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 imagination often improves comprehension. In this work, we propose IMAGINE, an imagination-based automatic evaluation metric for natural language generation. With the help of StableDiffusion (Rombach et al., 2022), a state-of-the-art text-to-image generator, 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 machine-generated images 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 both reference-based and reference-free evaluation scenarios.

1 Introduction

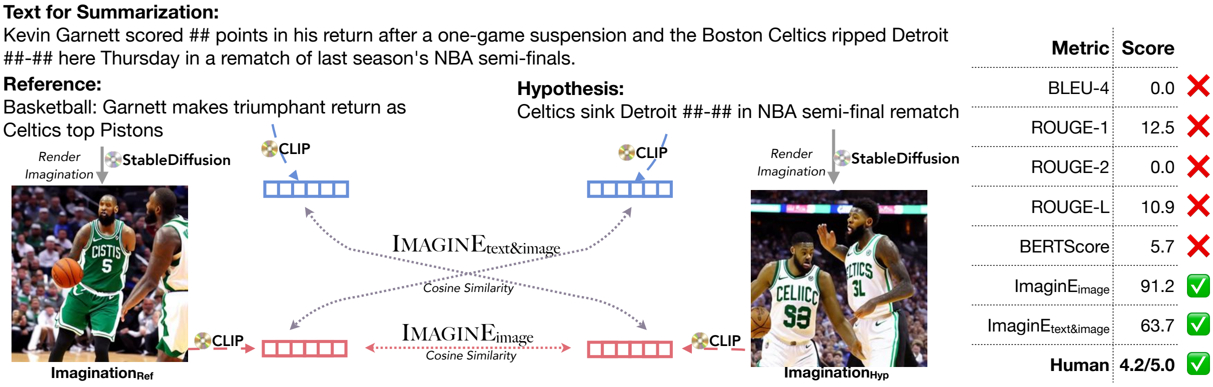
A major challenge for natural language generation (NLG) is to design an automatic evaluation metric that can align well with human judgments. To this end, many approaches have been investigated. Metrics that base on matching mechanisms such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), 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., 2015), SMD (Clark et al., 2019), have also been explored. Recently, BERTScore (Zhang* et al., 2020) and BLEURT (Sellam et al., 2020) attempt to leverage BERT (Devlin et al., 2019) to compare text embedding similarities, 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 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 use 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, we first use the state-of-the-art text-to-image generator StableDiffusion (Rombach et al., 2022) to visualize machine imagination from sentences, which is to generate descriptive images for the candidate text and the references. Then we receive the IMAGINE scores by computing two sets of similarity scores with the pre-trained CLIP model (Radford et al., 2021): the visual similarity of the generated images, and the cross-modal similarity between the text and the generated image. Figure 1 shows an example.

To understand the role the machine-generated images play in NLG evaluation, we conduct a series of experiments with IMAGINE on multiple NLG tasks and datasets, including machine translation, text summarization, and sentence completion for

Findings of the Association for Computational Linguistics: EACL 2023, pages 93–105
May 2-6, 2023 ©2023 Association for Computational Linguistics



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-4	0.0	✗
ROUGE-1	12.5	✗
ROUGE-2	0.0	✗
ROUGE-L	10.9	✗
BERTScore	5.7	✗
ImaginE$_{image}$	91.2	✓
ImaginE$_{text\&image}$	63.7	✓
Human	**4.2/5.0**	✓

Figure 1: An evaluation example on GigaWord for text summarization. IMAGINE visualizes machine imagination with StableDiffusion (Rombach et al., 2022) and extracts textual and visual representations with CLIP (Radford et al., 2021). While traditional evaluation metrics for natural language generation rely on n-grams matching or textual embeddings comparison, IMAGINE incorporates machine-generated images into the evaluation process and enhances the understanding of the text snippet as a whole through the integration of multi-modal information.

open-ended 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 do machine-generated images help?*
3. *What are the potentials and limitations of introducing machine-generated images with* IMAGINE *to NLG evaluation?*

Experimental results show that IMAGINE can serve as a complementary evaluation metric to text-based ones, and adding IMAGINE scores to existing metrics surprisingly improves most of the popular metrics' correlations with human performance on various text generation tasks. This holds for both reference-based evaluation and reference-free evaluation. 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. BLEU (Papineni et al., 2002), ROUGE-n (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam et al., 2015) are a few widely used n-gram based metrics 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 calculate the edit distance between the two text snippets with different optimizations. Embedding-based metrics (Kusner et al., 2015; Rubner et al., 1998; Clark et al., 2019; 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 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 imagination and apply our metric to various text-to-text generation tasks.

Mental Imagery The debate between pictorialists and propositionalists about how imagery information is stored in the human brain is still an open question in the neuroscience and psychology community (Troscianko, 2013). 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

This section describes how our IMAGINE metric evaluates the similarity between two pieces of text with the help of machine imagination. Figure 2 provides an overview of our method.

3.1 Model Details

CLIP We use the cross-modal retrieval model, CLIP (Radford et al., 2021), for our evaluation purposes. CLIP jointly trains an image encoder and a text encoder to predict the correct pairing of image-text pairs with InfoNCE (van den Oord et al., 2018) on 400M image-text pairs gathered from the web. We utilize the CLIP-ViT-B/32 variant, which consists of a 12-layer, 8-head Transformer text encoder with a hidden size of 512, and a Vision Transformer (ViT) (Dosovitskiy et al., 2021; Vaswani et al., 2017) image encoder adopting the BERT-base configuration and using a 32×32 input patch size. Both the text and image representations are normalized and projected into the multi-modal space before computing pairing likelihood through cosine similarity.

StableDiffusion We perform text-to-image generation with StableDiffusion (Rombach et al., 2022), which is a denoising diffusion probabilistic model (Ho et al., 2020). The model comprises three key components: a text encoder, a diffusion model, and an autoencoder. The text encoder, adopted from the frozen CLIP-ViT-L/14 (Radford et al., 2021), is utilized to encode the input text into textual embeddings. The diffusion model, which leverages UNet (Ronneberger et al., 2015) for noise estimation, is modified to attend to the input textual embeddings. We conduct experiments with StableDiffusion-v1-1, which was trained with LAION (Schuhmann et al., 2022), using 256×256 images for pre-training, followed by 512×512 images for fine-tuning.

3.2 IMAGINE Similarity Score

In our proposed approach, as depicted in Figure 2, the computation of IMAGINE consists of three sequential steps. Firstly, the StableDiffusion model (Rombach et al., 2022) is utilized to generate descriptive images, referred to as machine imagination, from the two text snippets being compared. Secondly, both the text snippets and the generated images are encoded using the CLIP model (Radford et al., 2021). Finally, IMAGINE is calculated by computing the cosine similarities of the resulting text and visual features, both in a mono-modal and cross-modal manner.

Step 1: Render Imagination For each image, StableDiffusion randomly initializes a latent matrix $\boldsymbol{H}$ from the standard normal distribution and uses the encoder of the pre-trained autoencoder to encode $\boldsymbol{H}$ into the lower-resolution latent map $\boldsymbol{z}_T$ (T is the total inference steps). At each step t, the diffusion model estimates the noise, ϵ, and subtracts it from $\boldsymbol{z}_t$. The decoder of the pretrained autoencoder takes the final noise-free latent map $\boldsymbol{z}$ and generates the image prediction $\boldsymbol{I}$ of size 512×512.

Step 2: Extract Feature In the previous step, we generate the corresponding images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ for the pair of text $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ for comparison with the text-to-image synthesis backbone. Then we pass the machine-generated images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ and the input text $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ through corresponding CLIP encoders to receive the visual representations $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and the textual representation $\boldsymbol{t}_1$, $\boldsymbol{t}_2$.

Step 3: Measure Similarity With $\mathrm{sim}(\cdot, \cdot)$ denoting the process of first normalizing the two vectors, then computing their cosine similarity, we compute two types of similarity scores for IMAGINE with the extracted textual and visual features:

(1) IMAGINE$_{image}$ computes the visual representation similarity between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$:

$$\text{IMAGINE}_{image} = \mathcal{F}\left(\mathrm{sim}(\boldsymbol{v}_1, \boldsymbol{v}_2)\right) \qquad (1)$$

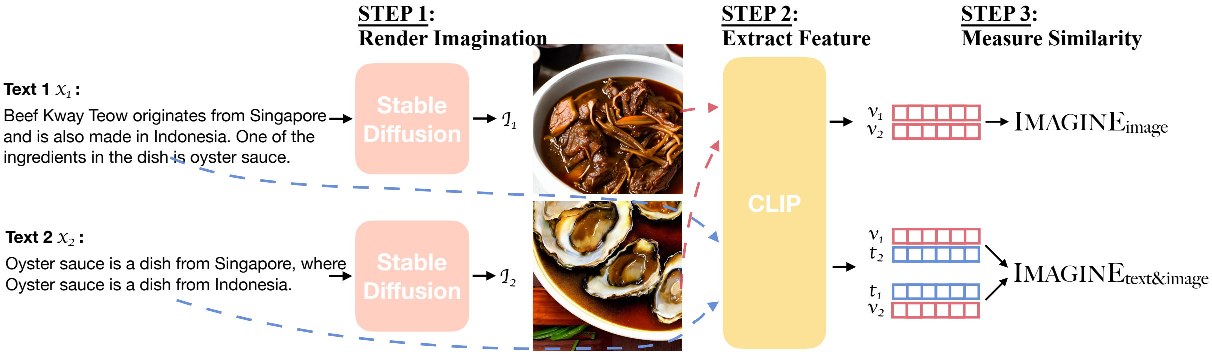

STEP 1:
Render Imagination
STEP 2:
Extract Feature
STEP 3:
Measure Similarity
Text 1 x_1:
Beef Kway Teow originates from Singapore and is also made in Indonesia. One of the ingredients in the dish is oyster sauce.
Stable Diffusion
I_1
Text 2 x_2:
Oyster sauce is a dish from Singapore, where Oyster sauce is a dish from Indonesia.
Stable Diffusion
I_2
CLIP
v_1
v_2
IMAGINE$_{\text{image}}$
v_1
t_2
t_1
v_2
IMAGINE$_{\text{text\&image}}$

Figure 2: Illustration of the computation process of the IMAGINE metric. Given the two pieces of text for comparison, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we render the machine imagination by generating two images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ with the pre-trained StableDiffusion (Rombach et al., 2022). We extract features of the input text and corresponding generated images with CLIP (Radford et al., 2021). We receive two variants of IMAGINE by computing the cosine similarity of the extracted features, in which IMAGINE$_{image}$ measures mono-modal similarities on the visual side, while IMAGINE$_{text\&image}$ conducts cross-modal matching.

(2) IMAGINE$_{text\&image}$ (IMAGINE$_{t\&i}$) takes both the text and the generated image into consideration, and conducts cross-modal comparisons between ($\boldsymbol{t}_1$, $\boldsymbol{v}_2$), as well as ($\boldsymbol{t}_2$, $\boldsymbol{v}_1$):

$$\text{IMAGINE}_{t\&i} = \mathcal{F}\left(\frac{\text{sim}(\boldsymbol{t}_1, \boldsymbol{v}_2) + \text{sim}(\boldsymbol{t}_2, \boldsymbol{v}_1)}{2}\right) \tag{2}$$

The cosine similarity between the text and image representations theoretically has a range of $[-1, 1]$. However, in practice, the IMAGINE similarity scores tend to cluster within a more narrow interval $[l, h]$. Following Hessel et al. (2021), we use a linear function $\mathcal{F}$ to stretch the similarity score distribution to the range of $[0, 1]$, which is also the score range for most of the automatic metrics covered in this study. Eq. (3) shows how we re-scale the similarity score s into s'. Appendix Figure 6 plots the two IMAGINE variants' distributions before and after rescaling.

$$\begin{aligned} s' &= \frac{s - l}{h - l}, \\ [l, h] &= \begin{cases} [0.1, 1.0], & \text{for IMAGINE}_{image}, \\ [0.1, 0.4], & \text{for IMAGINE}_{text\&image}. \end{cases} \end{aligned} \tag{3}$$

3.3 Integration with Existing Metrics

The IMAGINE similarity scores can serve as standalone automatic metrics. Additionally, IMAGINE can be incorporated as an extension to existing metrics, as it offers multimodal references and addresses the limitations of current text-only evaluations that only compare tokens or text embeddings. This mimics the human process of comprehending text, where both text and visual imagination are utilized. The integration of IMAGINE with other automatic metrics is straightforward, achieved by summing the IMAGINE similarity score with the other automatic metric's score for each example:

$$metric\_score' \mathrel{+}= \text{IMAGINE}_{similarity\_score} \tag{4}$$

4 Experimental Setup

4.1 Tasks, Datasets, and Models

We evaluate our approach on three popular natural language generation tasks: machine translation, abstractive text summarization, and open-ended text generation.

Machine Translation We use Fairseq (Ott et al., 2019) to generate English translation from German on IWSLT'14 (Cettolo et al., 2014) and WMT'19 (Barrault et al., 2019) datasets.

Abstractive Text Summarization We use the implementation of Li et al. (2017) to generate summarization on DUC2004[1] and use ProphetNet (Qi et al., 2020b) for generation on Gigaword.[2] Both datasets are built upon news articles.

Open-ended Text Generation We perform experiments on the ActivityNet (Heilbron et al., 2015) subset of HellaSwag (Zellers et al., 2019), which is a benchmark for commonsense natural language inference that ask the model to predict the most likely follow-up among several choices given a specific

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

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

Metric	IWSLT'14			WMT'19		
	Original	$+\text{IE}_{image}$	$+\text{IE}_{text\&image}$	Original	$+\text{IE}_{image}$	$+\text{IE}_{text\&image}$
BLEU-1	21.47	21.38±1.53	**21.86**±0.82	13.74	14.71±1.19	**16.40**±0.73
BLEU-2	20.82	21.17±1.45	**21.53**±0.68	12.50	12.93±1.13	**15.11**±0.64
BLEU-3	19.17	19.88±1.39	**20.31**±0.62	11.31	12.07±1.09	**13.90**±0.58
BLEU-4	17.60	18.57±1.36	**19.08**±0.60	9.10	9.15±1.06	**11.84**±0.54
METEOR	20.60	21.44±1.54	**21.30**±0.99	13.47	14.77±1.33	**16.80**±0.91
ROUGE	20.55	20.69±1.54	**21.26**±0.80	11.40	11.58±1.16	**14.34**±0.68
CIDEr	21.98	22.12±0.24	**22.25**±0.07	11.82	11.86±0.18	**12.05**±0.07
BERTScore	23.95	24.02±1.41	**24.09**±0.65	17.01	17.08±1.22	**18.88**±0.78
BLEURT	22.93	22.99±0.64	**23.40**±0.41	18.81	19.36±0.82	**19.59**±0.37

Table 1: The effect of applying our IMAGINE similarities on automatic metrics for machine translation, reflected in the Pearson correlation with human judgments. The image generation process is conducted over five different random seeds for each piece of text. We report the mean and standard deviation of the repeated runs. IE: IMAGINE.

context. The dataset is derived from ActivityNet video captions and we use it for the task of sentence completion, where the model is given a context and asked to complete the sentence. The predicted sentence endings generated by StoryEndGen (Guan et al., 2019) and GPT-2 (Radford et al., 2019) are collected and used in the following evaluation.

4.2 Automatic Metrics

Machine Translation & Summarization In the evaluation of machine translation and text summarization tasks, it is a common practice to compare the predicted text with the reference. Adhering to previous studies, we present results using reference-based metrics. For machine translation, we present scores using BLEU-n (n=1,2,3,4) (Papineni et al., 2002), METEOR(Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015). Meanwhile, for text summarization, we present ROUGE-n (n=1,2) (Lin, 2004) precision scores. Additionally, we report the scores of ROUGE-L (Lin, 2004), BERTScore (Zhang* et al., 2020), and BLEURT (Sellam et al., 2020) for both tasks.

Open-ended Text Generation In the context of open-ended text generation, where the number of possible answers for a given scenario can be inexhaustible, evaluating the quality of generated text through a comparison with a fixed set of references is challenging. To address this issue, previous studies have proposed to utilize reference-free metrics to evaluate the quality of the generated text. In this work, we experiment with the following reference-free metrics which assess model degeneration: (1) div-n = $\frac{|\text{unique } n\text{-grams}|}{|\text{total } n\text{-grams}|}$ measures sequence level repetition by computing the portion of duplicate n-grams (n=2,3,4) (Welleck et al., 2020). (2) diversity = $\prod_{n=2}^{4}$ rep-n measures the diversity of n-grams (Su et al., 2022), and assesses the model degeneration. (3) distinct-n = $\frac{|\text{unique } n\text{-grams}|}{|\text{length of text}|}$ measures the portion of distinct n-grams (here n=2) in the text (Li et al., 2016). In addition, we report results on BERTScore (Zhang* et al., 2020) and BLEURT (Sellam et al., 2020) for comparison of contextual similarity.

4.3 Human Evaluation

We invite Amazon Mechanical Turk[3] annotators to evaluate the quality of the generated text. Due to cost constraints, when conducting human evaluation, we randomly sample 1,000 test examples for each dataset, except for DUC2004 which has 500 examples in the test set. Each example is evaluated by three human judges using a 5-point Likert scale, which assessed the fluency, grammar correctness, and factual consistency of the generated text with the reference text. The overall human assessment score is calculated as the mean of the scores obtained from the three aspects. We compute the Pearson correlation (Freedman et al., 2007) between the human scores and the scores obtained from the automatic metrics, and the results are reported as a multiple of 100 for clarity.

5 Results and Analysis

5.1 Main Results

Machine Translation Table 1 presents the results of the system-level Pearson correlation with human judges when extending the IMAGINE similarity metric to various existing automatic natural language generation (NLG) metrics on the IWSLT'14 and WMT'19 German-to-English datasets. The results demonstrate that the addition of both IMAGINE$_{image}$ and IMAGINE$_{text\&image}$

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

Metric	DUC2004			GigaWord		
	Original	+IE$_{image}$	+IE$_{text\&image}$	Original	+IE$_{image}$	+IE$_{text\&image}$
ROUGE-1	13.66	**16.77**±1.31	13.45±0.80	12.90	**17.52**±0.73	16.78±0.66
ROUGE-2	9.74	**15.71**±1.65	11.19±1.08	7.75	**14.26**±0.83	13.33±0.77
ROUGE-L	13.14	**16.35**±1.47	13.17±0.95	14.31	**17.44**±0.77	16.78±0.70
BERTScore	19.44	**20.60**±1.29	20.26±0.78	19.59	**20.47**±0.64	20.10±0.57
BLEURT	23.59	**25.20**±0.72	24.46±0.42	20.23	**21.08**±0.39	20.74±0.35

Table 2: The effect of applying our ImaginE similarities on automatic metrics for text summarization, reflected in the Pearson correlation with human judgments. The image generation process is conducted over five different random seeds for each piece of text. We report the mean and standard deviation of the repeated runs. IE: ImaginE.

Metric	Reference-based			Reference-free		
	Original	+IE$_{image}$	+IE$_{text\&image}$	Original	+IE$_{image}$	+IE$_{text\&image}$
div-2	27.21	28.01±0.49	**28.08**±0.34	27.21	26.51±0.42	**27.29**±0.58
div-3	26.80	27.67±0.49	**27.78**±0.35	26.80	26.17±0.43	**26.98**±0.59
div-4	26.20	27.14±0.48	**27.28**±0.36	26.20	25.71±0.44	**26.55**±0.60
diversity	27.40	28.19±0.41	**28.23**±0.30	27.40	26.89±0.36	**27.55**±0.50
distinct-2	26.72	27.76±0.56	**27.90**±0.40	26.72	25.54±0.48	**26.49**±0.66
BERTScore	23.47	**25.92**±0.50	25.43±0.36	25.10	23.47±0.56	**25.26**±0.78
BLEURT	19.99	**22.47**±0.83	21.55±0.72	18.70	19.67±0.88	**20.56**±1.25

Table 3: The effect of applying our ImaginE similarities on ActivityNet for open-ended text generation, reflected in the Pearson correlation with human judgments. In the "Reference-based" setting, we compare the predictions with the references, while in the "Reference-free" setting, we compare the predictions with the input contexts. The image generation process is conducted over five different random seeds for each piece of text. We report the mean and standard deviation of the repeated runs. IE: ImaginE.

improves the Pearson correlation for all metrics listed. Among the two variants, the mean of ImaginE$_{text\&image}$ consistently performs better on both datasets. It is observed that there is a more substantial variance in ImaginE$_{image}$, which is attributed to the difference in the images generated by the StableDiffusion model (Rombach et al., 2022) due to varying random seed and initialization values. As a result, ImaginE$_{image}$, which compares two machine-generated images, has a higher standard deviation compared to ImaginE$_{text\&image}$.

Abstractive Text Summarization The results in Table 2 demonstrate the system-level Pearson correlation with human judges when incorporating our ImaginE similarity into existing automatic NLG metrics on the DUC2004 and Gigaword datasets. In alignment with the observations made in the machine translation task, the addition of both ImaginE$_{image}$ and ImaginE$_{text\&image}$ results in an improvement in Pearson correlation across all metrics. On the two summarization datasets, we notice that the correlation after incorporating ImaginE$_{image}$ exhibits higher mean values along with larger variances compared to the correlation with ImaginE$_{text\&image}$.

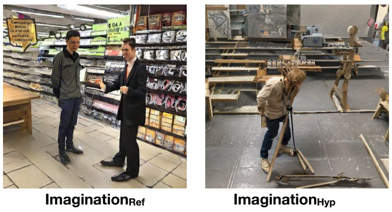
Open-ended Text Generation For the sentence completion task, we conduct evaluations in two setups. In the reference-based evaluation, we compare the predicted sentence ending with the ground-truth ending provided in the dataset. In reference-free evaluation, we compare the predicted sentence ending with the input context. This setup is designed to assess the coherence of the prediction with the input context, as it is hypothesized that a high-quality prediction for open-ended text generation should be consistent with the input context.

The results of extending our ImaginE similarity metric to existing automatic NLG metrics for the sentence completion task on the ActivityNet dataset are shown in Table 3. In the reference-based setting, both ImaginE variants demonstrate improvement over the listed metrics and exhibit comparable performances. In the reference-free setting, the introduction of ImaginE$_{text\&image}$ continues to enhance the Pearson correlation, while the implementation of ImaginE$_{image}$ results in a decrease in correlation. One possible reason for the decline in correlation when ImaginE$_{image}$ is used in the reference-free setting of the sentence completion task on ActivityNet (which is comprised of video captions) is that, despite the requirement for the predicted continuation to be coherent with the given context, the visual representation of the context and continued text may differ greatly in this scenario (e.g., due to a plot twist in the video). Consequently, direct comparison of images through ImaginE$_{image}$ may result in a decrease in correla-

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
ROUGE-L	50.00
BERTScore	58.88
BLEURT	55.73
ImaginE$_{\text{image\&text}}$	23.85

Figure 3: A case study on IWSLT'14 German-to-English translation with images rendered by StableDiffusion-v2-1. Src.: input source text. Ref.: reference text. Hyp.: generated hypothesis text.

tion. However, the inherent coherence between the input text and the continued text may be captured through cross-modal comparison, which may explain why IMAGINE$_{text\&image}$ still improves the correlation for the listed metrics.

5.2 Performance Analysis

Why is ImaginE helpful? As shown in Tables 1 to 3, the incorporation of certain variants of IMAGINE improves the correlation between the reference-based and reference-free metrics and human scores in the majority of cases. This indicates the usefulness of extending text-only metrics with multi-modal knowledge. However, how do these machine imaginations actually help text understanding and evaluation? In this section, we further explore 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.

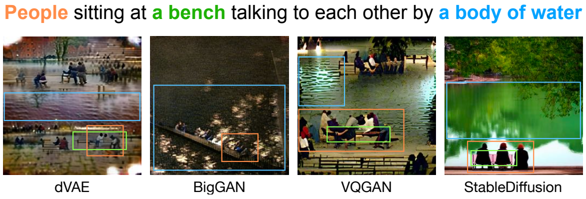
Case Study Figure 3 shows an example in which IMAGINE effectively detects the dissimilarity in keywords between two text snippets. Despite the similarity in sentence structure between the reference and hypothesis, the crucial distinction lies in the inclusion of the terms "manager" and "ladder". While traditional automatic metrics that rely on n-grams matching (BLEU, ROUGE) or textual embedding comparison (BERTScore, BLEURT) may exhibit high scores, the quality of the generated text remains questionable. In contrast, IMAGINE generates distinctive images and exhibits a relatively low cross-modal similarity score, which aligns with human perception.

Metric	**Original**	+IE$_{i(\text{dVAE})}$	+IE$_{i(\text{BigGAN})}$	+IE$_{i(\text{VQ-GAN})}$	+IE$_{i(\text{SD})}$
ROUGE-1	13.7	15.9 ± 0.9	15.7 ± 1.0	15.9 ± 0.8	**16.8** ± 1.3
ROUGE-2	9.7	14.9 ± 1.2	14.6 ± 1.3	14.9 ± 1.0	**15.7** ± 1.7
ROUGE-L	13.1	16.0 ± 1.0	15.8 ± 1.1	16.0 ± 0.9	**16.4** ± 1.5

Table 4: The Pearson correlations with human judges when using IMAGINE$_{image}$ (IE$_i$) to augment ROUGE-1/2 and ROUGE-L on DUC2004. We compute four sets of IMAGINE$_{image}$ similarity scores (mean±std) with dVAE, BigGAN, VQGAN, and StableDiffusion (SD).

	dVAE	BigGAN	VQGAN	StableDiffusion
Entity Recall	88.8%	41.2%	87.2%	94.1%

Table 5: Entity recall rate on the visualizations for Flickr30k captions. We report results for images generated by dVAE, BigGAN, VQGAN, and StableDiffusion.



People sitting at a bench talking to each other by a body of water
dVAE
BigGAN
VQGAN
StableDiffusion

Figure 4: An example caption from Flickr30k Entities, and images rendered by dVAE, BigGAN, VQGAN and StableDiffusion. The bounding boxes point to the visualizations of the entities marked in the same color.

Sensitivity to Different Image Generation Backbones In previous sections, we utilize StableDiffusion (Rombach et al., 2022) as the image generation backbone for IMAGINE. Here, we examine the influence of the image generation backbone on the evaluation performance of IMAGINE by conducting experiments on the DUC2004 dataset for summarization and comparing StableDiffusion with three alternative models: dVAE (Ramesh et al., 2021), BigGAN (Brock et al., 2019), and VQGAN (Esser et al., 2021). The results, as shown in Table 4, indicate comparable performance of IMAGINE$_{image}$ with dVAE and VQGAN, both of which outperform BigGAN across all metrics. StableDiffusion achieves the highest mean value, but also displays the largest variance among the models. These findings highlight the significance of considering the image generation architecture when evaluating text, as it can result in varying machine-generated images and affect the final evaluation outcomes.

Reliability of Machine-Generated Images The reliability of IMAGINE's visualization capability is further evaluated on the Flickr30k Entities dataset (Plummer et al., 2015), which consists of annotated image captions. We randomly sample

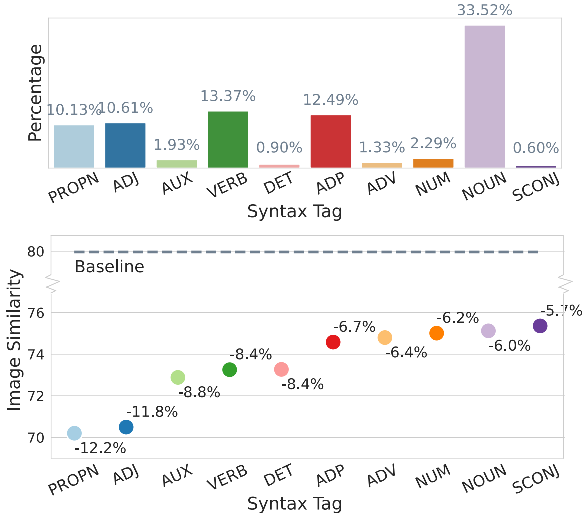


Figure 5: The influence on visualization when masking tokens of different syntax tags. Upper: The occurrence frequency of each syntax tag in DUC2004. Lower: The relative image similarity decrease after masking each syntax tag. Baseline: The average intra-group pairwise image similarity. The top-10 syntax tags that have the most significant impact on visualization are listed here.

100 captions and use the four generative backbones to render images. We present the captions and generated images to human annotators, and ask them to indicate if the entities mentioned in the captions are visually represented. The results, in terms of entity recall rates, are presented in Table 5. A higher recall rate indicates that the text-to-image generator is more capable of visualizing the content described in the text. The results show that StableDiffusion has the highest entity recall rate of approximately 94%, followed closely by dVAE and VQGAN. In contrast, BigGAN has the lowest recall rate of around 41%. An example of entity recall for a set of images generated by the four generative backbones is shown in Figure 4.

Syntax Importance to Machine-Generated Images We evaluate the significance of different syntax tokens in the image generation process using the DUC2004 summarization dataset. We utilized the Stanza (Qi et al., 2020a) part-of-speech (POS) tagger to parse the text and created ablated examples by masking out a token of a specific syntax tag.[4] The visual similarity of the images generated from the ablated examples is then compared to the visualization of the original text. The results, as reported in Table 5, indicated that the removal of PROPN and ADJ tags has a significant impact on the visualization results, resulting in a 12% decrease in visual similarity. Conversely, removing NOUN tokens has a comparatively smaller effect. The most frequent NOUN, PROPN, and ADJ tokens in the DUC2004 dataset were listed in Table 6. For DUC2004 built upon new clusters, PROPN and ADJ tokens cover concrete concepts such as nations, corporations, and celebrities, while 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. Further analysis for other dataset domains can be found in the Appendix.

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

Which ImaginE Variant to Report? From Tables 1 to 3, we can see a mixed trend of performance between the two ImaginE variants. In general, ImaginE$_{text\&image}$ has smaller variances among repeated runs. Nevertheless, we would still suggest reporting both ImaginE variants since they conduct comparisons from different aspects, with ImaginE$_{image}$ comparing similarity within the visual modality, while ImaginE$_{text\&image}$ compares cross-modal similarity.

ImaginE as a Standalone Metric Table 7 presents the Pearson correlation with human evaluations on each dataset when utilizing the two ImaginE variants as standalone metrics. The results reveal that both ImaginE variants demonstrate a lower correlation compared to other metrics as reported in Tables 1 to 3. Additionally, the scores produced by ImaginE are not determinate, given the probabilistic nature of text-to-image models that generate various images with different random seeds. Hence, ImaginE may not be an optimal choice as a standalone metric. Nonetheless, it is important to emphasize the capability of ImaginE in introducing multimodal aspects to traditional text-only metrics. In this study, integrating Imag-

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

	IWSLT'14	WMT'19	DUC2004	GigaWord	AN(w/ ref)	AN(w/o ref)
IE$_i$	19.1±1.5	13.8±1.7	10.6±1.5	15.9±1.1	18.9±1.5	16.8±1.9
IE$_{t\&i}$	18.0±1.5	12.9±1.8	9.6±1.6	15.3±1.1	18.4±1.6	18.2±1.8

Table 7: The Pearson correlation between IMAGINE variants and human assessments on each dataset. Here we use IMAGINE$_{image}$ (IE$_i$) and IMAGINE$_{text\&image}$ (IE$_{t\&i}$) as two individual metrics. AN: ActivityNet, "w/ ref": reference-based, "w/o ref": reference-free.

INE with text-only metrics leads to an improvement in the Pearson correlation with human evaluations. Future work may explore alternative methods of integrating multimodal information in text evaluation.

6 Conclusion

We present IMAGINE, a novel automatic evaluation metric for NLG that is based on machine imagination. Our experiments on five datasets across three different NLG tasks demonstrate the potential of incorporating IMAGINE similarity scores as a supplement to existing automatic NLG metrics, which can lead to improvement in their correlation with human evaluations in various scenarios. In the future, it is interesting to explore effective ways of visualizing abstract concepts, and how to incorporate machine imagination efficiently. We hope our work can contribute to the discussion and advancement of multi-modal studies.

Limitations

The current limitations of IMAGINE include the length constraint of the CLIP text encoder, which is limited to 77 BPE tokens (including [BOS] and [EOS]), thus limiting its applicability to longer text generation tasks such as story generation or document summarization. As a metric that relies on "machine imagination", IMAGINE is limited by the inherent limitations of the generative models for images. The non-determined nature of machine-generated images can lead to non-determined IMAGINE scores. Possible solutions to mitigate this issue includes but are not limited to fixing the random seeds or repeating the evaluation process several times to reduce the variance effect. Additionally, it remains a challenge for machines to properly visualize certain abstract concepts or numerical values, which could limit the scope of IMAGINE's applicability.

Ethical Statement

Our study has received IRB exempt status and the estimated hourly wage paid to MTurk annotators is $12. It is important to note that our "imagination" approach may raise questions of fairness if the training dataset for CLIP or StableDiffusion contains any biases. This could result in a tendency for IMAGINE to generate certain types of images based on what it has seen in the training data. While we did not observe such issues in our study, it is important to consider that such unfair behavior would undermine the effectiveness of IMAGINE as an evaluation tool.

All of the datasets used in our study on machine translation, abstractive text summarization and open-ended text generation tasks are publicly available. We use the public repositories to implement IMAGINE. The implementations of image generators used in our study are DALL-E(dVAE+CLIP),[5] Big-Sleep(BigGAN+CLIP),[6] VQGAN+CLIP,[7] and StableDiffusion.[8]

Acknowledgement

The research was sponsored by the U.S. Army Research Office and was accomplished under Contract Number W911NF-19-D-0001 for the Institute for Collaborative Biotechnologies. This work was also supported by the National Science Foundation award #2048122. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Gra-

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

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

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

[8] https://huggingface.co/CompVis/stable-diffusion-v1-1

ham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 1–61. Association for Computational Linguistics.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE.

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

David Freedman, Robert Pisani, and Roger Purves. 2007. *Statistics (international student edition)*.

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

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6473–6480. AAAI Press.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

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

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

Stephen M Kosslyn, Giorgio Ganis, and William L Thompson. 2001. Neural foundations of imagery. In *Nature reviews neuroscience*, volume 2, pages 635–642. Nature Publishing Group.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*

2015, Lille, France, 6-11 July 2015, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. ViLBERTScore: Evaluating image caption using vision-and-language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, Online. Association for Computational Linguistics.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 212–228. Springer.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2091–2100. Association for Computational Linguistics.

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.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 589–597. Association for Computational Linguistics.

Chi-kiu Lo. 2019. Yisi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 507–513. 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 *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Marilyn A Nippold and Jill K Duthie. 2003. Mental imagery and idiom comprehension: a comparison of school-age children and adults. In *Journal of Speech, Language, and Hearing Research*. ASHA.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Joybrata Panja and Sudip Kumar Naskar. 2018. ITER: improving translation edit rate through optimizable edit costs. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 746–750. Association for Computational Linguistics.

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. In *Proceedings of the National Academy of Sciences*, volume 112, pages 10089–10092. National Acad Sciences.

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. In *International Journal of Computer Vision*, volume 123, pages 74–93.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020a. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020b. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2401–2410. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. A metric for distributions with applications to image databases. In *Proceedings of the Sixth International Conference on Computer Vision (ICCV-98), Bombay, India, January 4-7, 1998*, pages 59–66. IEEE Computer Society.

Mark Sadoski and Allan Paivio. 1994. A dual coding view of imagery and verbal processes in reading comprehension. In *Theoretical Models and Processes of Reading*, pages 582–601. International Reading Association.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006*, pages 223–231. Association for Machine Translation in the Americas.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Christoph Tillmann, Stephan Vogel, Hermann Ney, A. Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*. ISCA.

Jesús Tomás, Josep Àngel Mas, and Francisco Casacuberta. 2003. A quantitative method for machine translation evaluation. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, pages 27–34.

Emily T Troscianko. 2013. Reading imaginatively: the imagination in cognitive science and cognitive literary studies. In *Journal of Literary Semantics*, volume 42, pages 181–198. De Gruyter Mouton.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 505–510. The Association for Computer Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

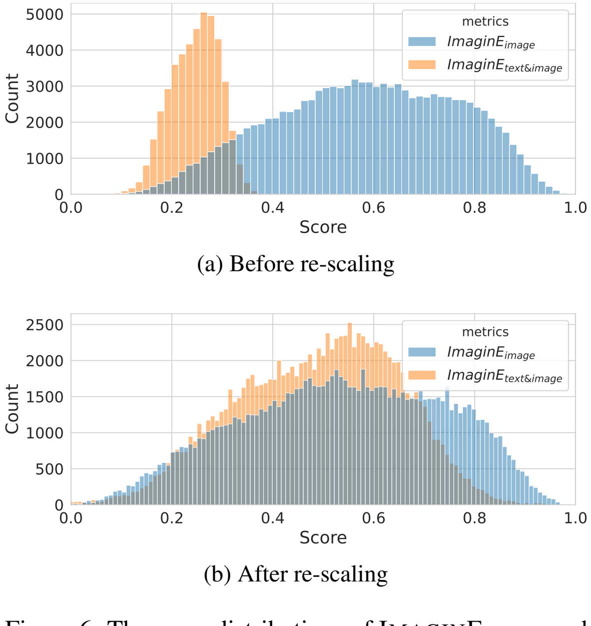
Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Score Distributions

In this study, we use cosine similarity to evaluate the similarity between features, which yields a score distribution in the range of $[-1, 1]$. However, our results indicate that negative scores were not observed when computing the similarities between the features generated by CLIP. The score distributions of the two IMAGINE variants are depicted in Figure 6. Prior to re-scaling, the scores generated by IMAGINE$_{image}$ typically fall within the range of $[0.1, 0.4]$, while those generated by IMAGINE$_{text\&image}$ are within $[0.1, 1.0]$. Following re-scaling, both IMAGINE metrics are linearly transformed to lie within the range $[0, 1]$.



(a) Before re-scaling

(b) After re-scaling

Figure 6: The score distributions of IMAGINE$_{image}$ and IMAGINE$_{text\&image}$ before and after re-scaling.

A.2 Syntax Importance to Imaginations

In Section 5.2, we discussed the impact of DUC2004 Part-of-Speech (POS) tags on the quality of generated images. In this section, we extend our examination to another dataset domain, the Flickr30k Entities dataset (Plummer et al., 2015), which is an image captioning corpus. While the domain of the Flickr30k Entities dataset is distinct from that of the DUC2004 (based on news articles), similar trends are observed. The results displayed in Figure 7 also suggest that concrete concepts are easier to be visualized and play a more significant role in the visualization process, similar to the results observed in Figure 5.

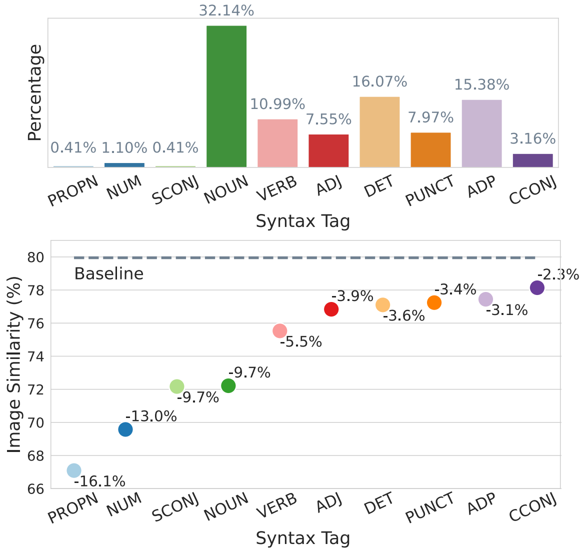


Figure 7: The influence on visualization when masking tokens of different syntax tags. Upper: The occurrence frequency of each syntax tag in Flickr30k. Lower: The relative image similarity decrease after masking each syntax tag. Baseline: The average intra-group pairwise image similarity. The top-10 syntax tags that have the most significant impact on visualization are listed here.