PHIVALBLIP
ARTISTIC PHOTO EVALUATION SYSTEM BASED ON BILINGUAL LANGUAGE-IMAGE PRE-TRAINING

Junjie Zhao & Zachary Tan
245-C418
University of Rochester
Rochester, NY 14627, USA
jzhao58@u.rochester.edu & ztan11@u.rochester.edu

ABSTRACT

Advancements in multi-modal Artificial Intelligence models have opened new avenues for creative feedback tools that cater to visual artists and photographers. Traditional approaches to improvement in these areas often require receiving professional artistic feedback. These methods are often resource-intensive, requiring significant time and financial investment. Here we present a photo evaluation system based on a multi-modal model trained to evaluate photos and deliver constructive feedback, focusing on holistic attributes of visual artwork, such as content, color, composition, and quality. We used BLIP (Bootstrapping Language-Image Pre-training) as a foundation and made improvements to the architecture while fine-tuning it with a custom dataset. It can interpret complex visual cues and translate them into text-based feedback that assists artists in refining their work independently. Our approach demonstrates the potential for multi-modal models to support and enhance the artist's creative process without necessitating frequent external consultations.

1 INTRODUCTION

Most current pre-trained image-to-text multi-modal models focus on high-level analysis of an image: image captioning, resolving queries, image-text retrieval, and more. However, there are few attempts to explore the capacity of image-to-text multi-modal models to understand images in a holistic way from the view of artists, and the ability to perform image evaluation specifically from an artistic perspective.

There are also very few data sets that can be used to train models in this manner. The majority of traditional datasets focus on image captioning, merely giving a one-sentence summary of what the image depicts. Therefore, we prepared a new dataset addressing the task of artistic image evaluation. We trained our model to do fine-tuning with different hyper-parameters related to the visual understanding of the base BLIP model. Furthermore, based on our image evaluation criteria, we tested innovations and attempted to make improvements to the architecture of the BLIP-based model. We tested a multi channel approach where the BLIP model would be divided in the last step to focus on the different paramters. to see whether the change we made to the architecture of the model can improve the performance of the model in the specific task of artistic image evaluation.

By addressing these challenges and introducing architectural improvements, our work aims to bridge the gap between multi-modal AI models and the nuanced needs of the artistic community, providing a robust tool for image evaluation from an artist's perspective.

2 RELATED WORK

2.1 BOOTSTRAPPING LANGUAGE-IMAGE PRE-TRAINING

Bootstrapping Language-Image Pretraining (BLIP) Li et al. (2022) is a multi-modal framework designed to connect vision and language understanding. Its implementation is discussed in Li et al. (2022), and we will give a brief overview of it here. Unlike other models that mainly focus on either retrieval-based or generative tasks, BLIP is capable of both, making it highly versatile for vision-language applications, and our model of choice for this project. BLIP operates by using a large amount of image-text pairs and aligning visual features with linguistic representations to achieve a unified understanding. Its design enables both supervised and unsupervised learning techniques, enabling it to generalize across various tasks even when faced with noisy or diverse datasets.

BLIP's architecture consists of three main components: the vision encoder, the text encoder, and the decoder. The vision encoder is usually implemented using pre-trained vision transformers (ViT). It extracts high-dimensional feature representations from images. The text encoder is based on transformers. It processes textual data to create corresponding embeddings. These embeddings are aligned through a shared multi-modal space using contrastive learning. The decoder is also transformer-based. It allows BLIP to generate text outputs, making it capable of tasks like image captioning and querry answering.

A critical feature of BLIP is its bootstrapped learning process. During pre-training, it employs a contrastive loss to align images and text, a captioning module to generate textual descriptions, and a filtering mechanism to refine noisy image-text pairs. This combination ensures that BLIP learns meaningful and accurate representations, paving the way for high performance across downstream tasks.

BLIP's training process starts with vision-language pre-training (VLP) on a large dataset of image-text pairs. It uses a dual-stream architecture during this phase, where the vision and text encoders operate independently before fusing their outputs. The contrastive learning objective ensures that semantically related image-text pairs are brought closer in the shared embedding space. Alongside, the captioning module generates descriptive text for each image, providing supervision for generative tasks.

During downstream fine-tuning, BLIP adapts to specific applications like visual question answering or text-based image retrieval. Its modular design allows fine-tuning of individual components, enhancing flexibility. For instance, one could adjust the decoder for better generative performance or fine-tune the encoder for more precise retrieval tasks. This adaptability makes BLIP a powerful tool for vision-language integration.

BLIP has a wide range of applications. In e-commerce, it aids in generating detailed product descriptions from images, improving online shopping experiences. For healthcare, it supports medical imaging by pairing visual scans with textual analyses, such as automated radiology reports. It is also extensively used in content creation, where it generates captions, enhances searchability, and provides context for visual media.

Moreover, BLIP has found a place in research and education, where it helps interpret complex diagrams and scientific visuals with textual annotations. It's also applied in accessibility technologies, enabling visually impaired users to understand images through text-to-speech systems.

2.2 VISION-LANGUAGE PRE-TRAINING

Vision-language pre-training (VLP) Gan et al. (2022) is a pre-trained model on large scale image-text pairs. It performs very well on downstream vision and language task. The framework of BLIP model is based on VLP, and BLIP offers more flexibility and perform better on a wide range of downstream tasks compared with VLP.

2.3 KNOWLEDGE DISTILLATION

Knowledge Distillation (KD) Hinton et al. (2015) is a technique that transforms "knowledge" from a model with a large number of parameters to a model with a smaller number of parameters but still

maintain relatively good performance. For the dataset BLIP used for training, the techniques of KD has been implemented. The researchers use CapFilt, a model that filters the text in the dataset to keep the text in a good narrative style and quality.

2.4 MEDBLIP

MedBLIP Chen et al. (2023) is a model trained with 3D medical images and text and aims for making computer-aided diagnoses(CAD) based on image scans and text descroptions in electronic health records, as done it partice. This study demonstrates the possibility and the capacity of BLIP model in many different scientific areas.

2.5 EVALUATION METRIC

The BLEU (Bilingual Evaluation Understudy) metric, introduced by Papineni et al. (2002), is a widely used automatic evaluation metric for natural language generation tasks. It is usually used in machine translation. It provides a quantitative measure of the similarity between two texts, enabling a method of evaluating the performance of language models.

BLEU operates on the principle of n-gram overlap. It measures the precision of n-grams in the generated text that also appears in the reference texts. This metric calculates precision for different n-gram lengths, for example, 1-gram, 2-gram, 3-gram, etc.) and combines them to provide a single score. To mitigate issues where overly shorter predictions score unrealistically high precision, BLEU includes a brevity penalty that penalizes outputs that are shorter than the reference texts. This ensures that the output will always be at least the length of the input.

The BLEU score is computed by the following equations:

Modified n-gram precision

BLEU computes precision by counting the n-grams in the generated text that match those in the reference texts, clipped to the maximum number of occurrences in the references. This avoids over-rewarding repeated n-grams in the hypothesis.

$$P_n = \frac{\sum_{C \in \text{Candidates}} \sum_{n\text{-gram} \in C} \min(\text{Count}_{\text{gen}}(n\text{-gram}), \text{Count}_{\text{ref}}(n\text{-gram}))}{\sum_{C \in \text{Candidates}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{gen}}(n\text{-gram})} \tag{1}$$

Brevity Penalty (BP)

The brevity penalty accounts for cases where a model might generate overly concise outputs. It is calculated as:

$$BP = \begin{cases} 1 & \text{if length}_{\text{gen}} > \text{length}_{\text{ref}} \\ e^{(1 - \frac{\text{length}_{\text{ref}}}{\text{length}_{\text{gen}}})} & \text{if length}_{\text{gen}} \leq \text{length}_{\text{ref}} \end{cases} \tag{2}$$

Final BLEU Score

The final BLEU score is the geometric mean of n-gram precisions (usually up to 4-grams), weighted by their respective importance, multiplied by the brevity penalty:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n\right) \tag{3}$$

Here, w_n are the weights for each n-gram, typically set equally (e.g., 0.25 for unigrams to 4-grams).

2.6 STRENGTHS AND LIMITATIONS

BLEU is computationally efficient, language-agnostic, and correlates well with human judgments in many contexts. However, it has several limitations. The first limitation is its insensitivity to

semantic similarity. BLEU only captures surface-level overlap and may undervalue paraphrased outputs. Another limitation is its dependence on reference quality. The choice and diversity of reference texts significantly impact BLEU scores. Finally, it tends to prefer longer text. While the brevity penalty addresses short hypotheses, BLEU can favor verbose outputs that maximize n-gram overlap.

Despite these limitations, we found that BLEU remained an accurate general sense of how close our generated text was to the reference text. We tested other evaluation metrics, such as ROUGE, and also included them in our results.

3 METHODS

3.1 DATASET

Existing datasets that capture specific artistic attributes were scarce and unreliable, so we prepared our own. This new dataset is split into 514 image-text pairs for training and 400 image-text pairs for testing. Each image is evaluated on four main criteria: content, color, composition, and quality. Content focuses on the objects that appear in the image. This performs similarly to other object identification models. Color focuses on how the colors in the image work together. It gives an in-depth description of color categories, such as bright colors or cool colors, and also focuses on how the color impacts the rest of the image. Composition focuses on more photographically elements such as framing, positioning, where the main object is placed, objects in relation to each other, and the layers. Quality describes how focused the image is. It mentions aspects such as noise level, and quality as a whole.

However, some potential issues with this dataset is the bias of certain types of pictures over others. In our dataset, a majority of the images depict food or animals. This leads the model to reacting more favorably to similar categories of images, while occasionally failing the same specificity and correctness on more diverse images. Furthermore, the majority of the images being used were high-quality and were evaluated positively. There was little variation in the quality of images used, so our model gives overwhelmingly positive feedback to the majority of images.

Figures 1 shows an example of an image from the dataset, and Figure 2 shows an example of the corresponding text.

Figure 1: ID1.jpeg

```
[
    {
        "image": "ID1.jpeg",
        "content": "The photograph captures an engaging street performance scene where the
                    musician is immersed in his craft, with the audience in the background
                    watching attentively. This moment highlights the vibrancy and intimacy of
                    urban life, creating a genuine connection between the performer and the
                    observers. The mood is both dynamic and personal, reflecting the
                    authenticity of public artistry.",
        "color": "The color palette in the image is vibrant and alive, with the bright yellows
                    of the musician's shorts standing out against the urban background filled
                    with glowing neon signs. The play of artificial and natural lighting creates
                    a dynamic harmony, making the scene feel warm and inviting while retaining
                    the energy of the street environment.",
        "composition": "The composition is thoughtfully structured, focusing on the musician in
                    the foreground while the slightly blurred audience creates a sense of depth
                    and context. The framing effectively draws the eye toward the central
                    subject while allowing the surrounding elements to add storytelling layers.
                    The balance between the performer and the audience creates a compelling
                    narrative.",
        "quality": "The quality of the photograph is impressive, with sharp details on the
                    musician and well-managed exposure that captures the intricate light
                    contrasts of the night scene. There is minimal noise, and the image retains
                    clarity, making it visually pleasing. The technical precision complements
                    the emotional and narrative depth, resulting in a polished and professional
                    image."
    },
.
.
.
]
```


Figure 2: Text Correspond with ID1.jpeg

3.2 MODELS

We explored four model variants off of the BLIP pre-trained model:

- Default Model without Fine-tuning
- Default Model with Fine-tuning
- More Visual Information Model with Fine-tuning
- New Architecture Model with Fine-tuning

3.2.1 DEFAULT MODEL WITHOUT FINE-TUNING

The default model without fine-tuning is the pre-trained BLIP base model. We used this as a baseline to compare our modified models to. This model follows what was discussed about in Li et al. (2022). Since the base model functions mainly as an image captioning model, the results from this model were rarely longer than a phrase or sentence.

3.2.2 DEFAULT MODEL WITH FINE-TUNING

The default model with fine-tuned is the pre-trained BLIP base model fine-tuned with our own dataset. We took the existing BLIP model, and trained it on our dataset, using supervised learning to adjust the parameters. Fine-tuning allows the model to adapt to the specific requirements of artistic image evaluation, such as understanding nuances in content, color, composition, and quality.

3.2.3 MORE VISUAL INFORMATION MODEL WITH FINE-TUNED

In this model, we have taken the previous model (the default model with fine-tuning), and provided more visual information to it. We resized the images from 224 x 224 to 384 x 384 to capture more information in the image encoder. Then, we trained the model on our custom dataset. The larger image-resize shape enables the model to capture more detailed and more minute visual information.

3.2.4 NEW ARCHITECTURE MODEL WITH FINE TUNED

New architecture model with fine-tuned is the pre-trained BLIP base model with larger image-resize shape (from 224 x 224 to 384 x 384), but we designed a linear layer for each of the four different image evaluation criterion and connect these four parallel linear layers with the last hidden layer in the model. Therefore, each of the image evaluation criterion, which are content, color, composition, and quality, will be trained differently on the last layer and has their own different weights.

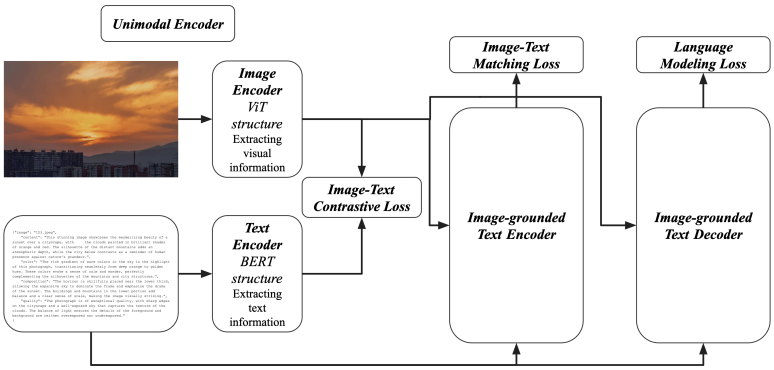

Figure 3: The Architecture of BLIP

3.3 TRAINING STRATEGY

We used the AdamW optimizer introduced by Loshchilov & Hutter (2019) for training our models. AdamW is a variant of the Adam optimizer that incorporates decoupled weight decay, which improves generalization by applying weight decay directly to the weights rather than through the gradient updates. This approach helps mitigate overfitting and ensures more stable convergence, particularly for models with a large number of parameters.

The AdamW optimizer updates weights using the following equations:

$$
\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\
\hat{m}_t &= \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \\
\theta_t &= \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \eta \lambda \theta_{t-1}
\end{aligned}
$$

Here: - g_t represents the gradient at time t, - m_t and v_t are the first and second moment estimates, - β_1 and β_2 are hyperparameters controlling the decay rates for these estimates, - $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected moment estimates, - η is the learning rate, - λ is the weight decay coefficient, - ϵ is a small constant for numerical stability, and - θ_t represents the weights at time t.

In our training, we utilized a dynamic learning rate strategy. The learning rate started at 1×10^{-5} and decayed by a factor of 0.5 every 5 epochs. This decay allows the optimizer to make larger updates during the early stages of training and finer adjustments as the model converges.

Additionally, we implemented early stopping to prevent unnecessary computations and over-fitting. If the training loss did not decrease by 1×10^{-3} for 5 consecutive epochs, the training was automatically terminated. This strategy ensures efficient training while maintaining model performance.

3.4 EVALUATION

To assess the performance of our models, we conducted both subjective and objective evaluations, acknowledging the inherent challenges in evaluating artistic image feedback.

3.4.1 SUBJECTIVE EVALUATION

The subjective evaluation was mainly based on human judgment, where we assessed the quality of the feedback generated by the models. We then made direct and high-level changes to the model to achieve better results. While human evaluation provides valuable insights, it is heavily influenced by individual biases, especially in subjective fields such as art. We encountered great difficulty in finding or creating a definitive evaluation metric for this task.

3.4.2 OBJECTIVE EVALUATION

For objective evaluation, we used automated metrics including BLEU and ROUGE scores Lin (2004). These metrics measure the overlap between the model-generated feedback and a set of reference texts, providing a quantitative assessment of textual similarity. BLEU evaluates n-gram precision, while ROUGE focuses on recall and overlap of key phrases. However, these metrics have limitations when applied to the nuanced nature of artistic evaluation, as they are primarily designed for tasks like machine translation and summarization. They struggle to capture deeper semantic similarities or contextual relevance in artistic feedback.

3.4.3 CHALLENGES IN EVALUATION

One of the key challenges in this study was designing an appropriate evaluation framework. Artistic feedback often involves subjective interpretation and contextual understanding that many current automated metrics fail to capture accurately. Additionally, human evaluators may be influenced by personal preferences or biases, further complicating the evaluation process. Developing more robust and nuanced evaluation methods remains an open problem for future work.

4 EXPERIMENTS

We conducted experiments on the models listed prior according to the training strategy mentioned in 3.3. We performed fine-tuning on the default model, the model with more visual information, and the model introducing a new architecture. Since we use the technique of early termination, all the models will automatically terminate training to prevent the issue of over-fitting. The results below show their performance:

Figure 4: Results of Default Model without Fine-tuning

content : this delightful photograph features a small bird perched on a rock, its delicate posture juxtaposed against a lush green leaf. the intimate focus on the bird ' s environment and its delicate gaze create a sense of connection with nature, making it both engaging and endearing.
color : the image features a natural and muted hues of the bird ' s orange against the earthy tones of the tree leaf. the soft glow of the light of the bird ' s head adds a subtle yet eye, enhancing the overall appeal of the scene.
composition : the composition places the bird on a slightly off - center, creating a dynamic composition that draws the viewer ' s attention to its environment. the surrounding elements create a natural and harmonious atmosphere, enhancing the sense of connection between the bird and its environment.
quality : the quality of the image is excellent, with clear details in the bird ' s fur and the surrounding environment. the lighting is soft and even, highlighting the delicate glow of the bird ' s feathers while the surrounding environment adds a touch of tranquility..

Figure 5: Results of Default Model with Fine-tuning

content : this image captures a tranquil forest setting, focusing on small mushrooms amidst a forested environment. the juxtaposition of natural life and man - made structures evokes a sense of coexistence and human development, inviting the viewer to reflect a quiet coexistence of nature.
color : the warm orange of the mushrooms against the dark tones of the forest floor and the muted greens of the surrounding trees. the subtle glow of the mushrooms adds a fresh and natural color palette, enhancing the overall visual.
composition : the mushrooms are positioned slightly off - center, drawing the viewer ' s eye into the scene while allowing the surrounding elements to frame the scene.
quality : the quality of the image is excellent, with sharp details that highlight the textures of the mushrooms and leaves. the lighting is natural and even, ensuring the mushrooms are well - balanced and well - balanced, making the scene visually captivating..

Figure 6: Results of More Visual Information Model with Fine-tuning

content : this image photograph a a an ofling a a the the by perched natural of a lush sense. the, of the and backdrop of an traditional ju the textures into -. the,, moment of, and the its sense draws the water beauty.
color : the and features of tones of image dominate golden the the complement vibrant ' and green of the and hue blurred, muted of complement the of. muted enhance. enhance the in of a adds that lighting the balance vi atmosphere overall depth. enhance the atmosphere
composition : the composition placement balancedlys cat,, drawing the a viewer, diagonal and the to framing subject adding. background its drawing elements to fore the creates reflection sense the to the sense from point thats engaging balance. and activity toward the both
quality : the quality of is sharp, high, impressive, intricate on cat textures s and sharply and in slightly the fur the s ensuring - on balanced well colors is ofs the even fields leaves subject the resulting professional the the a scene to to

Figure 7: Results of New Architecture Model with Fine-tuning

BLEU Score	Default Model with Fine-tuning	More Visual Information Model with Fine-tuning	New Architecture Model with Fine-tuning
1-gram	0.1172	0.1171	0.3558
2-gram	0.0308	0.0331	0.0378
3-gram	0.0099	0.0087	0.0009
4-gram	0.0032	0.0026	0.0000

Figure 8: BLEU Score for Models

Default Model with Fine-tuning

	r	p	f
rouge-1	0.900	0.920	0..909
rouge-2	0.692	0.757	0.719
rouge-1	0.856	0.876	0.865

More Visual Information Model with Fine-tuning

	r	p	f
rouge-1	0.931	0.844	0.883
rouge-2	0.789	0.476	0.584
rouge-1	0.891	0.806	0.844

New Architecture Model with Fine-tuning

	r	p	f
rouge-1	0.875	0.897	0.885
rouge-2	0.555	0.737	0.632
rouge-1	0.823	0.844	0.832

Figure 9: ROUGE Score for Models

Based on the human evaluation, we can see that between the default model with fine-tuning, the model with more visual information and fine-tuning, and the new architecture model with fine-tuning, the model with more visual information and fine-tuning performs the best. It is able to accurately capture and describe information in the photo in great detail, and generate evaluation in natural language. The default model with fine-tuning is also able to generate evaluation in human language, however, it has errors in recognizing the content in the photo, such as misinterpreting the mushroom as a bird. The model with our proposed architecture and fine-tuning performs the worst. It is unable to generate evaluation in human language and is unable to correctly recognize the content in the photo. We believe that this is because after adding the layers, the computational complexity of the model increases and makes it more difficult for the weights of the layers further away from the input to change. Therefore, the performance of the model becomes worse.

Based on numerical evaluation, BLEU score Papineni et al. (2002), we can see that the new architecture model with fine-tuning performs the best in 1-gram and 2-gram, but performs the worst in 3-gram and 4-gram. In contrast, the default model with fine-tuning and more visual information model with fine-tuning performs moderately well in 1-gram and 2-gram, but performs better in 3-gram and 4-gram.

By combining human evaluation and numerical evaluation Papineni et al. (2002) Lin (2004), we believe that the model with more visual information and fine-tuning performs the best in the task of generating artistic photo evaluation.

5 Conclusion

This study demonstrates the potential of multi-modal AI models in providing tailored and constructive feedback for visual artists. By fine-tuning the BLIP framework with a custom dataset and introducing architectural innovations, we have demonstrated the capability of such models to evaluate artistic images as a whole, while focusing on more discrete details such as content, color, composition, and quality. Our experiments have revealed that increasing the size of image inputs to give the model more information significantly improved the model's performance in generating detailed and relevant feedback.

While promising, this work also highlights several challenges. These include dataset biases, the difficulty in constructing an accurate evaluation metric, and the need for broader and more diverse training data to enhance the model's generalizability. Addressing these limitations in future research can further improve the gap between AI-driven tools and human sentiment, while also addressing the needs of the artistic community. Since many artists are concerned about their livelihoods due to the advancements in AI, we hope that this study demonstrates how AI can also be used as a tool to assist artists.

Ultimately, our approach demonstrates the transformative potential of leveraging AI in creative processes, assisting artists by providing them with accessible and intelligent tools to refine their craft. By enabling independent artistic growth and fostering innovation, multi-modal models like ours pave the way for broader applications across creative and professional domains.

REFERENCES

Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts, 2023. URL https://arxiv.org/abs/2305.10799.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends, 2022. URL https://arxiv.org/abs/2210.09263.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022. URL https://arxiv.org/abs/2201.12086.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.