RECAP: Reconstructing Images for Caption Evaluation

Adrian Chang
Brown University
adrian_chang@brown.edu

Sheridan Feucht
Brown University
sheridan_feucht@brown.edu

Shreyas Sundara Raman
Brown University
shreyas_sundara_raman@brown.edu

Abstract

Traditional text-text caption evaluation uses heuristics to perform token matching between a generated caption and a limited set of reference captions; these metrics (BLEU, METEOR, CIDer and SPICE) are restricted to textual modalities, requiring "ground truth" captions, which may not contain diverse semantic representations that cover the variance of possible image captions. These metrics might also favor adversarial captions designed to have high token similarity. Thus, a caption evaluation metric that leverages visual semantic information might be more robust. We propose a metric that scores captions on the the ability to recreate the imageś visual semantic information (in latent or pixel space). We evaluate different image-image comparison metrics against standard text-text metrics for robustness and correlation with human preference. Though our image reconstructions in pixel space are promising and latent reconstructions are sensitive to adversarial captions, we find our method has shortcomings due to current text-to-image models' invariance to prompt syntactic structure and poor learning signals from certain adversarially designed captions.

1 Introduction

Automatic image captioning methods attempt to generate a textual description of an image that is both semantically valid and contains sufficient visual descriptions. Such methods have numerous possible applications in human computer interaction, video surveillance, and medical imaging. In addition to these demands, image captioning models can also provide quality of life improvements to visually impaired individuals. An ideal image caption evaluation metric reflects these use cases, but for the case of the visually impaired, current ground truth captions often underspecify the level of detail in the image. See the appendix for examples of such captions.

A single image can have multiple valid captions and vice versa, a single caption can describe multiple different images. Caption evaluation is poorly defined because of the numerous factors such as specificity, number of objects, length, and semantic correctness which must be accounted for. We address this limitation by defining a good caption as a text description containing enough information to reconstruct an image with semantic and visual information similar to the original image. We predict that an image captioning captioning metric which incorporates visual information will be both more robust and have a stronger correlation with human judgement than a purely text-text metric.

In order to achieve both high quality text to image reconstruction and image-image comparison, we leverage the power of foundation models Stability Diffusion [Rombach et al., 2021] and CLIP [Radford et al., 2021b]. We generate images from ground truth captions with Stability Diffusion and

compare them to images the original caption belongs to. In addition to explicitly reconstructing the image, we also attempt to reconstruct the image embeddings projected to GPT-2 latent space, as a latent reconstruction evaluation metric. We then perform several experiments comparing our two proposed pipelines to current standard text-text metrics.

2 Related Work

Caption evaluation assess the quality of generated captions against a reference image and/or a set of annotated reference captions. The evaluation of a natural language generation (NLG) system is a fundamentally difficult task. In practice, there are three common approaches Bernardi et al. [2017]

2.1 Common Approaches for Caption Evaluation

1) Human evaluation relies on the subjective scoring of generated captions by human annotators e.g. on Amazon Turk (MTurk), where captions are judged on grammar, syntax, relevance, correctness, logical inconsistencies and specificity.Bernardi et al. [2017]

2) Text-text evaluation uses heuristics to automatically compare generated captions with a small (1-5 samples) set of reference captions; each of these metrics output a score capturing a different notion of "similarity" with reference captions. Most text-text metrics were originally developed for evaluating machine translation or text summarization, except CIDer, which was explicitly designed to evaluate image captions. This approach for caption evaluation has historically been subject to extensive critique, with multiple findings of poor correlation with human preference, particualrly for BLEU and METEOR. (Bernardi et al. [2017])

- BLEU measures the geometric mean of multiple clipped n-gram precisions between generated and reference captions, with "n" between 1 to 4. Higher BLEU score implies generated captions directly match ground-truth captions at the token-level (Papineni et al. [2002]).
- METEOR constructs an alignment between unigrams in generated and reference captions over multiple stages including: exact match, stemmed match and synonymous match; the mean F1 score of the alignment is calculated, with a penalty for non-adjacent mappings (Banerjee and Lavie [2005]). METEOR is more robust to inflexional forms than BLEU.
- SPICE constructs semantic propositions that capture the objects, attributes and relationships in captions. Semantic propositions are combined by their syntatic dependency and predefined semantic role labeling (SRL) rules into propositional scene graphs; the conjunction of semantic propositions over all semantic propositions tuples present in both generated and reference scene graphs is taken. SPICE prioritizes abstract entity relationships, and has shown high pearson's correlations (0.88) with human judgment (Anderson et al. [2016]).
- CIDEr computes TF-IDF scores for tokens in the reference and generated captions. The average cosine similarity between TF-IDF vectors of generated and reference caption is taken. CIDEr ensures that generated captions match the diversity of reference captions (Ramakrishna Vedantam [2014]).

Whilst NLG caption systems are trained on next token prediction, these text-text metrics use sentence-level comparisons; they are also non-differentiable and cannot be used to influence caption generation during training. Hence, there is a mismatch between training objectives for captioning and traditional caption evaluation metrics (Ghandi et al. [2022]).

3) Machine Learning methods such as (Cui et al. [2018]) and (tig) train a caption evaluator using embedded representations of the image, reference captions and generated captions. Cui et al. [2018] uses Compact Billinear Pooling on the Fourier transformations of the embedded image and captions to predict binary classification differentiating machine v.s. human generated captions; tig compares spatial similarity of the embeddings of reference and candidate captions that are cross-attended with a latent image representation. Our learned approach differs in that we attempt to reconstruct the image embedding in latent space and we do not utilize reference captions as inputs to the model.

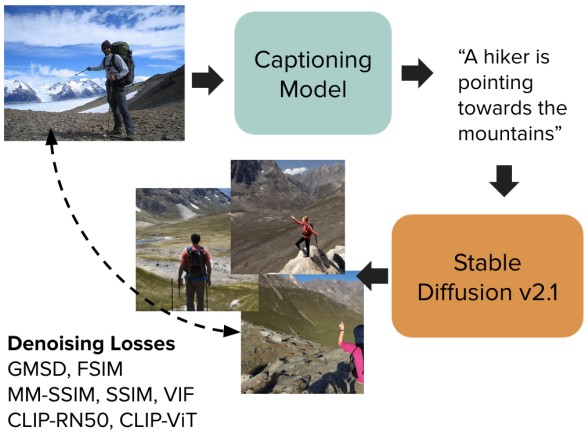


Figure 1: Explicit reconstruction of images to evaluate captions.

2.2 Visual-Language Models for Image Captioning

Visual-language models like CLIP (Radford et al. [2021a]) provide powerful joint embedding spaces between natural language and visual features. This has transformed approaches to captioning and caption evaluation. CLIPScore (Jack Hessel and Choi [2022]) directly uses CLIPś cosine similarity, scaled by a constant factor, as a numerical caption "score". ClipCap (Ron Mokady and Bermano [2021]) trains a small mapping network (transformer or MLP) to project CLIP image embeddings to GPT-2 space; the projected embedding acts like a "soft prompt" for a GPT-2 decoder head and is trained on the language modelling objective.

3 Methodology

3.1 Explicit Reconstruction

Our first approach is that of explicit reconstruction of an image in order to evaluate a given caption (Figure 1). We first take the caption in question (generated by some captioning model under evaluation) and feed it through a text-to-image model, in this case, Stable Diffusion v2.1. We sample three images from Stable Diffusion. For each sampled image, we can then perform image-to-image comparisons to attempt to evaluate the validity of the reconstructions. Results reported in this paper represent the mean over all generated images.

There are two types of metrics we use to compare two images: low-level denoising metrics from GAN literature as well as higher-level comparisons of CLIP image embeddings. For low-level denoising metrics, we experimented with SSIM [Wang et al., 2012], MS-SSIM [Md and Channappayya, 2016], VIF [Sheikh and Bovik, 2004], GMSD [Xue et al., 2013], and FSIM [Zhang et al., 2011]. All of these were implemented in the PyTorch Image Quality library [Kastryulin et al., 2022]. The two CLIP comparisons we made were based on CLIP-RN50 and CLIP-ViT-L/14 embeddings, using OpenAI's CLIP interface. While the low-level metrics used were designed as loss functions for the task of denoising an image, the hope was that they might capture information about similarity of lower-level image features (for example, giving a better rating to images that both have grass textures versus an image with a snow texture and a grass texture).

3.2 Latent Reconstruction

Due to the high resource and time requirements with explicit image reconstructions, we also explore reconstructing the image embedding in latent space to evaluate a given caption (Figure 2). We pre-train and then freeze the ClipCap (Ron Mokady and Bermano [2021]) mapping network (transformer) using CLIP's ViT-L/14 image encoder and remove the GPT-2 decoder head. As mentioned in subsection 2.2, ClipCap's mapping network allows us to project images to GPT-2 latent space whilst preserving semantic information for captioning. On the caption side, GPT-2's pretrained text encoder

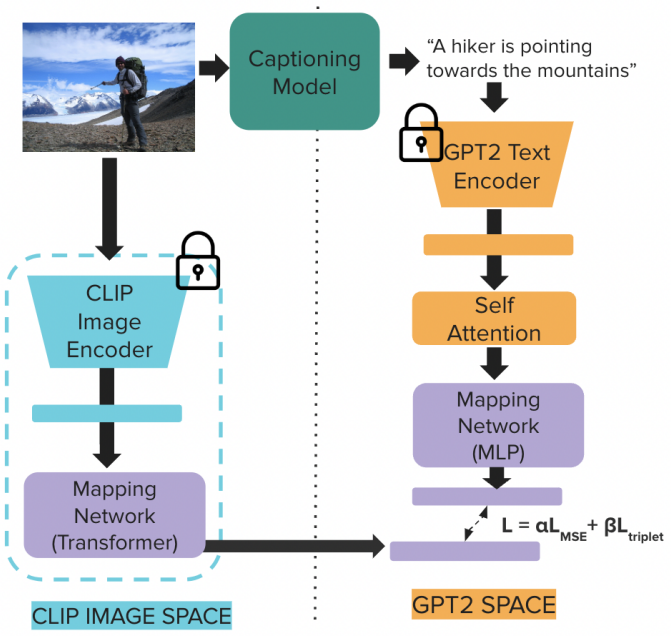


Figure 2: Latent reconstruction of images to evaluate captions.

feeds the embedded caption to a self-attention layer (with 8 heads and embedding size 768) followed by a multi-layer perceptron (MLP with 3M parameters); only the self-attention layer and MLP are trained to "reconstruct" the image embedding in GPT-2 latent space. To motivate the reconstruction, we use a weighted combination of MSE and a Triplet Margin loss 1. The MSE loss ensures the mapped caption embedding is pushed closer to the mapped image embedding; the Triplet loss tries to bring the "original" caption embedding (positive anchor) closer to image embedding than and a "noisy" caption embedding (negative anchor). The "noisy" caption is either a random caption (selected from a different image to the one being compared) or a word permutation; one type of noise is randomly chosen and applied for each batch. During test time, the distance between reconstructed embeddings serve as the image-caption evaluation metric: we consider the average MSE distance (Latent (Mean)) and the L2 distance (Latent (L2)) between image and reconstructed latent vectors.

$$
\begin{aligned}
\mathrm{L} &= \alpha \mathcal{L}_{MSE} + \beta \mathcal{L}_{triplet} \\
\mathrm{L}_{MSE} &= (x - \hat{x})^2 \\
\mathrm{L}_{triplet} &= max(d(a_i, p_i) - d(a_i, n_i) + m, 0)
\end{aligned}
\tag{1}
$$

where $\hat{x}$ is the latent reconstruction; α and β are weighting coefficients to prioritize reconstruction similarity and resistance to pathological captions, respectively. We use $\alpha = 1$ and $\beta = 10$ to bring both losses to the same order of magnitude, so they have similar influence on the summed loss.

4 Experiments

We perform a number of experiments to determine the robustness of both proposed captioning metrics. The first two sections (Section 4.1 and 4.2) describe experimental designs where pathological captions are evaluated using our metric; we apply pathological transformations to $0\%, 20\%, 50\%, 75\%$ and 90% of the Flickr8k test set, with the hope that our metrics will be sensitive to deleterious changes that make caption quality worse. Since we present the normalized scores for fair comparison, we expect performance degradation to be linearly proportional to the extent of pathological transforma-

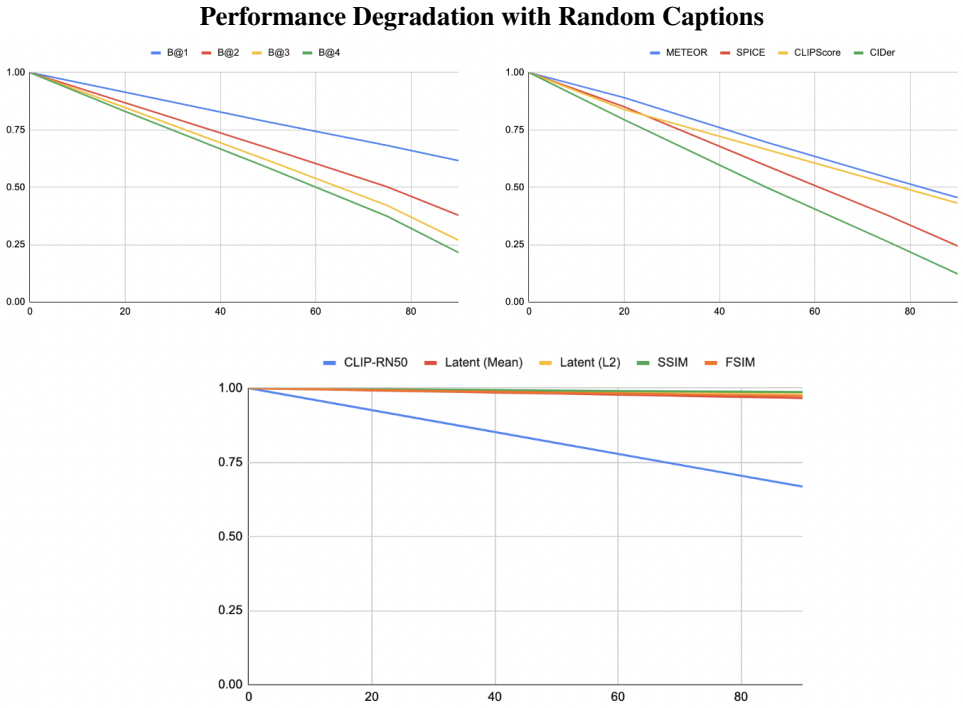


Figure 3: Results when we swap which captions are associated with which images. If a caption does not belong to an image, scores should drop accordingly. For all of our metrics, explicit and latent included, swapping of captions does little to degrade performance indicating a lack of sensitivity to this pathological transformation. Only the CLIP-RN50 image-image metric seems to be sensitive to swapped captions.

tions applied. The third section (Section 4.3) measures the correlation of each metric with human preferences of "correctness" for different image-caption pairs, using expert evaluations from a subset of Flickr8k.

4.1 Resistance to Random Captions

Figure 3 shows how the score assignment of our mertics change when we swap which captions are associated with which images. If a caption does not belong to an image and the original image-caption pairing is perturbed, an ideal captioning metric should be sensitive to semantic changes and reduce score assignment proportional to the percentage of examples that are swapped. Most optimal degradation is achieved by CIDer that nearly scale 1:1 with more transformation. Other baseline text-text metrics are sensitive to random swaps, but the only metric of ours that drops performance (to 0.67 with 90% random captions) is cosine similarity comparison using CLIP-RN50 image embeddings. This indicates that our proposed metrics are not sensitive to random swaps and cannot distinguish semantic changes in the caption relative to visual semantics of the image.

4.2 Resistance to Word Permutations

Figure 4 shows how the score assignment of our mertics change when we preserve image-caption pairings but shuffle the tokens within a caption and then evaluate for quality. An ideal captioning metric should be sensitive to the loss of syntactic structure in the shuffled caption, and reduce score assignment proportional to the percent of captions shuffled. Most optimal degradation is achieved by BLEU-3 and BLEU-4 that nearly scale 1:1 with increase permutation. Our explicit reconstruction metrics are not sensitive to word order; we find that the latent reconstructions do degrade performance when more captions are shuffled, though "Latent (Mean)" metric disproportionately reduces average score for the degree of transformation and "Latent (L2)" metric does not degrade fast enough.

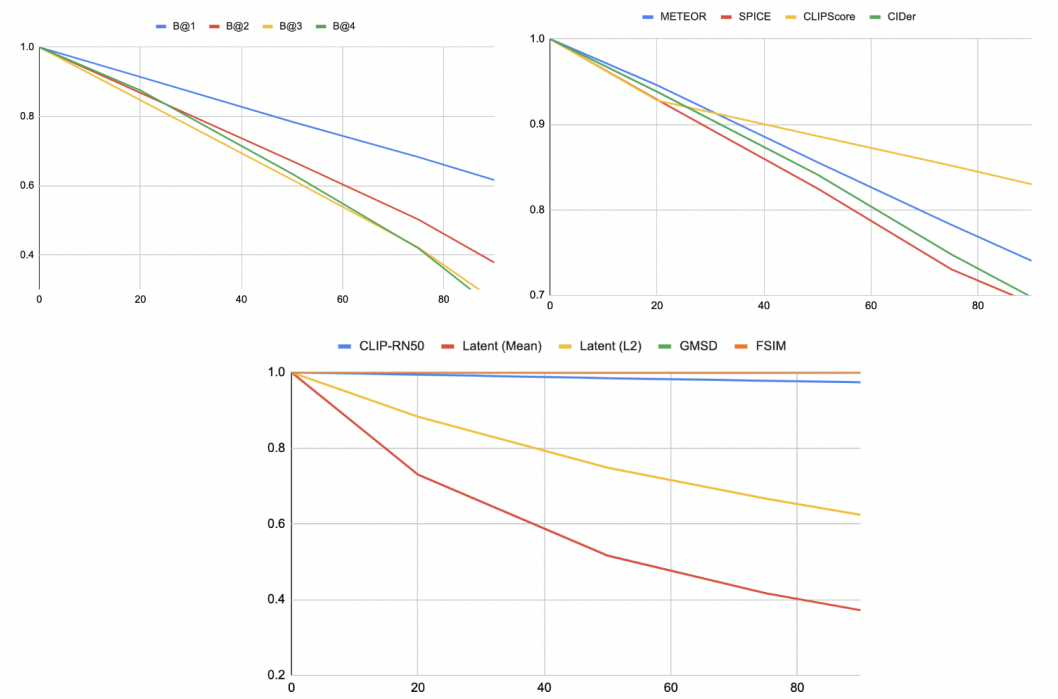


Figure 4: The two top figures show drops in baseline scores (BLEU-{1,2,3,4}, METEOR, SPICE, CLIPScore, and CIDer) as the percentage of captions with tokens shuffled increases. The bottom shows that all image-image metrics (CLIP-RN50, GMSD, FSIM) lack sensitivity to caption word ordering and syntax. However, performance *does* degrade when using latent reconstruction.

Surprisingly, image-text CLIPScore baseline is far less sensitive than our reconstruction metrics and baselines that operate purely in text space.

Figure 5 also validates our observations. The left and right boxplots show distributions of scores for unshuffled captions and shuffled captions, respectively. Most baseline metrics tend to have substantially lower median score with lower IQR when captions are shuffled, with the exception of CLIPScore. There are no significant differences in median or IQR of scores between our image-image metrics. However our latent methods show significantly lower (better) score assignment with $3\times$ smaller IQR without shuffling; the score range and IQR are far wider with shuffling, though the bottom 25% still overlaps in value.

4.3 Human Preferences

Finally, we use human evaluations from a subset of Flickr8k, which consists of 5822 image-caption pairs rated on "how well the caption describes the image" by three expert annotators. The ratings are ordinal in the range 1-4 with higher score implying a more descriptive caption; we compute the the mean score across the annotators and find the correlation between these "human prefernce" scores and the scores assigned by each of our metrics, including the baseline metrics.

We use Kendall's Tau [Kendall, 1938], a nonparameteric method of measuring the strength of association (monotonicity) of two variables, to assess correlation. An absolute value greater than 0.3 shows strong correlation; absolute values between 0.1-0.3 have notable but weaker correlation; absolute values below this show no correlation. Table 1 shows these results. Certain "loss based" metrics (marked with $\downarrow$) reduce score assignment for captions of higher quality, so a negative correlation is desirable for them. For our explicit image reconstruction metrics, we perform an ablation that correlates the scores assigned to human preference, when the original images are replaced by random noise. This helps us assess whether these image-image metrics capture features that are aligned with or important to human preference.

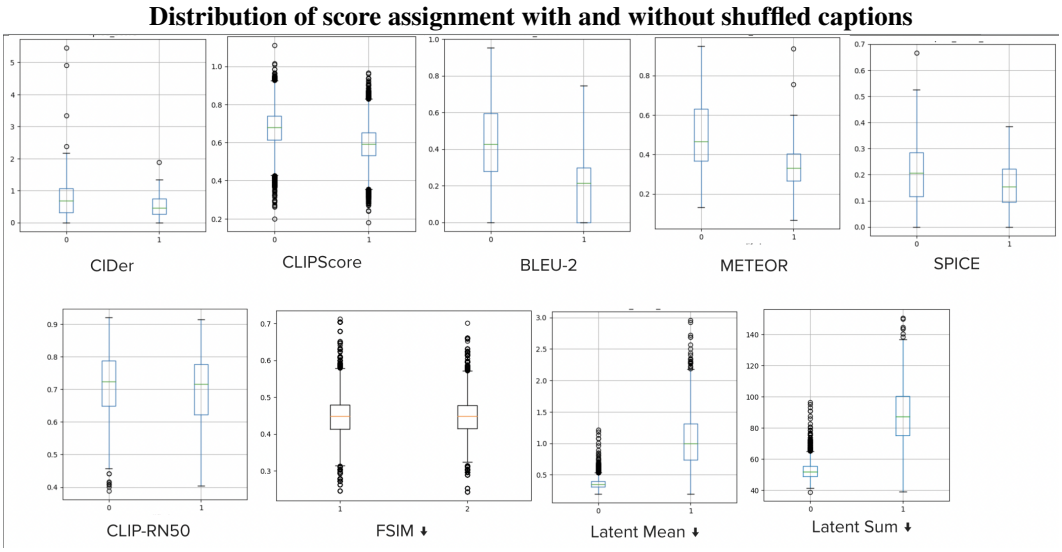


Figure 5: In each subfigure, the left and right boxplots show distribution of scores for unshuffled and shuffled captions respectively. There are no significant differences in score distribution with and without shuffling for image-image metrics, validating results from Figure 4. However, our latent method is sensitive to caption shuffling, even more than baseline methods in the top row.

Method	τ	p-value
BLEU-1	0.3759	0.0
BLEU-2	0.3923	0.0
BLEU-3	0.3819	0.0
BLEU-4	0.3759	0.0
METEOR	0.4115	0.0
CIDer	0.0055	0.6
SPICE	**0.5913**	0.0
CLIPScore	**0.5947**	0.0

Method	τ	p-value	τ with noise	p-value
SSIM ↓	-0.0196	0.0	-0.0597	0.0
MS-SSIM ↓	-0.0313	0.0	-0.0319	0.0
VIF ↓	-0.0011	0.9	-0.0242	0.0
GMSD ↓	*-0.1024*	0.0	-0.0729	0.0
FSIM ↓	***-0.1106***	0.0	-0.0517	0.0
CLIP-RN50 ↑	***0.4722***	0.0	-0.0174	0.0
CLIP-ViTL ↑	***0.4201***	0.0	-0.0374	0.0
Latent ↓	-0.1075	0.0	–	–

Table 1: Correlations of assigned scores with human preference, for different metrics. The left table shows correlations for baseline text-text and text-image metrics, which all have strong correlations, except for CIDer. The right table shows correlations for our explicit and latent reconstruction metrics, where most low-level image comparison metrics have no correlation.

We find that correlation is above chance (replacing images with random noise) for GMSD, FSIM, CLIP-RN50 and CLIP-ViTL image-image metrics, shown in italics. The latent method, GMSD and FSIM have weak correlations to and using CLIP-RN50 embeddings for image-image comparison seems to offer the strongest correlation with human preference of our metrics, outperforming most text-text baselines. However CLIP-RN50 does not outperform direct text-image comparison between caption and image using CLIPScore, which has the highest overall correlation with human preference.

5 Discussion

Our proposed metrics for caption evaluation, which use image reconstruction as a basis for score assignment, are not generally robust across all evaluated paradigms i.e. pathological captions (random or shuffled captions) and human preference correlation. Nearly all image-image metrics do not outperform CLIPScore or traditional text-text metrics in human preference; the image-image metrics are also unable discern changes in syntax and semantics of captions relative to visual semantics, as shown with experiments using pathological captions. However, the strong correlation (0.4722) and performance degradation seen with random captions makes us believe that reconstruction is a viable approach that might have better potential with carefully designed image-image comparison metrics and text-image models that are more sensitive to syntax.

The metrics adopted from GAN literature perform low-pixel-level comparisons between image-image, so perhaps incorporating language concepts (e.g. using SegmentAnything) when searching for high-level matching visual features might help connect the visual-language domain and be a fruitful direction for incorporating visual information in a captioning metric. The image-image metrics are also limited by the fidelity and accuracy of text-image models. Diffusion models are notorious for struggling with relations in text prompts and do not usually require syntactically accurate descriptions to generate meaningful images, resulting in possibly inaccurate image reconstructions even if the caption and/or image-image comparisons are adequate.

Our proposed latent reconstruction metric is also not generally robust, though it exhibits weak correlation with human preference and strong sensitivity (performance degradation) to shuffled pathological captions. The model achieves stable reduction in both MSE (reconstruction) and tripley margin loss 7. The median score (for Latent (Mean)) reduces from 1.0 to 0.47 (implying higher quality) without shuffling and the median with shuffling is far above the maximum (lowest quality) score without shuffling. However there are shuffled caption examples that are assigned low score (higher quality) in the first-quartile. Qualitative visualizations of reconstructions in Figure 8 show that shuffled captions have generally low activations, suggesting the triplet loss might be pushing shuffled captions embeddings towards the origin of the embedding space. There is also only a weak visual correlation between image embedding and reconstruction in the GPT-2 latent space.

We believe there is limited sensitivity to random captions because sampling completely random captions (from other images) provide a weak learning signal (weak negatives) to triplet margin loss compared to strong negatives from random shuffling; as a result, the model might be more motivated to be sensitive to shuffling transformations and the projections might actually lose linguistic semantic information in the original GPT-2 embedding. Using human-feedback to train selectively on high-preference captions or to choose more hard-negatives for random captions (e.g. "a boy riding a bicycle" (positive) and "a man riding a motorbike" (negative)) might improve human correlation and lead to stronger sensitivity to random captions. Additionally, projecting to a larger (or richer) language space like GPT-2 XL or GPT-3 might better capture semantic variations.

6 Division of Labour

Shreyas: ideation of latent reconstruction method, coding latent reconstruction method, training/fine-tuning latent method, coding scripts to run pathological caption experiments, running pathological captions evaluation, generating boxplots, writing paper, making presentation

Sheridan: ideation of explicit reconstruction method, coding explicit reconstruction method using Stable Diffusion, coding scripts to run human preference experiments, running evaluations for human preference, writing paper, making presentation

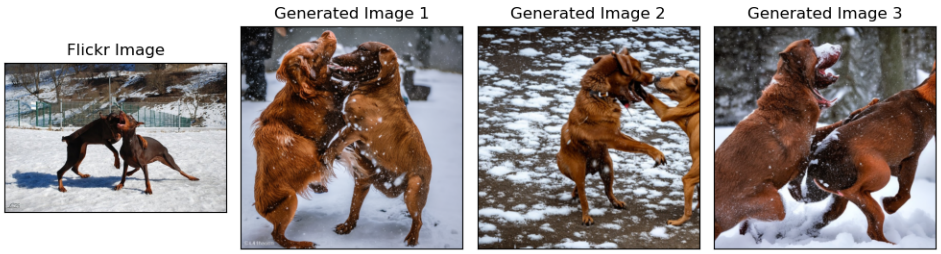
Adrian: ideation of explicit reconstruction method, coding explicit reconstruction method using Stable Diffusion, coding scripts to run human preference experiments, running evaluations for human preference, writing paper, making presentation

7 Appendix

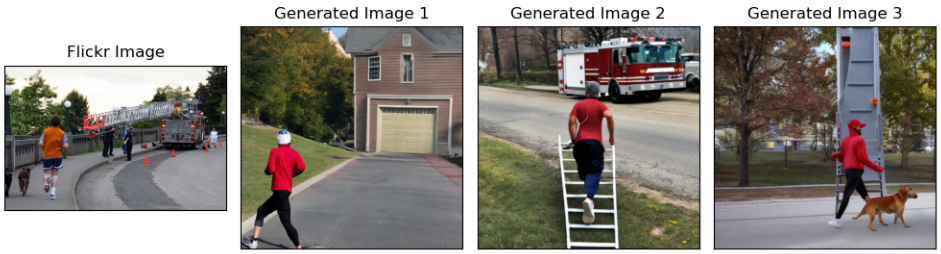
7.1 Examples of explicit image reconstructions

In this section we provide examples of ground truth image caption pairs alongside the images generated by the caption.

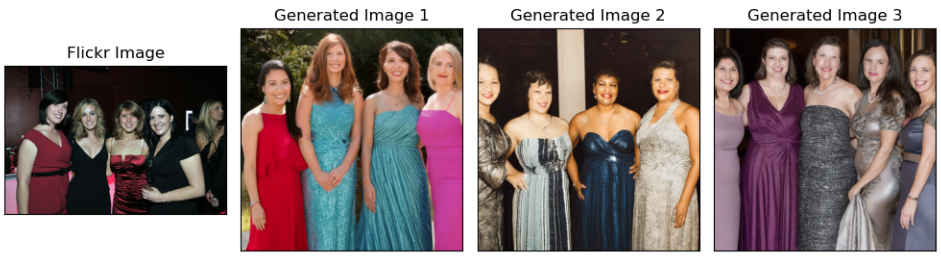
Two brown dogs playfully fight in the snow .



A jogger with a dog approaches a ladder truck .



Four woman wearing formal gowns pose together and smile .



Two lizards fighting .

Flickr Image

Generated Image 1

Generated Image 2

Generated Image 3

A man is hiking on a mountaintop on a cloudy day .

Flickr Image

Generated Image 1

Generated Image 2

Generated Image 3

7.2 Training loss for ClipCap (ViT-L/14) pretraining

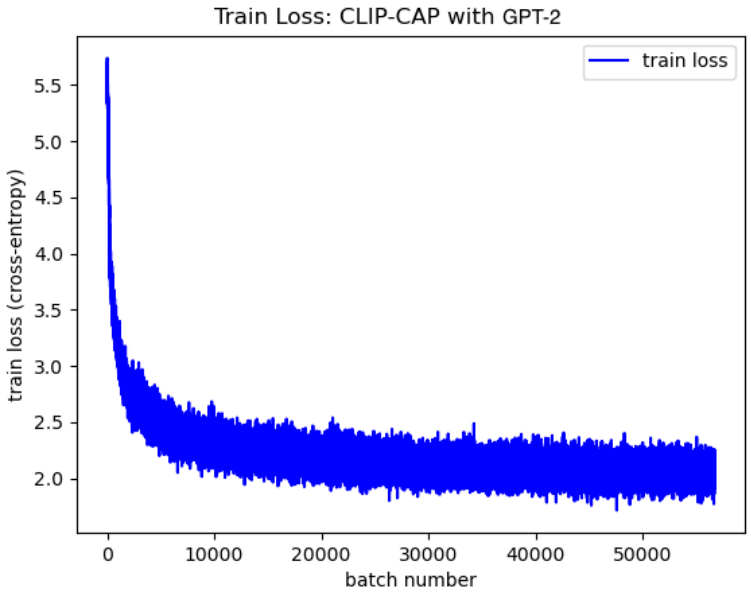


Figure 6: Pre-training the ClipCap mapper (transformer) on COCO for a caption generation. There is a sharp reduction in cross-entropy loss by batch 5000 after which the loss is volatile but bounded, though the moving average reduces slowly

7.3 Training loss curves for latent reconstruction

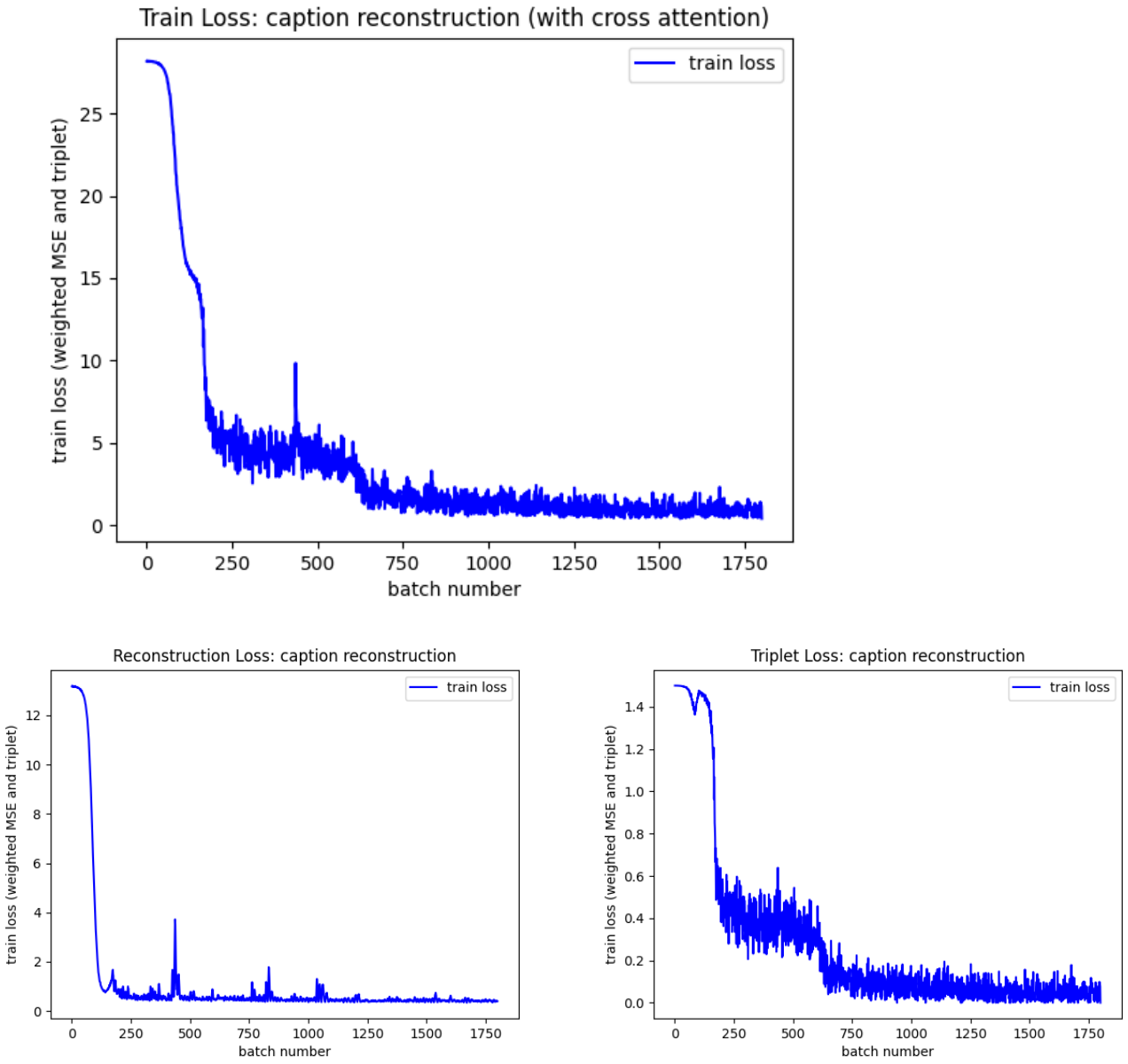


Figure 7: The latent image reconstruction model is capable of reducing total loss, with the MSE reconstruction loss reducing more sharply (within 250 batches) than the triplet loss. The triplet loss shows staggered reduction in three plateauing phases, eventually becoming volatile but bounded; we notice that the volatility of the triplet loss is very sensitive to the β in loss function 1

7.4 Comparing embedding visualizations of latent reconstructions

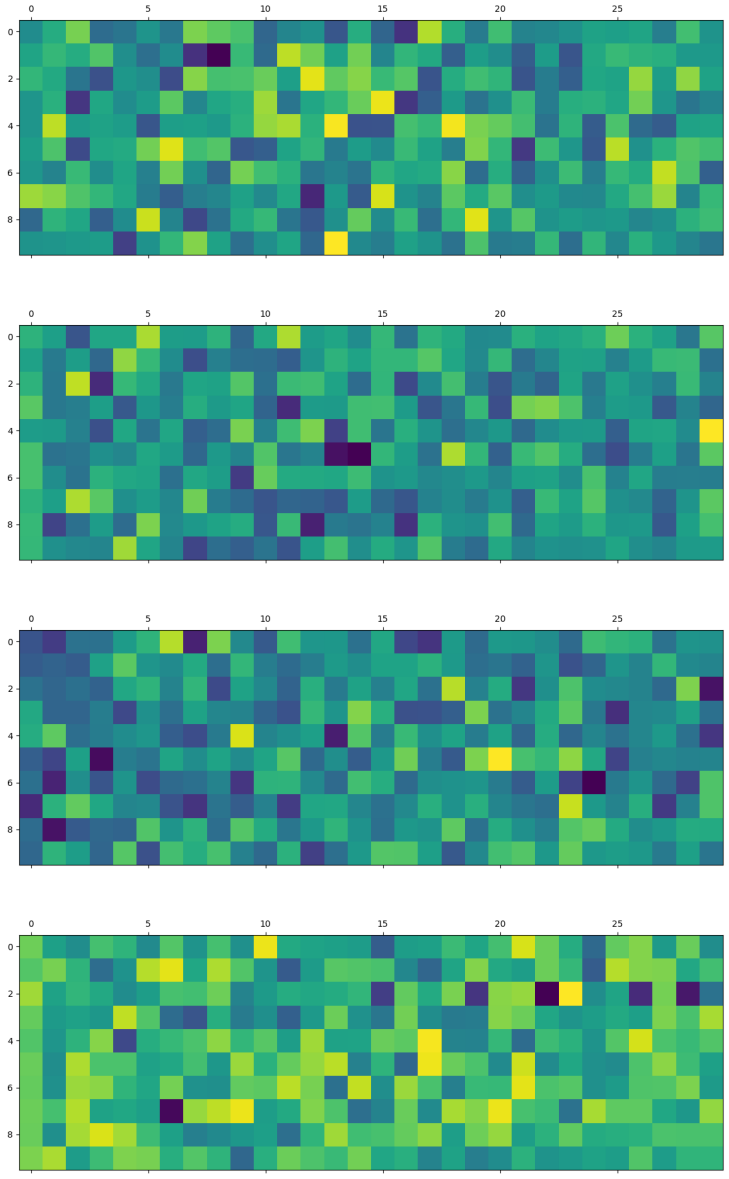


Figure 8: Visualizations of the latent embeddings for (from top to bottom): the reference (ground-truth) captions, reconstructed embeddings, reconstructions for captions with word permutations, reconstructions for captions from different images. Dimensionality reduction via PCA to $d = 30$ was applied to a normalized embedding vector (size 10×768) to make visualizations more interpretable

References

URL https://aclanthology.org/D19-1220.pdf.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. *arXiv preprint arXiv:1607.08822*, 2016.

Satanjeev Banerjee and Alon Lavie. 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, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures, 2017.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. Conference on Computer Vision and Pattern Recognition, 2018.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *arXiv preprint arXiv:2201.12944*, 2022.

Maxwell Forbes Ronan Le Bras Jack Hessel, Ari Holtzman and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. 2022. URL https://aclanthology.org/2021.emnlp-main.595v2.pdf.

Sergey Kastryulin, Jamil Zakirov, Denis Prokopenko, and Dmitry V. Dylov. Pytorch image quality: Metrics for image quality assessment, 2022. URL https://arxiv.org/abs/2208.14818.

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

Sameeulla Khan Md and Sumohana S. Channappayya. Multiscale-ssim index based stereoscopic image quality assessment. *2016 Twenty Second National Conference on Communication (NCC)*, pages 1–5, 2016.

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

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. Learning transferable visual models from natural language supervision, 2021a.

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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021b.

Devi Parikh Ramakrishna Vedantam, C. Lawrence Zitnick. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*, 2014.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

Amir Hertz Ron Mokady and Amit H. Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Hamid R. Sheikh and Alan Conrad Bovik. Image information and visual quality. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:iii–709, 2004.

Shiqi Wang, Abdul Rehman, Zhou Wang, Siwei Ma, and Wen Gao. Ssim-motivated rate-distortion optimization for video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 22:516–529, 2012.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan Conrad Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23:684–695, 2013.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20:2378–2386, 2011.