Assessing Image-Captioning Models: A Novel Framework Integrating Statistical Analysis and Metric Patterns

Qiaomu Li, Ying Xie, Nina Grundlingh, Varsha Rani Chawan, Cody Wang
Kennesaw State University, The Home Depot
Marietta, GA; Smyrna, GA
{qli12, ngrundli}@students.kennesaw.edu, yxie2@ kennesaw.edu
{varsha_rani_chawan, cody_wang1}@homedepot.com

Abstract
In this study, we present a novel evaluation framework for image-captioning models that integrate statistical analysis with common evaluation metrics, utilizing two popular datasets, FashionGen and Amazon, with contrasting dataset variation to evaluate four models: Video-LLaVa, BLIP, CoCa and ViT-GPT2. Our approach not only reveals the comparative strengths of models, offering insights into their adaptability and applicability in real-world scenarios but also contributes to the field by providing a comprehensive evaluation method that considers both statistical significance and practical relevance to guide the selection of models for specific applications. Specifically, we propose Rank Score as a new evaluation metric that is designed for e-commerce image search applications and employ CLIP Score to quantify dataset variation to offer a holistic view of model performance.

Keywords: image-captioning model, image-based search, evaluation metric

1. Introduction

Image-captioning, the process of generating descriptive textual summaries for visual content, has emerged as an important AI capability with applications such as providing context for visually impaired users, automated alt-text generation, and enhanced image search. However, rigorously evaluating image-captioning models remains challenging. Traditionally, evaluating these models has involved using several evaluation metrics to score their performance, followed by comparing which model leads in these metrics to determine superiority.

Besides, as is known to all, each evaluation metric has its own focus, only a model that matches an evaluation metric's preference could get high scores. But it's increasingly clear that no single model could consistently outperform others across all metrics and datasets. This divergence highlights the challenge in declaring one model definitively superior to others. On the other hand, now that a single model cannot win all, we could only identify each model's comparative strengths when evaluating several models simultaneously, acknowledging each one's pros and cons. This full understanding is essential, as different application scenarios may require different strengths, making it imperative to match a model from several potential one to the situation where it's comparatively more suitable.

Furthermore, the choice of benchmark dataset influences evaluation outcomes. To take advantage of model potential, we generally need to finetune or train models on the target dataset. However, in practical situations, image-caption alignment can diverge across datasets, influencing final results. We account for this by employing CLIP Score to quantify dataset variation, that is, the overall alignment between images and corresponding captions within the dataset. By measuring dataset variation, we gain insights into dataset complexity and noise levels. This allows us to deduce model ability based on dataset qualities, and then figure out comparative model strengths.

In this paper, we propose an integrated evaluation framework that combines the statistical analysis of various metrics to identify models with comparative strengths. By analyzing statistical significance across metric results, we reveal relative advantages of models and suitable applications based on evaluation metric patterns. We summarize our primary contributions as follows:

- We utilize CLIP Score to assess dataset variation in overall image-caption alignment, providing a basis for model evaluation.
- We come up with a novel evaluation framework that merges statistical significance with reverse reasoning from metric patterns, extracting comparative model strengths.
- We introduce a novel Rank Score metric, a simple yet powerful metric to evaluate image-captioning models by assessing generated text quality through comparative ranking against reference captions.

2. Related Work

Recent years have seen diverse image-captioning models developed based on generative techniques. Representative examples include Video-LLaVa (Lin et al., 2022), BLIP (Li et al., 2021), CoCa (Yao et al., 2021), and ViT-GPT2 (Kumar, 2022). Among these, Video-LLaVa

The Seventh Workshop on e-Commerce and NLP (ECNLP 7), pages 79–87

extends language models to video for dynamic content understanding, BLIP uses bootstrapped pre-training for improved visual-language synergy, CoCa utilizes cross-modal contrastive learning to enhance image understanding and caption generation, and ViT-GPT2 combines Vision Transformer (ViT) (Dosovitskiy et al., 2021) and GPT-2(Radford, Alec, et al., 2019) for efficient image and text processing. While adopting different approaches, rigorous comparative evaluation is needed to reveal the comparative strengths of these models to match them to suitable applications.

A range of automated metrics have been proposed for evaluating image-captioning models by comparing candidate captions to references. Popular metrics include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), (Anderson et al., 2016), and METEOR (Banerjee & Lavie, 2005). These measures rely on lexical, grammatical and semantic similarities. CLIP Score (Hessel et al., 2021) supplements these by comparing captions directly to images. However, no single model consistently outperforms across all metrics.

Our hypothesis is that different models may exhibit comparative strengths aligning with particular use cases based on precision, efficiency, adaptability etc. This work provides a comprehensive framework combining statistical analysis and tailored datasets to reveal model capabilities, guiding selection for applications using different metrics.

3. Methodology

The methodology involves using CLIP Score to quantify dataset variation based on the alignment between images and captions of each data set. For model evaluation, a set of metrics including BERT Score, BART Score, METEOR, SPICE, BLEU, CLIP Score, and the proposed Rank Score are applied on the output of each image-captioning model on each data set. Statistical analysis using paired t-tests with Bonferroni correction is then conducted on the evaluation results to identify models with comparative strengths based on statistically significant performance. The preferences of metrics are analyzed to infer model capabilities. By combining statistical significance with reasoning from evaluation patterns, the framework identifies specialized strengths of models and their suitable applications.

3.1 Evaluation Metrics and Preferences

3.1.1 BLEU (Bilingual Evaluation Understudy)

BLEU measures the precision of n-grams in the generated text compared to the reference texts, adjusting for the proper length and penalizing overly short translations (Papineni et al., 2002). It does this by calculating the n-gram precision for several n-gram lengths (usually 1 to 4) and then combining these precisions geometrically, applying a brevity penalty for translations that are too short.

The BLEU is calculated as:

- *N-gram Precision (P_n): For each n-gram length (n=1 to 4), calculate the count of n-grams in the candidate translation that appear in any reference translation, divided by the count of all n-grams in the candidate translation.*
- *Brevity Penalty (BP): To penalize short machine-generated translations, a brevity penalty is applied. If the length of the candidate translation is less than the effective reference corpus length, the brevity penalty applies:*

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases}$$

$$c: Length\ of\ the\ candidate\ translation$$

$$r: Length\ of\ the\ effective\ reference\ corpus$$

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

$$w_n: weights\ for\ each\ n-gram\ precision$$

$$P_n: the\ precision\ for\ n-grams$$

$$N: the\ maximum\ n-gram\ length$$

Preferences: BLEU calculates the score by matching the n-grams of the candidate text with the reference texts and applying a brevity penalty for short candidate texts. This method, focusing on surface lexical matches without considering semantic context or synonyms, inherently favors models that are excellent at producing exact n-gram matches with the reference texts.

3.1.2 METEOR (Metric for Evaluation of Translation with Explicit Ordering)

Meteor is a machine translation evaluation metric, which is calculated based on the harmonic mean of precision and recall, with recall weighted more than precision (Banerjee & Lavie, 2005).

The METEOR is calculated as:

F-Score: The harmonic mean of Precision and Recall, given more importance to recall. It's calculated as:

$$F_{score} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P}$$

Penalty: A penalty is applied for poor word order, computed based on the largest common subsequence of matched words between the candidate and the reference:

$$Penalty = 0.5 \cdot \left(\frac{number\ of\ chunks}{number\ of\ unigrams\ matched}\right)^3$$

$$METEOR = F_{score} \cdot (1 - Penalty)$$

Preferences: METEOR assesses translations by accounting for exact word matches, synonyms,

stemming, and word order, calculating a harmonic mean of precision and recall adjusted for these factors. This approach indicates a preference for models that understand and utilize linguistic nuances, including synonymy and grammatical structure.

3.1.3 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE is a set of metrics used for evaluating automatic summarization and machine translation software in natural language processing (Lin, 2004). It works by comparing an automatically produced summary or translation against a reference or a set of reference summaries (typically human-produced). It includes several variants, such as ROUGE-N, ROUGE-L, and ROUGE-W, each focusing on different aspects of the comparison:

ROUGE-N measures the overlap of n-grams between the system-generated text and the reference texts. It is defined as:

$$ROUGE - N = \frac{\sum_{s \in \{Reference\ Summaries\}} \sum_{gram_n \in s} Count_{match}(gram_n)}{\sum_{s \in \{Reference\ Summaries\}} \sum_{gram_n \in s} Count(gram_n)}$$

$Count_{match}(gram_n)$: *max number of n-grams co-occurring in a candidate summary and and reference summary.*

$Count(gram_n)$: *the count of n-grams in the reference summaries.*

ROUGE-L measures the longest common subsequence (LCS) between the system-generated summary and the reference summaries. It considers sentence-level structure similarity naturally and identifies the longest co-occurring in-sequence n-grams of words. ROUGE-L is defined as:

$$ROUGE - L = \frac{(1 + \beta^2) \cdot Recall_{LCS} \cdot Precision_{LCS}}{Recall_{LCS} + \beta^2 \cdot Precision_{LCS}}$$

Where:

$$Recall_{LCS} = \frac{LCS(System\ Summary, Reference\ Summary)}{Length(Reference\ Summary)}$$

$$Precision_{LCS} = \frac{LCS(System\ Summary, Reference\ Summary)}{Length(System\ Summary)}$$

$\beta\ is\ the\ set\ to\ favor\ recall (e.g., \beta^2 = 1.2), because\ recall\ is\ more\ important.$

ROUGE-W is based on the weighted longest common subsequence, which considers the length of the LCS and the gaps between consecutive LCS matches. It is defined as:

$$ROUGE - W = \frac{(1 + \beta^2) \cdot Recall_{WLCS} \cdot Precision_{WLCS}}{Recall_{WLCS} + \beta^2 \cdot Precision_{WLCS}}$$

Where:

$Recall_{WLCS}$: $weighted\ Recall_{LCS}$

$Precision_{WLCS}$: $weighted\ Precision_{WLCS}$

Preferences: ROUGE measures the overlap of n-grams and the longest common subsequences between the generated text and reference texts, primarily focusing on recall. This metric's emphasis on recall over precision suggests a preference for models that ensure no information is lost in summarization, even if it leads to less concise outputs. Therefore, models that excel under ROUGE are those capable of capturing the breadth of content in reference texts, making them particularly suitable for summarization tasks where the completeness and coverage of the source material are paramount, rather than stylistic conciseness or linguistic innovation.

3.1.4 CIDEr (Consensus-based Image Description Evaluation)

CIDEr is a metric used to evaluate the quality of generated textual descriptions of images (Vedantam et al., 2015). It measures the similarity between a generated caption and the reference captions.

The CIDEr is calculated as:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_n \cdot sim_n(c_i, S_i)$$

Where:

c_i: $The\ candidate\ caption\ for\ image\ i.$

S_i: $A\ set\ of\ reference\ captions\ for\ image\ i$

N: $Max\ n-gram\ length.$

w_n: $The\ weight\ of\ each\ n-gram\ length.$

sim_n: *The cosine similarity between the tf-idf weighted n-gram vectors of the candidate caption and the reference captions.*

Preferences: CIDEr evaluates image-captioning quality by calculating the consensus between a candidate caption and reference captions using term frequency-inverse document frequency (TF-IDF) weighting for n-grams. This approach emphasizes the importance of unique and descriptive terms that are relevant to the image, favoring models that generate detailed and image-specific captions.

3.1.5 SPICE (Semantic Propositional Image Caption Evaluation)

SPICE is an automated caption evaluation metric that uses scene graphs to measure the semantic similarity of reference and candidate captions (Anderson et al., 2016). The SPICE score is calculated as the F1 score between the sets of tuples extracted from the candidate and reference captions' scene graphs. The tuples represent objects, attributes, and relations.

The SPICE is calculated as:

$$SPICE = 2 \cdot \frac{P \cdot R}{P + R}$$

P: $Proportion\ of\ tuples\ in\ the\ candidate\ caption\ that\ also\ appear\ in\ the\ reference\ captions.$

R: $Proportion\ of\ tuples\ in\ reference\ captions\ that\ are\ captured\ in\ the\ candidate\ caption.$

Preferences: SPICE evaluates image captions based on semantic fidelity, comparing structured scene graphs derived from candidate and reference captions to assess the presence and accuracy of depicted objects, attributes, and relationships. This focus on semantic content leads SPICE to favor models adept at deep visual understanding and generating captions that accurately reflect complex visual scenes in natural language.

3.1.6 BERT Score

BERT Score evaluates the quality of text by calculating the cosine similarity between the BERT embeddings (Devlin et al., 2019) of the candidate text and the reference text (Zhang et al., 2020).

The BERT Score is calculated as:

$$BERT\ Score = 2 \cdot \frac{P \cdot R}{P + R}$$

$$P(Precision) = \frac{1}{|C|}\Sigma_{c \in C} max_{r \in R} cosine(c, r)$$

$$R(Recall) = \frac{1}{|R|}\Sigma_{r \in R} max_{c \in C} cosine(r, c)$$

$$C: Cadidate\ text; \quad R: Reference\ text$$

Preferences: BERT Score leverages the contextual embeddings from BERT model to compare the semantic similarity between candidate and reference texts, focusing on the match at a deeper semantic level rather than surface lexical similarity. This metric's design prefers models that generate contextually rich and semantically accurate text, reflecting a deep understanding of language nuances.

3.1.7 BART Score

BART Score, which stands for BLEU Artifact Reduction Test score, is a metric for evaluating the quality of text generation models (Yuan et al., 2021). It focuses on measuring how well a model's generated text preserves the factual content and overall meaning of a reference text.

The BART Score is calculated as:

$$BART\ Score = w_1 \cdot BLEU\text{-}4 + w_2 \cdot ROUGE\text{-}L + w_3 \cdot BERT\ Embedding\ Similarity + w_4 \cdot Factual\ Consistency + w_5 \cdot Informativeness;$$

$w_1 \sim w_5$: $weighted\ score\ to\ each\ component,\ generally\ take\ the\ average.$

BLEU-4: *This component measures n-gram overlap between the generated and reference texts, similar to traditional BLEU score.*

ROUGE-L: *This part assesses the longest matching subsequences between the texts, capturing more meaningful phrases.*

BERT Embedding Similarity: *This measures the semantic similarity between the generated and reference texts using pre-trained language models like BERT.*

Factual Consistency: This component analyzes the factual inconsistencies between the texts, ensuring generated information aligns with the reference.

Informativeness: This portion gauges the level of new information added by the generated text compared to the reference.

Preferences: BART Score leverage BART's architecture to assess coherence, fluency, and contextual relevance of generated text against reference texts. It would naturally prefer models that are adept at producing text that is contextually relevant, coherent across longer passages, and syntactically fluent.

3.1.8 BLEURT (Bilingual Evaluation Understudy with Representations from Transformers)

BLEURT is an evaluation metric for Natural Language Generation (Sellam et al., 2020). It takes a pair of sentences as input, a reference and a candidate, and it returns a score that indicates to what extent the candidate is fluent and conveys the meaning of the reference.

Calculation procedure:

1. ***Feature Extraction****: The model takes a pair of sentences (a reference and a candidate) as input and passes them through a BERT model to obtain their embeddings. These embeddings are high-dimensional feature vectors that capture the semantic information of the sentences.*
2. ***Regression Model****: The regression model compares the embeddings of the reference and candidate sentences to calculate a similarity score. This comparison is done using a linear layer that predicts the similarity between the feature vectors of the original sentence and its re-translation.*
3. ***Training****: The regression model is trained on a dataset of human ratings. The training process involves adjusting the parameters of the model to minimize the difference between the model's predictions and the actual human ratings.*
4. ***Output****: The model returns a score that indicates to what extent the candidate is fluent and conveys the meaning of the reference.*

Preferences: BLEURT scores texts by leveraging a BERT-based model fine-tuned on human judgment data, evaluating semantic similarity and naturalness of language. This mechanism inclines BLEURT to prefer models that produce text closely mirroring human writing styles and semantic richness, capturing nuances in meaning and context.

3.1.9 CLIP Score

CLIP Score is a reference free metric that can be used to evaluate the correlation between a generated caption for an image and the actual content of the image. It has been found to be highly correlated with human judgement (Hessel et al., 2021).

The CLIP Score is calculated as:

$$CLIP\ Score(I, C) = cosine\ (E_I, E_C)$$
$$E_I: embedding\ for\ image\ I\ by\ CLIP\ Model$$
$$E_C: embedding\ for\ caption\ C\ by\ CLIP\ Model$$

Preferences: CLIP Score uses the cosine similarity between the embeddings of images and corresponding textual descriptions generated by the CLIP model, this score would measure how well the text describes the image, considering both semantic content and visual details. Given this approach, CLIP Score would favor models that excel in generating accurate, detailed, and semantically rich descriptions of images. These models are able to understand and interpret complex visual scenes and translate this understanding into coherent, contextually relevant text.

3.1.10 Rank Score

Rank Score is designed to evaluate image-captioning models for the application of image-based product search. Given a query product image, the model generates a text caption. This caption is then used to retrieve relevant products textually, and the rank of the original queried product in the results list is used to calculate the Rank Score.

Specifically, the candidate caption for the query image is compared against product captions in the dataset via BERT embeddings to retrieve a ranked list of products by similarity. If the original product image ranks 1st, the Rank Score is 1, indicating the highest performance in returning the queried product. If the product ranks last, the Rank Score is close to 0. Other ranks will have values between 0 and 1, with higher values indicating better performance in retrieving the original product.

The Rank Score is calculated as:

$$Rank\ Score = 1 - \frac{True\ Rank - 1}{N}$$
$$N: Total\ num\ of\ original\ texts\ in\ the\ dataset$$
$$True\ Rank: the\ position\ of\ the\ true\ original\ text\ in\ the\ list\ sorted\ by\ descending\ similarity\ to\ the\ generated\ text.$$

Preferences*:* This approach aims to quantify how well the generated caption captures the visual essence of the query image in a way that enables accurate text-based retrieval of the original product. The metric favors models that produce captions semantically aligned with the visual content to support precise image search.

3.2 Dataset Variation with CLIP Score

In evaluating image-captioning models, the quality of benchmark datasets greatly influences the outcomes. To address this, we employ CLIP Score (Hessel et al., 2021) to measure the congruence between images and captions, offering a method to assess the dataset's overall alignment, shown in Figure 1. By calculating the standard deviation of these scores for each dataset, we can get comparative variations of datasets. A dataset with lower standard deviation indicates less variation, and higher standard deviation points to greater variation. In general, a model performs better in a dataset with less dataset variation is inclined to have more precision and efficiency; and a model performs better in dataset with more dataset variation is inclined to have more robustness and adaptability. We could extract more model patterns and suitable applications based on comparative dataset variations and evaluation metric patterns, shown in TABLE I.

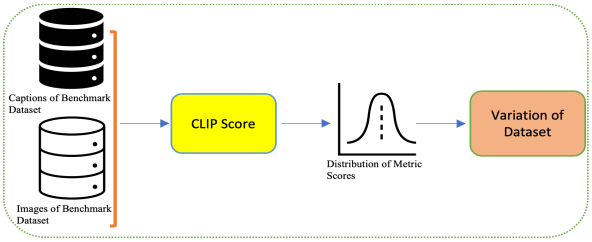


Figure 1: Check dataset variation with CLIP Score

Metrics	*Less Variation*	*More Variation*
BLEU	**Strength:** High lexical matching precision. **Applications:** Formal document translation, legal document replication.	**Strength:** Adaptability to lexical diversity. **Applications:** Multilingual social media content translation, diverse genre text localization.
ROUGE	**Strength:** Detailed content coverage. **Applications:** Executive summary generation for business reports, focused news article summarization.	**Strength:** Flexibility in content extraction. **Applications:** Summarizing user-generated content, variable style news aggregation.
METEOR	**Strength:** Precision in detailed captions. **Applications:** Ideal for archival systems where accuracy is paramount.	**Strength:** Versatility in language adaptation. **Applications:** Suited for dynamic content such as diverse social media platforms.
SPICE	**Strength:** In-depth scene analysis. **Applications:** Suitable for educational tools requiring detailed image explanations.	**Strength:** Recognition of features in complex visuals. **Applications:** E-commerce platforms, highlighting product features amidst visual clutter.
BLEURT	**Strength:** Nuanced tone and language detection. **Applications:** Luxury branding where subtlety in captions can influence perception.	**Strength:** Adaptability to a range of linguistic styles. **Applications:** User-generated content platforms needing accurate captions for diverse submissions.
BERT Score	**Strength:** Deep semantic alignment with reference texts. **Applications:** Content-centric websites that require aligned thematic narratives.	**Strength:** Robust contextual understanding. **Applications:** News and information sites with varied topical content.
CIDEr	**Strength**: Precision in detail-oriented image description. **Applications**: Cataloging for digital archives, precise product descriptions for e-commerce.	**Strength**: Ability to highlight unique image features. **Applications**: Dynamic caption generation for social media platforms, interactive educational content.
BART Score	**Strength:** Mastery in generating coherent, contextually relevant text. **Applications:** Narrative-driven media, adding depth to visual stories.	**Strength:** Flexibility in text generation across diverse styles and formats. **Applications:** Creative writing tools, adaptive marketing content generation across various platforms.
CLIP Score	**Strength:** Precise visual-text alignment. **Applications:** Art galleries or databases requiring accurate image cataloging.	**Strength:** Creativity and adaptability in describing diverse visual content. **Applications:** Social media content generation and enhancement.
Rank Score	**Strength:** Precision in semantic alignment with target texts. **Applications:** Customized news feed generation, precise document retrieval in legal and academic research databases.	**Strength:** Adaptability in understanding and matching a wide range of semantic contexts. **Applications:** Chatbots and virtual assistants tailored to diverse user queries.

TABLE 1: Model's Comparative Strength and Suitable Applications under Each Evaluation Metric

3.3 Comparative Evaluation with Statistical Analysis

Our method employs t-tests with Bonferroni correction on model results across various evaluation metrics. This determines if there is a model that achieves higher scores with statistical significance than others in a specific evaluation metric. If there exists such a model, we harness the specific preferences of that metric and corresponding benchmark dataset variation to infer the model's comparative strengths. The whole procedure is shown in Figure 2.

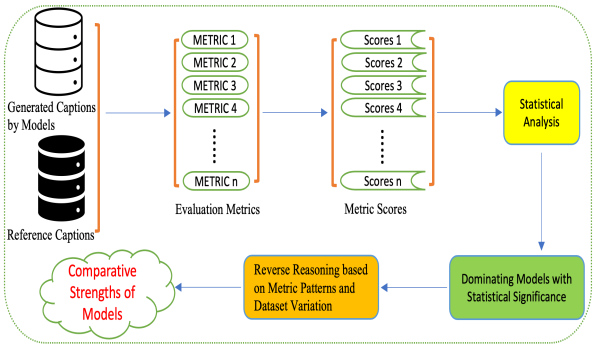


Figure 2: Procedure of Comparative Evaluation with Statistical Analysis

4. Experiments

We conduct experiments on two distinct datasets to validate the real-world efficacy of our novel evaluation framework for enhanced evaluation of image-captioning models.

4.1 Datasets

FashionGen (Rostamzadeh et al. 2022) - This dataset contains 360K apparel image-text pairs with high alignment. Typically, a product could have several different images but the same caption. This dataset features images of clothing items worn by models against clean backgrounds and consistent captions detailing visual attributes of clothing items. We randomly sample a subset of 50,000 rows for training and 500 rows for testing, which are mutually exclusive.

Amazon Product Dataset (Tools and Home Improvement) (Ni et al., 2019) - The home & kitchen product dataset of 51K rows from Amazon consists of seven subcategories, exhibiting greater noise - multiple images per product with different perspectives, backgrounds, and even sketch maps. Captions could contain peripheral details less related to corresponding products, such as product histories and usage scenarios not directly extractable from images alone. Our experiment uses 500 rows proportionally sampled across subcategories for testing, with the remainder forming the training dataset.

4.2 Evaluation Metrics

For our experiments, we specifically chose not to use BLEU, ROUGE, and CIDEr due to their limitations. BLEU, primarily focused on n-gram overlap, is adept at evaluating literal translations but falls short in assessing the contextual alignment in image-captioning. ROUGE, while valuable for summarization tasks, does not cater to our objective of generating descriptive captions to describe images. Lastly, CIDEr, despite its utility in comparing a generated caption against a set of references, does not suit our model's aim of producing a singular, optimal caption for each image.

4.3 Models

CoCa(**Contrastive Captioners**) (Yao et al. 2021) - a state-of-the-art model designed to excel in both image understanding and captioning tasks by leveraging contrastive learning. It combines the strengths of powerful visual encoders with language models to generate descriptive, accurate captions for images.

Video-LLaVa (Video Language-Large Model) (Lin et al. 2022) **-** extends the capabilities of language models to the domain of video understanding and captioning. By processing video inputs alongside textual descriptions, Video-LLaVa aims to capture the dynamic aspects of video content, translating them into coherent and comprehensive text.

BLIP (Bootstrapped Language Image Pre-training) (Li et al. 2021) - a model that emphasizes the pre-training phase to enhance the synergy between visual perception and language understanding. By bootstrapping from large-scale datasets, BLIP learns to generate captions that are not only accurate in depicting the visual content but also engaging and informative.

ViT-GPT2 - it combines the Vision Transformer (ViT) (Dosovitskiy et al. 2021) with the GPT-2 (Radford, Alec, et al. 2019) language model to create a hybrid system capable of processing images and generating corresponding captions. ViT extracts and processes visual information from images, transforming it into a format that the GPT-2 model can use to generate textual descriptions.

4.4 Experimental Procedure

To fully leverage each model's capabilities and provide a robust evaluation, we adopted distinct measures tailored to each model.

For CoCa, it's a structural model without prior training, we train it from scratch separately on both the FashionGen and Amazon training datasets.

For BLIP, to sharpen their domain knowledge and optimize performance for our specific datasets,

we finetune the pretrained model (*Salesforce/blip-image-captioning-large*) from HuggingFace, separately on both the FashionGen and Amazon training datasets and rename it Finetuned-BLIP.

For ViT-GPT2, same as BLIP, we finetune the pretrained model (*vit-gpt2-image-captioning*) from HuggingFace, separately on both the FashionGen and Amazon training datasets and rename it Finetuned-ViT.

For Video-LLaVa, it is a LLM with extensive pre-training. Given its big size and intricate internal structure, fine-tuning was not a viable option. Instead, we utilize prompting to direct it to generate appropriate captions. The respective prompts as shown below:

For FashionGen:

Create a caption for the fashion image that accurately describes the garment. Concentrate on the clothing item design, material, cut, patterns, and any distinctive features. Provide a clear and precise description that could serve as a product description, ignoring any other elements in the image.

For Amazon:

Generate a concise and accurate caption describing key details in product images from Amazon. Focus on correctly identifying the main product or object in the image frame and list only the most salient attributes and components. Avoid subjective impressions or editorial comments. Stick to factual, neutral descriptions of visual elements reflecting the product name, materials, parts, colors, text, or other relevant details a shopper would need to make a purchase decision.

Having generated candidate captions from these 4 models for both test datasets, we next apply our suite of evaluation metrics, including BERT Score, BART Score, METEOR, SPICE, BLEU, CLIP Score, Rank Score, to get results.

After getting results, we then conduct paired t-tests with Bonferroni correction separately on the evaluation results of these two testing datasets, shown in Figure 3 and Figure 4. We collect results for each metric to check whether there exists a significant leading model on an evaluation metric, the results are shown in TABLE 2 and TABLE 3, and the statistical significance criteria is (p-value < 0.01).

Given the divergence across different datasets, we use CLIP Score to evaluate the overall alignment of training datasets, shown in Figure 5. FashionGen shows an approximately normal distribution with a lower standard deviation, indicating less dataset variation. Conversely, the Amazon dataset displays a wider distribution with a higher standard deviation, reflecting more dataset variation.

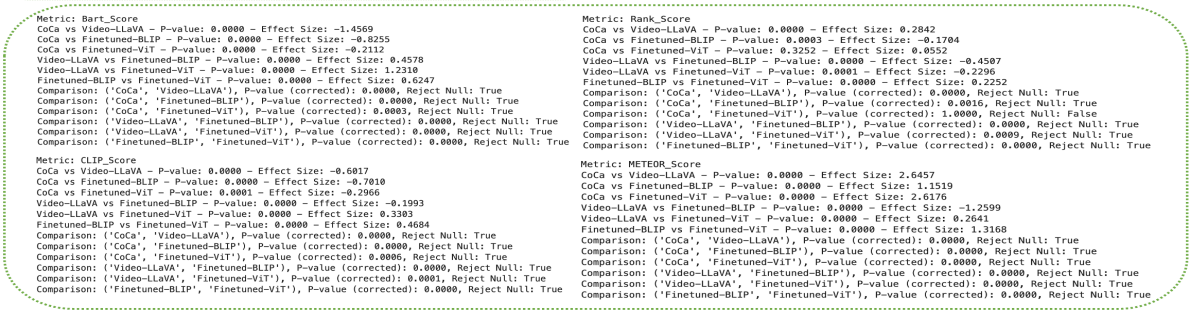


Figure 3: Results of Statistical Analysis with Statistical Significance (FashionGen)

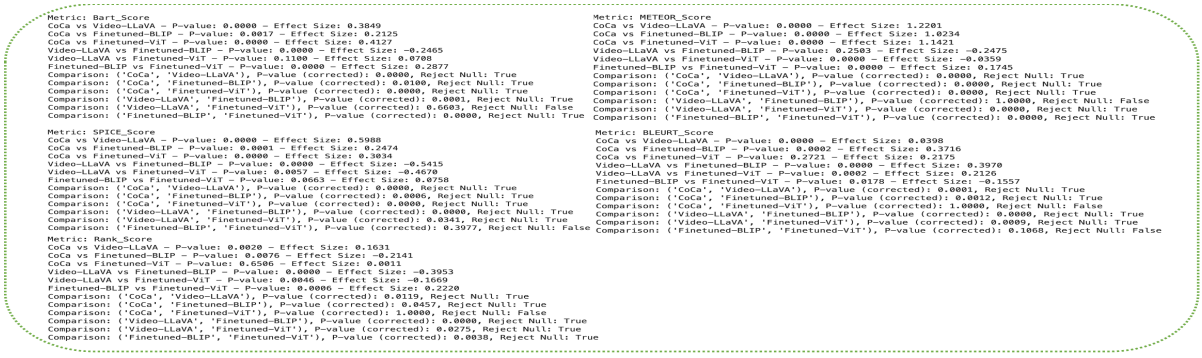


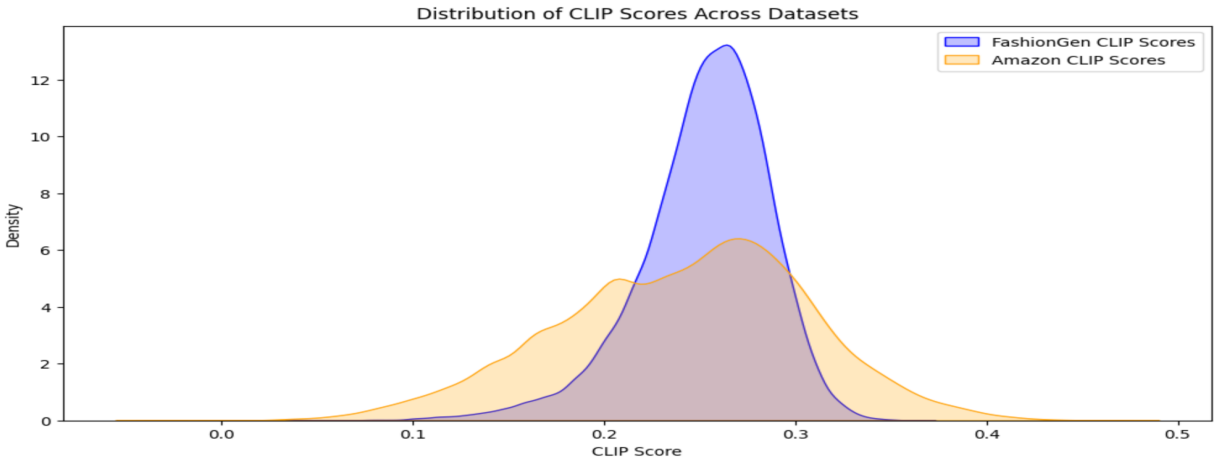
Figure 4: Results of Statistical Analysis with Statistical Significance (Amazon)

Evaluation Metric	*Significant Leading Model*	*Statistical Significance*
BART Score	Video-LLaVa	Yes
METEOR	CoCa	Yes
CLIP Score	Finetuned-ViT	Yes
Rank Score	Finetuned-BLIP	Yes

TABLE 2: Statistical Analysis OF Evaluation Metric Scores IN FashionGen Testing Dataset.

Evaluation Metric	*Significant Leading Model*	*Statistical Significance*
BART Score	CoCa	Yes
BLEURT	CoCa	Yes
SPICE	CoCa	Yes
METEOR	CoCa	Yes
Rank Score	Finetuned-BLIP	Yes

TABLE 3: Statistical Analysis OF Evaluation Metric Scores IN Amazon Testing Dataset




```
Mean CLIP Score (FashionGen): 0.25236592904051
Mean CLIP Score (Amazon): 0.24148941073665947
Standard Deviation (FashionGen): 0.034193109111506505
Standard Deviation (Amazon): 0.06530483782377532
```

Figure. 5: Distribution Plot of CLIP Scores across Two Benchmark Datasets

4.5 Experimental Results

Table 2 and Table 3 present significant leading models on metrics across the FashionGen and Amazon Product test datasets. Observed results include:

- Video-LLaVA demonstrates top performance on BART Score in FashionGen dataset while CoCa leads on BART Score in Amazon dataset.
- Fine-tuned ViT takes the lead on CLIP Score of FashionGen dataset.
- CoCa shows dominance in other metrics with statistical significance across datasets.
- Fine-tuned BLIP consistently beat others on Rank Score metric in both datasets.

4.6 Model Analysis on Evaluation Results

Analyze the experimental results and refer to TABLE 1, we get insights below:

CoCa: CoCa takes the lead in various metrics across both datasets, reflecting its comprehensive understanding of image-caption relationships. Its success could be attributed to its contrastive learning framework, enabling interpretation and generation of captions that are both contextually relevant and linguistically precise. From these results, we conclude that CoCa is exceptionally capable of adapting to diverse dataset qualities, making it a versatile choice for applications requiring detailed and accurate image captions, from content creation to archival description.

Video-LLaVa: Video-LLaVa's stands out on BART Score in FashionGen dataset, demonstrating its ability to generate coherent and contextually relevant text. This proficiency suggests that Video-LLaVa effectively interprets and translates visual content into meaningful captions, leveraging its advanced understanding of both visual elements and textual narrative. From this result, we conclude that Video-LLaVa is well-suited for enriching narrative-driven media and adding depth to visual stories, making it a valuable tool for applications aiming to provide engaging and enriched content narratives.

ViT-GPT2: ViT-GPT2 excels in the FashionGen dataset on the CLIP Score, showcasing its strength in precise visual-text alignment. This performance could be attributed to its use of Vision Transformer for visual analysis combined with GPT-2 for text generation, ensuring accurate and relevant captions for images. Consequently, we conclude that ViT-GPT2 could be effective for applications like art gallery databases or specialized image catalogs, where accurate and detailed image descriptions are crucial for user engagement and information retrieval.

BLIP: BLIP excels in the Rank Score metric across datasets, showcasing its precision and adaptability in semantic alignment. This performance indicates BLIP's robust capability for detailed semantic interpretation and flexible application across different content needs. Consequently, BLIP is ideal for creating precise, customized content in areas like news feeds and document retrieval, as well as for developing responsive chatbots and virtual assistants capable of handling a wide range of queries.

Our research, informed by experimental results and analysis to evaluation metrics, aimed to evaluate, not rank, various image-captioning models to discern their comparative strengths and suitable applications. This methodology highlights comparative advantages of each model, facilitating an informed selection for specific image-captioning needs. The findings offer a strategic framework for choosing models that best match the required competencies for targeted applications.

5. Conclusion

In our paper, we introduced a handy framework to evaluate image-captioning models. By conducting experiments on two datasets with contrasting variation in image-caption alignment, we demonstrated how our approaches can reveal the inherent strengths and practical applicability of different models. Our integration of statistical analysis with reverse reasoning on evaluation metrics, providing a comprehensive framework that not only assesses accuracy but also provides practical applications. The insights extracted from this procedure underscore the versatility of our evaluation framework in discerning the comparative capabilities of image-captioning models in varied contexts.

Future work: Future efforts could aim to expand the range of evaluation metrics for a deeper analysis of model capabilities. There is also significant potential in refining the training or finetuning procedures for the image-captioning models under study. Perfecting these models to their optimal performance is key to accurately harnessing their full potential in real-world applications. Besides, our method shall not be limited to image-captioning, once there are multiple evaluation metrics, we could always apply the framework, such as evaluations on video-captioning and text-to-image under e-commerce industry.

References

Kumar, A. (2022). The Illustrated Image Captioning using transformers. ankur3107.github.io.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4566-4575).

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

Banerjee, S., & Lavie, A. (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 (pp. 65-72).

Hessel, J., Holtzman, A., Forbes, M., Woolley, A., Rajani, N. F., Rocktäschel, T., ... & Lee, L. (2021). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv preprint arXiv:2102.08535.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (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) (pp. 4171-4186).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTscore: Evaluating text generation with BERT. In International Conference on Learning Representations.

Yuan, W., Neubig, G., & Liu, P. (2022). BARTScore: Evaluating Generated Text as Text Generation. arXiv preprint arXiv:2206.04793.

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 7881-7892).

Rostamzadeh, N., Hosseini, S.A., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C. et al. (2022). Fashion-Gen: The Generative Fashion Dataset and Challenge. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 525-534).

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., & Yuan, L. (2022). Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 801-810).

Li, X., Yin, X., Li, C., Hu, X., Yang, P., Zhang, L., Hu, H., Zhang, L., Wang, L., & Hu, B. (2021). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. European Conference on Computer Vision (ECCV).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. & Uszkoreit, J. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. 9th International Conference on Learning Representations (ICLR).

Radford, Alec, et al. Language models are unsupervised multitask learners. OpenAI blog 1.8 (2019): 9.

Jianmo Ni, Jiacheng Li, Julian McAuley. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. Empirical Methods in Natural Language Processing (EMNLP), 2019.