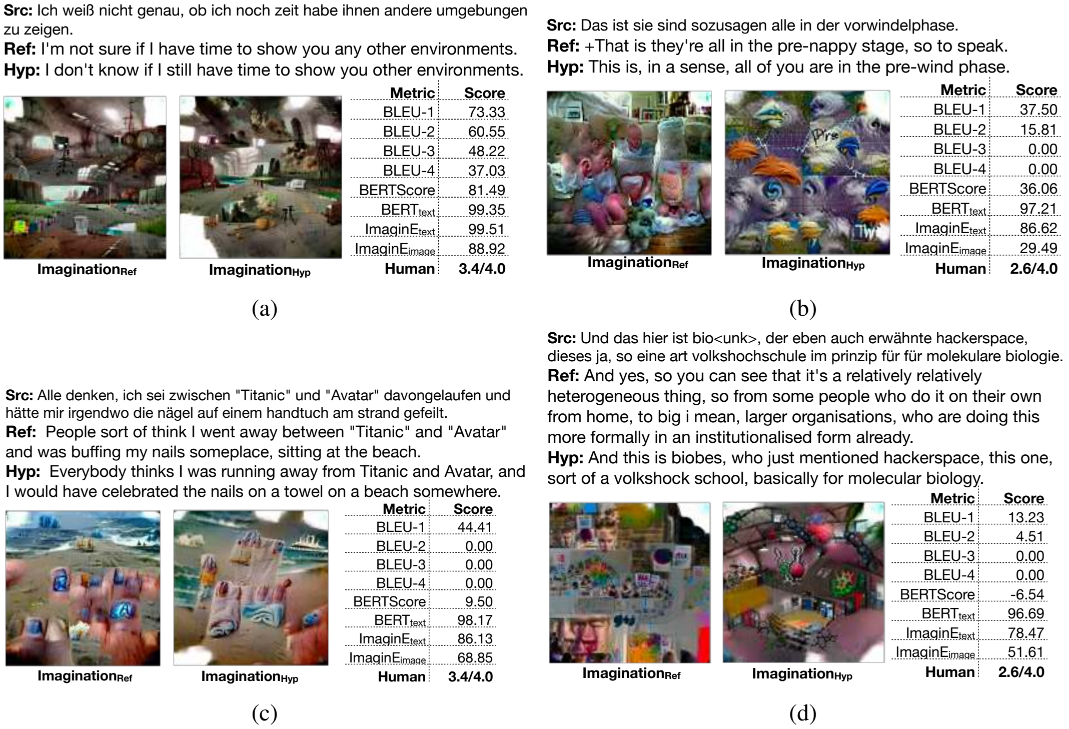
Figure 10: More examples for the machine translation task on WMT'19. **Src**: the German text to be translated. **Ref**: the reference translation. **Hyp**: the generated translation candidate.

A.3 Human Evaluation

Figure 17 shows an example of instructions provided to MTurk annotators. Our study on vision and language annotation is approved for IRB exempt. The estimated hourly wage paid to participants is $9.

A.4 Computation Expenses

We conduct experiments on 8 Titan RTX GPUs. It takes ~ 150 hours to generate all the imagination figures used in our study.



Src: Ich weiß nicht genau, ob ich noch zeit habe ihnen andere umgebungen zu zeigen.
Ref: I'm not sure if I have time to show you any other environments.
Hyp: I don't know if I still have time to show you other environments.

Metric	Score
BLEU-1	73.33
BLEU-2	60.55
BLEU-3	48.22
BLEU-4	37.03
BERTScore	81.49
$\text{BERT}_{\text{text}}$	99.35
$\text{ImaginE}_{\text{text}}$	99.51
$\text{ImaginE}_{\text{image}}$	88.92
Human	**3.4/4.0**

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

(a)

Src: Das ist sie sind sozusagen alle in der vorwindelphase.
Ref: +That is they're all in the pre-nappy stage, so to speak.
Hyp: This is, in a sense, all of you are in the pre-wind phase.

Metric	Score
BLEU-1	37.50
BLEU-2	15.81
BLEU-3	0.00
BLEU-4	0.00
BERTScore	36.06
$\text{BERT}_{\text{text}}$	97.21
$\text{ImaginE}_{\text{text}}$	86.62
$\text{ImaginE}_{\text{image}}$	29.49
Human	**2.6/4.0**

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

(b)

Src: Alle denken, ich sei zwischen "Titanic" und "Avatar" davongelaufen und hätte mir irgendwo die nägel auf einem handtuch am strand gefeilt.
Ref: People sort of think I went away between "Titanic" and "Avatar" and was buffing my nails someplace, sitting at the beach.
Hyp: Everybody thinks I was running away from Titanic and Avatar, and I would have celebrated the nails on a towel on a beach somewhere.

Metric	Score
BLEU-1	44.41
BLEU-2	0.00
BLEU-3	0.00
BLEU-4	0.00
BERTScore	9.50
$\text{BERT}_{\text{text}}$	98.17
$\text{ImaginE}_{\text{text}}$	86.13
$\text{ImaginE}_{\text{image}}$	68.85
Human	**3.4/4.0**

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

(c)

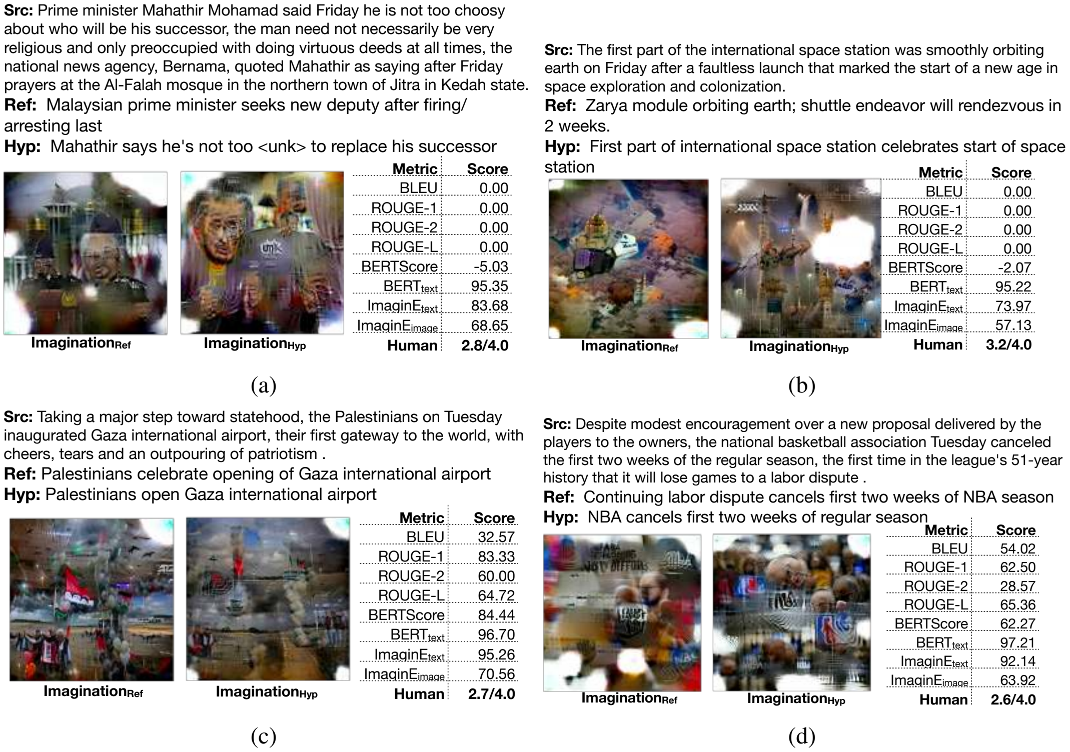
Src: Und das hier ist bio<unk>, der eben auch erwähnte hackerspace, dieses ja, so eine art volkshochschule im prinzip für für molekulare biologie.
Ref: And yes, so you can see that it's a relatively relatively heterogeneous thing, so from some people who do it on their own from home, to big i mean, larger organisations, who are doing this more formally in an institutionalised form already.
Hyp: And this is biobes, who just mentioned hackerspace, this one, sort of a volkshock school, basically for molecular biology.

Metric	Score
BLEU-1	13.23
BLEU-2	4.51
BLEU-3	0.00
BLEU-4	0.00
BERTScore	-6.54
$\text{BERT}_{\text{text}}$	96.69
$\text{ImaginE}_{\text{text}}$	78.47
$\text{ImaginE}_{\text{image}}$	51.61
Human	**2.6/4.0**

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

(d)

Figure 11: 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.



Src: Prime minister Mahathir Mohamad said Friday he is not too choosy about who will be his successor, the man need not necessarily be very religious and only preoccupied with doing virtuous deeds at all times, the national news agency, Bernama, quoted Mahathir as saying after Friday prayers at the Al-Falah mosque in the northern town of Jitra in Kedah state.
Ref: Malaysian prime minister seeks new deputy after firing/ arresting last
Hyp: Mahathir says he's not too <unk> to replace his successor

Metric	Score
BLEU	0.00
ROUGE-1	0.00
ROUGE-2	0.00
ROUGE-L	0.00
BERTScore	-5.03
$\text{BERT}_{\text{text}}$	95.35
$\text{ImaginE}_{\text{text}}$	83.68
$\text{ImaginE}_{\text{image}}$	68.65
Human	**2.8/4.0**

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

(a)

Src: The first part of the international space station was smoothly orbiting earth on Friday after a faultless launch that marked the start of a new age in space exploration and colonization.
Ref: Zarya module orbiting earth; shuttle endeavor will rendezvous in 2 weeks.
Hyp: First part of international space station celebrates start of space station

Metric	Score
BLEU	0.00
ROUGE-1	0.00
ROUGE-2	0.00
ROUGE-L	0.00
BERTScore	-2.07
$\text{BERT}_{\text{text}}$	95.22
$\text{ImaginE}_{\text{text}}$	73.97
$\text{ImaginE}_{\text{image}}$	57.13
Human	**3.2/4.0**

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

(b)

Src: Taking a major step toward statehood, the Palestinians on Tuesday inaugurated Gaza international airport, their first gateway to the world, with cheers, tears and an outpouring of patriotism .
Ref: Palestinians celebrate opening of Gaza international airport
Hyp: Palestinians open Gaza international airport

Metric	Score
BLEU	32.57
ROUGE-1	83.33
ROUGE-2	60.00
ROUGE-L	64.72
BERTScore	84.44
$\text{BERT}_{\text{text}}$	96.70
$\text{ImaginE}_{\text{text}}$	95.26
$\text{ImaginE}_{\text{image}}$	70.56
Human	**2.7/4.0**

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

(c)

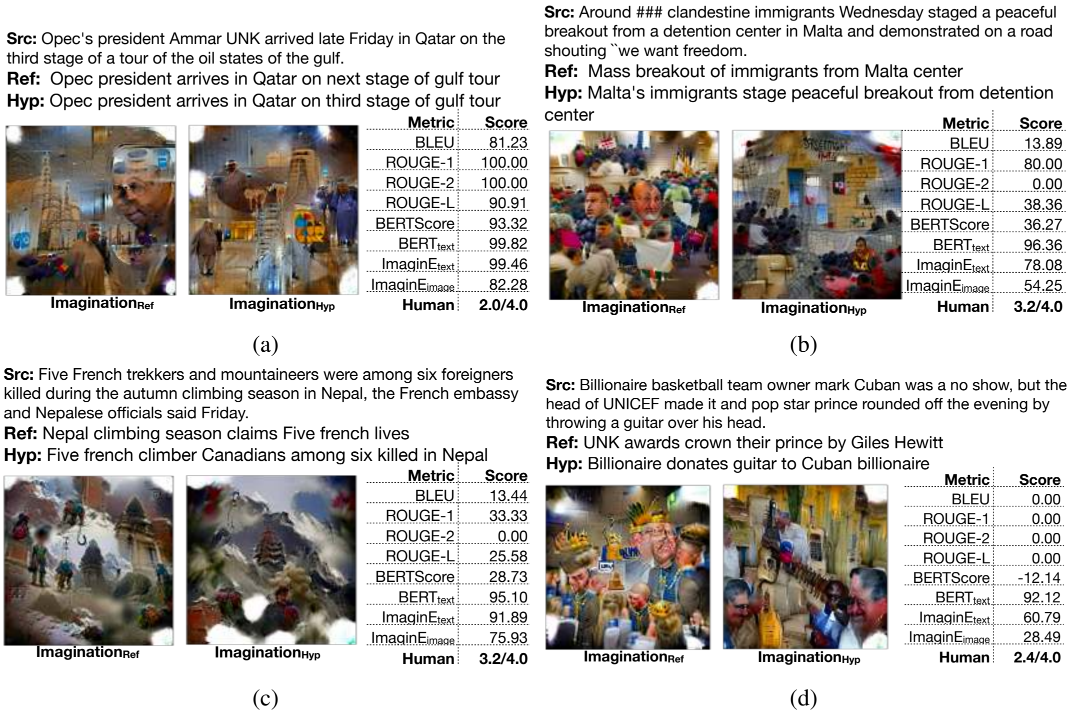
Src: Despite modest encouragement over a new proposal delivered by the players to the owners, the national basketball association Tuesday canceled the first two weeks of the regular season, the first time in the league's 51-year history that it will lose games to a labor dispute .
Ref: Continuing labor dispute cancels first two weeks of NBA season
Hyp: NBA cancels first two weeks of regular season

Metric	Score
BLEU	54.02
ROUGE-1	62.50
ROUGE-2	28.57
ROUGE-L	65.36
BERTScore	62.27
$\text{BERT}_{\text{text}}$	97.21
$\text{ImaginE}_{\text{text}}$	92.14
$\text{ImaginE}_{\text{image}}$	63.92
Human	**2.6/4.0**

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

(d)

Figure 12: 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.



(a)

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

Metric	Score
BLEU	81.23
ROUGE-1	100.00
ROUGE-2	100.00
ROUGE-L	90.91
BERTScore	93.32
BERT$_{text}$	99.82
ImaginE$_{text}$	99.46
ImaginE$_{image}$	82.28
Human	**2.0/4.0**

(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

Metric	Score
BLEU	13.89
ROUGE-1	80.00
ROUGE-2	0.00
ROUGE-L	38.36
BERTScore	36.27
BERT$_{text}$	96.36
ImaginE$_{text}$	78.08
ImaginE$_{image}$	54.25
Human	**3.2/4.0**

(c)

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

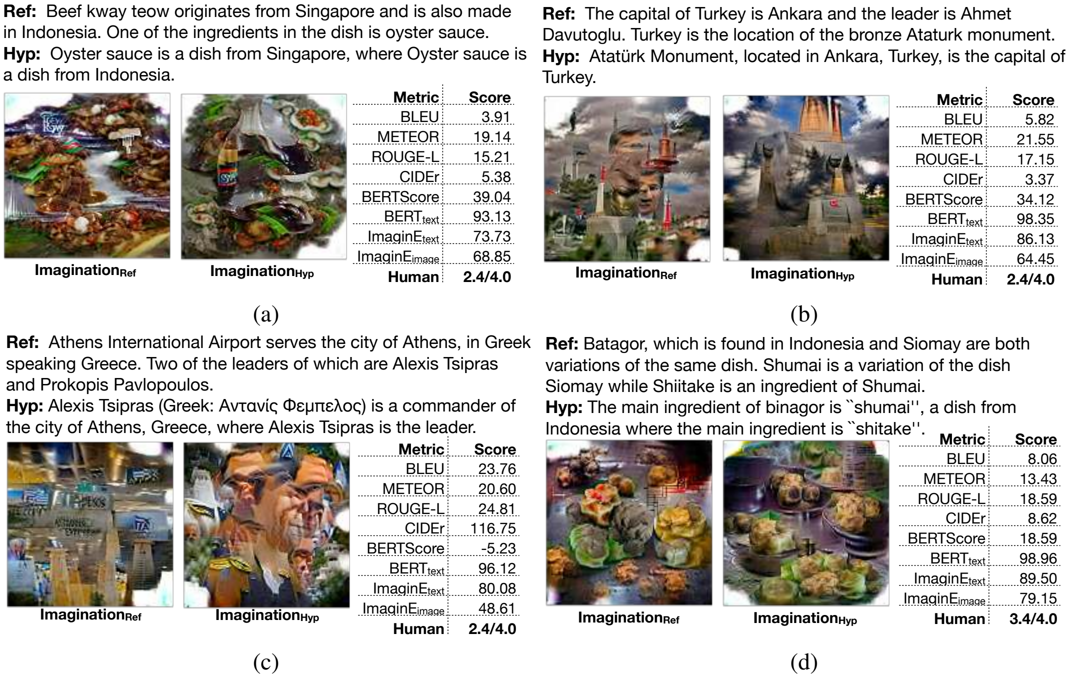
Metric	Score
BLEU	13.44
ROUGE-1	33.33
ROUGE-2	0.00
ROUGE-L	25.58
BERTScore	28.73
BERT$_{text}$	95.10
ImaginE$_{text}$	91.89
ImaginE$_{image}$	75.93
Human	**3.2/4.0**

(d)

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

Metric	Score
BLEU	0.00
ROUGE-1	0.00
ROUGE-2	0.00
ROUGE-L	0.00
BERTScore	-12.14
BERT$_{text}$	92.12
ImaginE$_{text}$	60.79
ImaginE$_{image}$	28.49
Human	**2.4/4.0**

Figure 13: 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.



(a)

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.

Metric	Score
BLEU	3.91
METEOR	19.14
ROUGE-L	15.21
CIDEr	5.38
BERTScore	39.04
BERT$_{text}$	93.13
ImaginE$_{text}$	73.73
ImaginE$_{image}$	68.85
Human	**2.4/4.0**

(b)

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.

Metric	Score
BLEU	5.82
METEOR	21.55
ROUGE-L	17.15
CIDEr	3.37
BERTScore	34.12
BERT$_{text}$	98.35
ImaginE$_{text}$	86.13
ImaginE$_{image}$	64.45
Human	**2.4/4.0**

(c)

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.

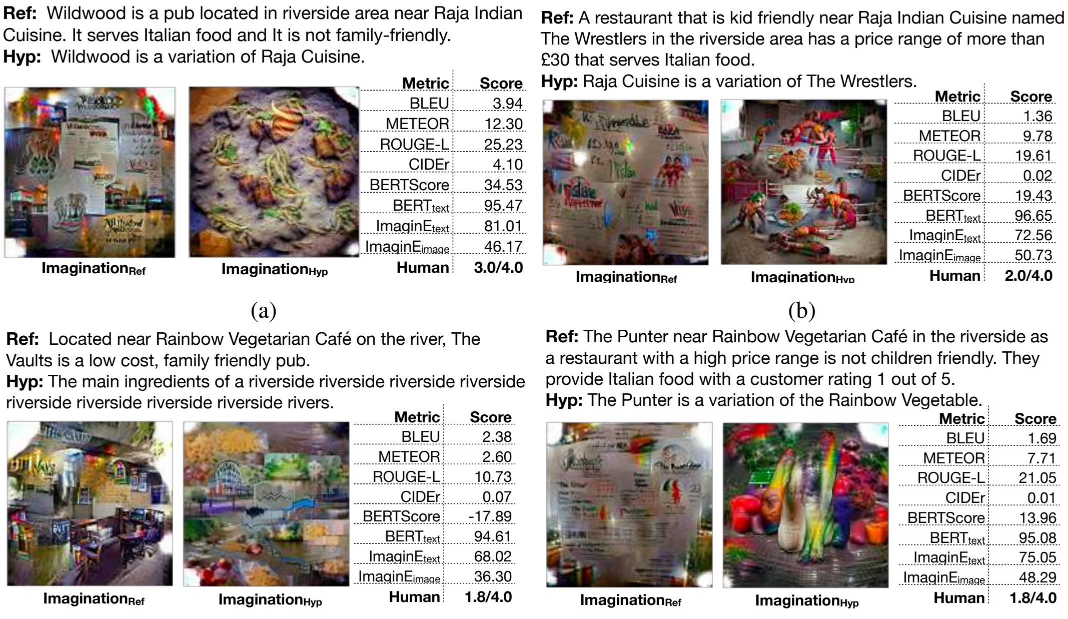
Metric	Score
BLEU	23.76
METEOR	20.60
ROUGE-L	24.81
CIDEr	116.75
BERTScore	-5.23
BERT$_{text}$	96.12
ImaginE$_{text}$	80.08
ImaginE$_{image}$	48.61
Human	**2.4/4.0**

(d)

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''.

Metric	Score
BLEU	8.06
METEOR	13.43
ROUGE-L	18.59
CIDEr	8.62
BERTScore	18.59
BERT$_{text}$	98.96
ImaginE$_{text}$	89.50
ImaginE$_{image}$	79.15
Human	**3.4/4.0**

Figure 14: More examples for the data-to-text task on WebNLG. **Ref**: the reference text. **Hyp**: the generated text candidate.



(a)

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.

Metric	Score
BLEU	3.94
METEOR	12.30
ROUGE-L	25.23
CIDEr	4.10
BERTScore	34.53
$\text{BERT}_{\text{text}}$	95.47
$\text{ImaginE}_{\text{text}}$	81.01
$\text{ImaginE}_{\text{image}}$	46.17
Human	**3.0/4.0**

(b)

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.

Metric	Score
BLEU	1.36
METEOR	9.78
ROUGE-L	19.61
CIDEr	0.02
BERTScore	19.43
$\text{BERT}_{\text{text}}$	96.65
$\text{ImaginE}_{\text{text}}$	72.56
$\text{ImaginE}_{\text{image}}$	50.73
Human	**2.0/4.0**

(c)

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 riverside rivers.

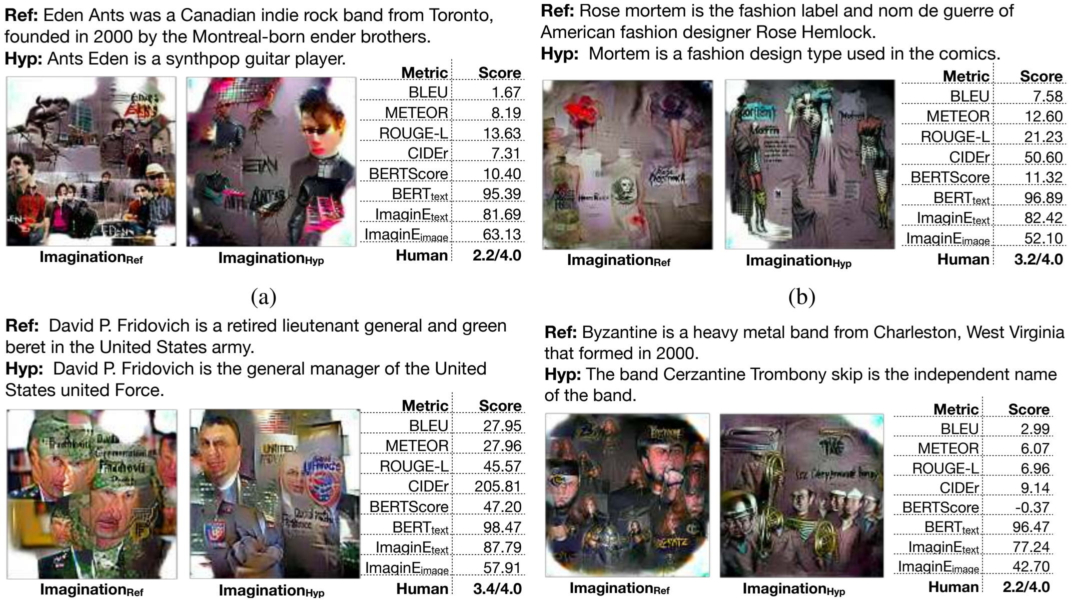
Metric	Score
BLEU	2.38
METEOR	2.60
ROUGE-L	10.73
CIDEr	0.07
BERTScore	-17.89
$\text{BERT}_{\text{text}}$	94.61
$\text{ImaginE}_{\text{text}}$	68.02
$\text{ImaginE}_{\text{image}}$	36.30
Human	**1.8/4.0**

(d)

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.

Metric	Score
BLEU	1.69
METEOR	7.71
ROUGE-L	21.05
CIDEr	0.01
BERTScore	13.96
$\text{BERT}_{\text{text}}$	95.08
$\text{ImaginE}_{\text{text}}$	75.05
$\text{ImaginE}_{\text{image}}$	48.29
Human	**1.8/4.0**

Figure 15: More examples for the data-to-text task on E2ENLG. **Ref**: the reference text. **Hyp**: the generated text candidate.



(a)

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.

Metric	Score
BLEU	1.67
METEOR	8.19
ROUGE-L	13.63
CIDEr	7.31
BERTScore	10.40
$\text{BERT}_{\text{text}}$	95.39
$\text{ImaginE}_{\text{text}}$	81.69
$\text{ImaginE}_{\text{image}}$	63.13
Human	**2.2/4.0**

(b)

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.

Metric	Score
BLEU	7.58
METEOR	12.60
ROUGE-L	21.23
CIDEr	50.60
BERTScore	11.32
$\text{BERT}_{\text{text}}$	96.89
$\text{ImaginE}_{\text{text}}$	82.42
$\text{ImaginE}_{\text{image}}$	52.10
Human	**3.2/4.0**

(c)

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.

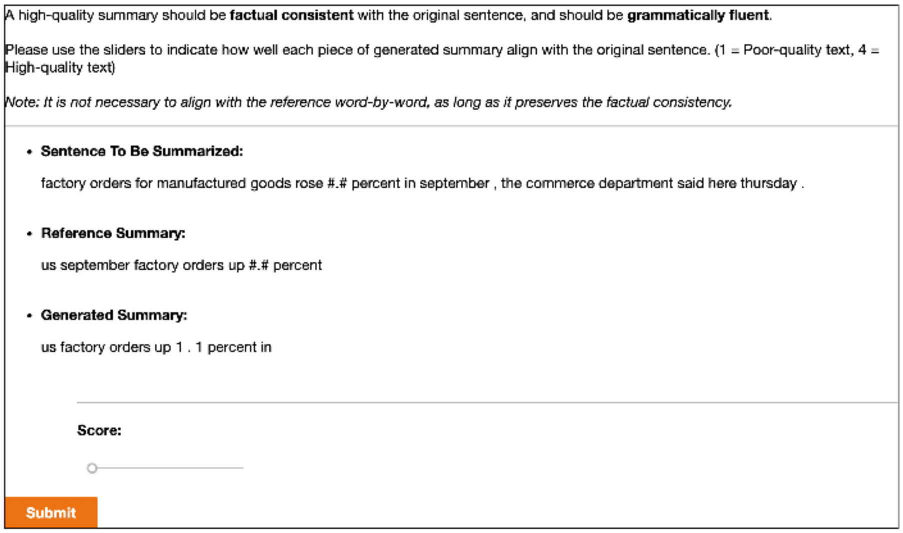
Metric	Score
BLEU	27.95
METEOR	27.96
ROUGE-L	45.57
CIDEr	205.81
BERTScore	47.20
$\text{BERT}_{\text{text}}$	98.47
$\text{ImaginE}_{\text{text}}$	87.79
$\text{ImaginE}_{\text{image}}$	57.91
Human	**3.4/4.0**

(d)

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.

Metric	Score
BLEU	2.99
METEOR	6.07
ROUGE-L	6.96
CIDEr	9.14
BERTScore	-0.37
$\text{BERT}_{\text{text}}$	96.47
$\text{ImaginE}_{\text{text}}$	77.24
$\text{ImaginE}_{\text{image}}$	42.70
Human	**2.2/4.0**

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



A high-quality summary should be **factual consistent** with the original sentence, and should be **grammatically fluent**.

Please use the sliders to indicate how well each piece of generated summary align with the original sentence. (1 = Poor-quality text, 4 = High-quality text)

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

- **Sentence To Be Summarized:**

  factory orders for manufactured goods rose #.# percent in september , the commerce department said here thursday .

- **Reference Summary:**

  us september factory orders up #.# percent

- **Generated Summary:**

  us factory orders up 1 . 1 percent in

Score:

Figure 17: The instructions for MTurk annotators to evaluate the summary generated for the abstractive text summarization task.