ImageIntegrity: Verifying Image-Caption Consistency

Atharva Bhide asbhide@usc.edu
Cheryl Khau ckhau@usc.edu
Jake Treska treska@usc.edu
Jeannie Jiang jeanniej@usc.edu
Rahul Sura rsura@usc.edu
Linghao Jin (TA) linghao@usc.edu

Abstract

The rise of multimodal AI has introduced new challenges in aligning images with text, especially in areas like misinformation detection, accessibility, and content validation. While methods like TIFA have improved object presence evaluation, they often struggle with explainability and finer-grained inconsistencies such as attributes or actions. We present ImageIntegrity, a modular framework for evaluating image-caption consistency without relying on human-annotated ground truth. Our initial design used BLIP-based captioning, object detection, and question-answering via TIFA, but was refined into a streamlined pipeline focused on CLIP-based semantic alignment for efficient and interpretable evaluation. Validated on MS-COCO, our system shows that model-generated captions align more closely with images than user-written ones, highlighting CLIP's effectiveness for real-world, reference-free caption evaluation.

1 Introduction

As text-to-image models like Midjourney, DALL·E 3, and Stable Diffusion become ubiquitous—in fields from education and journalism to marketing and accessibility—ensuring that their outputs faithfully reflect the input prompts is critical. Even small misalignments can mislead readers, propagate bias, or undermine trust in applications that depend on precise visual storytelling. Much of the existing work on alignment relies on surface-level checks such as cosine similarity in embedding space or simple n-gram overlap. These methods often miss subtler inconsistencies, such as omitted attributes (e.g., color or size), incorrect object counts, or mismatched actions (e.g., "running" versus "standing"). Our original proposal explored this challenge through a multi-metric evaluation framework—focusing on correctness, completeness, and fluency—along with supporting tools like buzzword-based scoring and visual Q/A validation. Building on this foundation, our project evolved to introduce ImageIntegrity, a streamlined and scalable system for evaluating image-caption alignment. Through empirical testing, we refined our pipeline to rely on CLIP-based semantic similarity as the primary measure of caption fidelity. This design enables effective evaluation without dependence on human-written ground truth, making the system suitable for real-world deployment and extensible to future captioning models.

2 Related Work

2.1 TIFA

TIFA (Hu et al., 2023) breaks a prompt into fine-grained Q/A pairs, feeds them to a VQA model over the generated image, and computes a consistency score from answer accuracy. It outperforms simple embedding methods at capturing detailed misalignments but relies on older GPT back-ends, struggles with abstract or anime-style inputs, and is considerably slower than CLIP-based checks.

2.2 ContextCLIP

ContextCLIP (Grover et al., 2022) augments the CLIP objective with a contextual loss that aligns image regions to individual words. Trained on Conceptual Captions, it yields more structured embeddings, surpasses CLIP in text–image retrieval, and shows stronger zero-shot transfer on datasets where CLIP suffers negative transfer.

2.3 OPT2I

OPT2I (Mañas et al., 2024) uses a large language model to iteratively refine text prompts—without any model fine-tuning—to maximize a consistency score. By automating prompt optimization, it substantially improves alignment between generated images and their descriptions while preserving output quality and diversity.

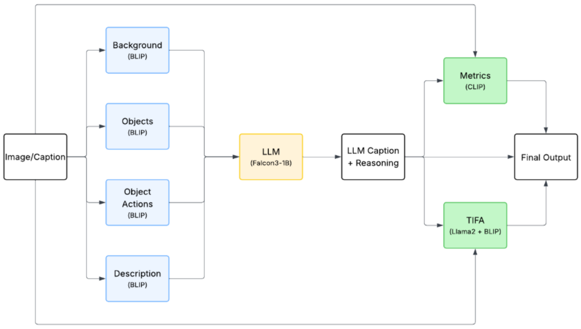

Figure 1: System architecture of our modular pipeline.

2.4 BLIP and BLIP-2

BLIP(Li et al., 2022) and BLIP-2 (Li et al., 2023) connects frozen vision encoders to large language models via a lightweight Querying Transformer. This design achieves state-of-the-art performance on vision–language benchmarks with minimal trainable parameters, making it an efficient backbone for downstream consistency-checking tasks.

2.5 Detectron2

Detectron2 (Wu et al., 2019) is Facebook AI's modular object detection library, offering bounding box, attribute, and keypoint extraction. Its high performance and extensible design make it a key component for grounding textual content in images.

2.6 CLIP

CLIP (Radford et al., 2021) maps images and text into a shared embedding space, allowing similarity comparisons without explicit supervision. Its ability to evaluate semantic alignment across diverse inputs with minimal computation made it central to our final evaluation approach.

3 Problem Description

Our objective is to assess whether a caption accurately matches an image. Rather than comparing captions to each other, we evaluate the alignment of each caption directly with the visual input. We apply the following steps:

- Select an image from MS-COCO.
- Use CLIP to compute a similarity score for an MS-COCO reference caption.
- Generate a new caption using BLIP.
- Score the new caption using CLIP and TIFA (If necessary).
- Compare the two scores to determine alignment improvement.

Higher scores indicate stronger alignment. The difference in scores (ΔCLIP) highlights the comparative performance of model-generated captions.

4 Methods

4.1 Materials

We curated images from four categories using the Microsoft Common Objects in Context (MS COCO) (Lin et al., 2015) dataset:

- **Animals**: such as horses in a field,
- **Actions**: people performing activities like running or jumping,
- **Objects**: static items such as tools or furniture,
- **Indoor backgrounds**: including cluttered or messy rooms.

4.2 Processing Pipeline

Our evaluation process follows these five stages:

1. Select an image from the MS-COCO dataset.
2. Use CLIP to compute a similarity score for the human-written caption.
3. Generate a candidate caption using BLIP-2 in zero-shot mode.
4. Compute a CLIP score for the generated caption.
5. Measure the difference (ΔCLIP) between the two captions to assess relative alignment.

4.3 Model Usage

- **BLIP** is used for image caption generation in zero-shot mode, without fine-tuning. It produces descriptive captions directly from image inputs.
- **CLIP** maps images and captions into a shared embedding space. We compute cosine similarity between embeddings to evaluate the alignment between image and text.

4.4 Evaluation Metrics

We explored both traditional language-based metrics and modern vision-language techniques:

- **BLEU** (Papineni et al., 2002), **ROUGE-L** (Lin, 2004), and **METEOR** (Banerjee and Lavie, 2005) assess surface-level textual similarity between captions and references. These metrics help evaluate overlap in word choice, structure, and phrasing.

- **CLIP Score** measures semantic alignment between image and text by comparing their vector representations in a shared embedding space. We selected CLIP as our final metric due to its ability to capture deeper conceptual consistency, even when phrasing differs.

4.5 Design Philosophy

The system is designed to be modular and script-based, supporting flexible integration of new models and scalable evaluation across image types. It accommodates batch comparisons and is generalizable to future captioning systems beyond the MS COCO dataset.

5 Experimental Results

We conducted a qualitative evaluation across 50 images by comparing user-written (reference) captions with machine-generated (candidate) alternatives. Each candidate caption was assessed for visual alignment using a Vision-Language Model (VLM), which analyzed image-caption consistency and flagged inaccuracies or omissions.

Candidate captions improved over the reference in **76%** of cases, often by providing richer scene descriptions or incorporating visually grounded details absent from the original. In many instances, user-written captions were minimal or vague (e.g., *"Donuts"* or *"A man with some sheep"*), lacking specific elements visible in the image. The corresponding candidate captions enriched these descriptions by including concrete visual details such as *"a box of donuts containing several donuts"* or *"a man in a blue robe and a sheep standing in a field surrounded by trees"*, which better reflected the visual context as illustrated in Figure 2. Other improvements included clarifying settings (e.g., transforming *"People discussing in a meeting"* into *"A group of people sitting around a computer, engaged in discussions"*) and adding spatial or compositional cues like *"lush, green trees"* in the background. These enhancements contributed to higher semantic alignment and more accurate visual grounding overall.

Quantitatively, candidate captions led to improvements in the majority of cases:

- **Images with improved scores: 75.0%**
- **Average Reference Score: 65.43%**
- **Average Candidate Score: 72.65%**
- **Average Score Improvement: 7.22%**
- **Average Increase (only improved cases): 13.73%**

Notably, some candidate captions introduced factual errors (e.g., describing a parked plane for a stunt or misidentifying a woman instead of a man), which led to score decreases. These cases highlight the importance of grounding generated text not only in linguistic quality but also in accurate scene understanding.

Overall, the evaluation demonstrates that machine-generated captions can enhance descriptive clarity and alignment when carefully guided, but still require safeguards against hallucinations and misinterpretations.

6 Challenges and Insights

Our use of CLIP-based similarity provided a more semantically grounded method to evaluate image-caption alignment than traditional text-based metrics. While models like BLEU, ROUGE, and METEOR offer helpful surface-level comparisons, they often penalize conciseness and fail to detect missing visual elements. For instance, candidate captions averaged just 3.01% BLEU, 16.92% METEOR, and had a high average perplexity of 20,629, indicating issues with linguistic fluency and coverage. CLIP, on the other hand, enabled a more nuanced understanding of how well the content of a caption matched what was present in the image—75.0% of candidate captions showed improved alignment, with an average candidate score of 72.65% compared to 65.43% for reference captions. On average, this yielded a 7.22 percentage point improvement, and a 13.73% increase when considering only the improved cases.

We found that captions describing vivid scenes with identifiable subjects and added context—such as the horse grazing near trees or the man and sheep

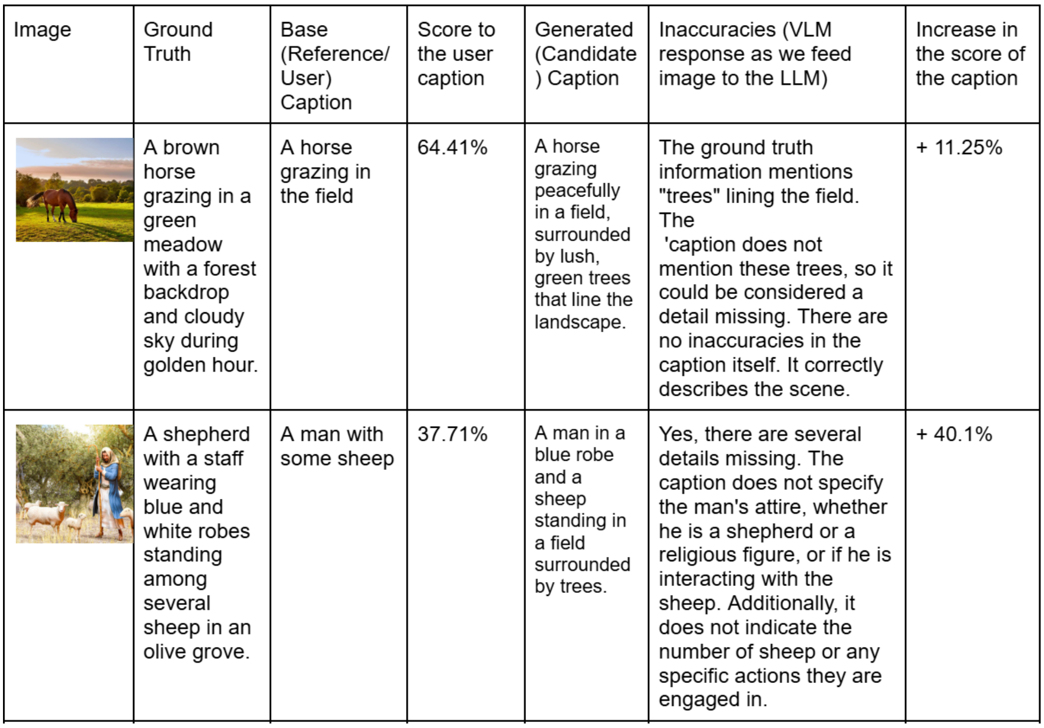

Image	Ground Truth	Base (Reference/ User) Caption	Score to the user caption	Generated (Candidate) Caption	Inaccuracies (VLM response as we feed image to the LLM)	Increase in the score of the caption
	A brown horse grazing in a green meadow with a forest backdrop and cloudy sky during golden hour.	A horse grazing in the field	64.41%	A horse grazing peacefully in a field, surrounded by lush, green trees that line the landscape.	The ground truth information mentions "trees" lining the field. The 'caption does not mention these trees, so it could be considered a detail missing. There are no inaccuracies in the caption itself. It correctly describes the scene.	+ 11.25%
	A shepherd with a staff wearing blue and white robes standing among several sheep in an olive grove.	A man with some sheep	37.71%	A man in a blue robe and a sheep standing in a field surrounded by trees.	Yes, there are several details missing. The caption does not specify the man's attire, whether he is a shepherd or a religious figure, or if he is interacting with the sheep. Additionally, it does not indicate the number of sheep or any specific actions they are engaged in.	+ 40.1%

Figure 2: Comparison of user vs. model-generated captions with faithfulness scoring and interpretable feedback.

example—benefited significantly from our generation process. In these cases, CLIP score improvements reached up to $+40.1\%$ (for the shepherd scene) and $+11.25\%$ (for the horse image), demonstrating that generated captions can capture scene richness that user captions often miss. However, CLIP's scoring still showed limitations in capturing fine-grained visual details like specific clothing, interactions, or object count. For instance, the generated caption for the shepherd image omitted key contextual cues such as the man's role, attire, and the number of sheep, yet still received a high score increase—highlighting CLIP's occasional insensitivity to semantic precision.

From a system design perspective, our modular pipeline worked well for adapting new models and scaling across different image domains. Still, we observed that high CLIP scores do not always correlate with completeness or truthfulness—while 75.0% of candidate captions showed improved alignment scores, some of these gains came from captions that were more fluent or concise but not necessarily more informative. Additionally, while our use of MS-COCO helped validate the model, some reference captions were overly generic (e.g., "A kitchen stove" or "Donuts"), which may have skewed early evaluations by enabling minimalist captions to superficially align with them in terms of CLIP similarity without adding meaningful scene detail.

Technical constraints also presented challenges. Running larger models like BLIP-2 and integrating evaluation tools required optimization to fit available computational resources. In future iterations, combining CLIP with human-in-the-loop feedback or question-answer-based models like TIFA could further enhance the reliability and depth of our scoring framework.

7 Conclusion and Future Works

ImageIntegrity evaluates how well captions match their images using a reference-free approach. Though we validated it using MS-COCO, the pipeline can be extended to other datasets, and further fine tuned to specific domains such as medical or justice system, with a large enough set of labelled data for that fine tuning. Future directions include refining action detection, integrating visual Q/A models such as TIFA, and exploring new scoring mechanisms that incorporate visual attributes. For more complex images, there can be human evaluation for scoring validation, especially for more ambiguous images that are hard to describe or not detailed enough. A diverse dataset would make the model stronger in the future, since MS COCO was intended to contain a lot of images even a child could recognize.

8 Division of the Labor between the Teammates

Atharva Bhide: Atharva led the integration of image segmentation and action detection modules using Detectron2, RCNN (Girshick et al., 2014), and MMAction2 (MMAction2). He experimented with CLIP and BLIP models, explored multiple segmentation techniques (keypoint, instance, panoptic), and assessed their usefulness for generating detailed base captions. He also tested BLEU and cosine similarity to identify scoring gaps. Created the test data set for the model evaluation. Also, he was responsible for compiling the reports.

Cheryl Khau: As project manager, Cheryl coordinated team meetings, documented progress, edited reports and presentation, and helped define the project scope. She researched academic literature to develop the evaluation framework and tested various BLIP-2 model variants for caption generation. She also explored promptless captioning and supported technical members by aligning research with implementation goals.

Jeannie Jiang: Jeannie implemented the evaluation metrics module, including CLIP Score, BLEU with smoothing, ROUGE-L, and GPT-2 perplexity. She created a reusable testing script using six curated image samples to allow consistent evaluation and quick extensibility. Her module outputs clearly report performance metrics per image-caption pair.

Rahul Sura: Rahul initially worked with llava (Liu et al., 2023) for action recognition and compared large and small model variants, trying to determine whether action detection benefits more from a specialized model than a generic captioner, although it was not used, since the pipeline didn't require it. He also conducted research on metric effectiveness for action- and attribute-heavy captions, as well as created a weighted metric evaluation model to place more weight on certain metrics, such as clip, bleu, meteor, perplexity, etc. (Banerjee and Lavie, 2005).

Jake Treska: Jake developed the structure and implementation of the pipeline utilizing BLIP for image analysis and Falcon3-1B for caption/result generation. Jake also developed the TIFA-based consistency checker by integrating TIFA with a VQA model (BLIP). He tested its ability to verify captions through generated question-answer pairs. Finally he also handled technical bottlenecks related to running LLM-based models on constrained hardware and proposed solutions like remote execution and quantization.

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.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *Preprint*, arXiv:1311.2524.

Chanda Grover, Indra Deep Mastan, and Debayan Gupta. 2022. Contextclip: Contextual alignment of image-text pairs on clip visual representations. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP'22, page 1–10. ACM.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. 2024. Improving text-to-image consistency via automatic prompt optimization. *Preprint*, arXiv:2403.17804.

MMAction2. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. [link].

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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. *Preprint*, arXiv:2103.00020.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. `https://github.com/facebookresearch/detectron2`.