Cross-Modal Attribute Insertions for Assessing the Robustness of Vision-and-Language Learning

Shivaen Ramshetty[*] and **Gaurav Verma**[*] and **Srijan Kumar**
Georgia Institute of Technology
{sramshetty3, gverma, srijan}@gatech.edu

Abstract

The robustness of multimodal deep learning models to realistic changes in the input text is critical for their applicability to important tasks such as text-to-image retrieval and cross-modal entailment. To measure robustness, several existing approaches edit the text data, but do so without leveraging the cross-modal information present in multimodal data. Information from the visual modality, such as color, size, and shape, provide additional attributes that users can include in their inputs. Thus, we propose cross-modal attribute insertions as a realistic perturbation strategy for vision-and-language data that inserts visual attributes of the objects in the image into the corresponding text (e.g., "girl on a chair" → "little girl on a wooden chair"). Our proposed approach for cross-modal attribute insertions is modular, controllable, and task-agnostic. We find that augmenting input text using cross-modal insertions causes state-of-the-art approaches for text-to-image retrieval and cross-modal entailment to perform poorly, resulting in relative drops of ∼ 15% in MRR and ∼ 20% in F_1 score, respectively. Crowd-sourced annotations demonstrate that cross-modal insertions lead to higher quality augmentations for multimodal data than augmentations using text-only data, and are equivalent in quality to original examples. We release the code to encourage robustness evaluations of deep vision-and-language models: https://github.com/claws-lab/multimodal-robustness-xmai.

1 Introduction

The ability to model the interaction of information in vision and language modalities powers several web applications — text-to-image search (He et al., 2016), summarizing multimodal content (Zhu et al., 2018), visual question answering (Antol et al.,
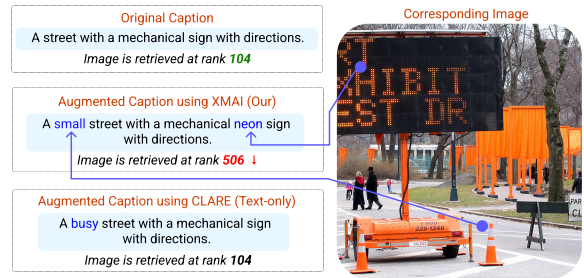

Figure 1: We propose Cross-Modal Attribute Insertions (XMAI) — an approach that leverages cross-modal interactions in multimodal data to obtain meaningful text augmentations that methods using text-only information (e.g., CLARE) cannot provide. These augmentations highlight vulnerabilities of multimodal models; in this case, the corresponding image is retrieved at a worse rank (104 → 506) for the modified caption.

2015), and editing images using language commands (Shi et al., 2021). Ensuring satisfactory user experience within such applications necessitates the development of multimodal models that can *robustly* process text and image data, jointly.

Existing research has demonstrated the brittle reasoning mechanism of text-only and image-only models by introducing variations in the inputs (Evtimov et al., 2020; Li et al., 2021a). Furthermore, prior work have established controlled generation methods for text (Ross et al., 2022), including counterfactuals for model assessment (Madaan et al., 2021; Wu et al., 2021). However, beyond applying modality-specific perturbations to multimodal (image + text) data (Qiu et al., 2022), existing research has not studied the robustness of models to *likely* augmentations in text that leverage cross-modal interactions. Specifically, current research on text augmentation considers the following likely variations: skipping certain words, introducing typographical errors, inserting noun or verb modifiers, or using synonyms. Consequently, to study the robustness of deep models, several automated methods have been developed to intro-

---

[*]Equal contribution.

duce these variations in the text. However, while these *text-only* perturbations can cover more variations, they are by no means exhaustive with respect to multimodal data. In the context of multimodal data, the text accompanying an image can be meaningfully perturbed to include information from the image. For instance, users can issue a query on a search engine that specifies attributes of the desired image(s); '*a male driver posing with a red car*' instead of '*a driver posing with a car.*' Existing augmentation approaches can only model text-only data and cannot introduce relevant cross-modal information (like 'male' and 'red' in the above example) while generating augmentations.

We propose novel text variations that leverage the image modality to insert relevant information into the text, which we call *cross-modal attribute insertions*. Our method inserts attributes of objects that are both present in the image and mentioned in the text. To do so, *cross-modal attribute insertion* uses object detection to capture objects and their attributes in the image (Anderson et al., 2018), and masked-language modeling to place those attributes prior to the object's mentions in the text (Devlin et al., 2019) (see Figure 1). Additionally, we use embedding similarities to expand the search space of possible augmentations, and introduce an adversarial component to estimate the robustness of multimodal models.

Our proposed approach is highly modular, controllable, and task-agnostic. Different modules govern attribute selection from images, cross-modal object matching, attribute insertion in text, and adversarial strength of the augmented example. The contribution of these modules toward the final augmented text can be controlled using weights that can be tuned as hyper-parameters. Finally, our approach for generating augmentations does not involve any parameter training, which makes it task-agnostic and broadly applicable.

We demonstrate the applicability of our cross-modal attribute insertion approach by generating augmentations for assessing the robustness of models for two different multimodal tasks — *(a)* text-to-image retrieval and *(b)* cross-modal entailment. Together, these two tasks are representative of ranking and classification multimodal tasks. Our evaluation comprises assessing the robustness of state-of-the-art multimodal learning approaches for these tasks to our augmentations as well as quantifying the relevance of generated augmentations to unmodified

examples. We contrast our cross-modal attribute insertions with several baseline approaches that model text-only information.

Our key contributions and findings are:

• We propose cross-modal attribute insertions as a new realistic variation in multimodal data. Our proposed approach introduces these variations in a modular, controllable, and task-agnostic manner.

• We demonstrate that state-of-the-art approaches for text-to-image retrieval and cross-modal entailment are not robust to cross-modal attribute insertions, demonstrating relative drops of $\sim 15\%$ and $\sim 20\%$ in MRR and F_1 score, respectively.

• While being as effective as existing text-only augmentation methods in highlighting model vulnerabilities, our approach produces augmentations that human annotators perceive to be of better quality than the most competitive text-only augmentation method. Furthermore, our method matches the quality of unmodified textual examples, while being at least $9\times$ faster than the most competitive baseline across the two multimodal tasks.

Overall, we find that cross-modal attribute insertions produce novel, realistic, and human-preferred text augmentations that are complementary to current text-only perturbations, and effectively highlight the vulnerabilities of multimodal models. Future work could employ our augmentation strategy to evaluate and develop more robust multimodal models.

2   Related Work

Text Augmentations in Multimodal Data: Existing research investigating the robustness of deep learning models for natural language processing has proposed several automated approaches to introduce plausible variations in the text. Ribeiro et al. (2020) and Naik et al. (2018) propose a comprehensive list of perturbations that NLP models should be robust to — including distracting phrases, URLs, word contractions and extensions. Many of these perturbations are task-agnostic and hence can be used to modify the text in multimodal (image + text) data as well. Similarly, other task-agnostic approaches to modify text data include random deletion, swapping, and insertion of words (Wei and Zou, 2019) and replacing, inserting, and merging words or phrases using masked language modeling (Li et al., 2021a). TextAttack (Morris et al., 2020) provides a comprehensive categorization of such methods and a framework to implement them.
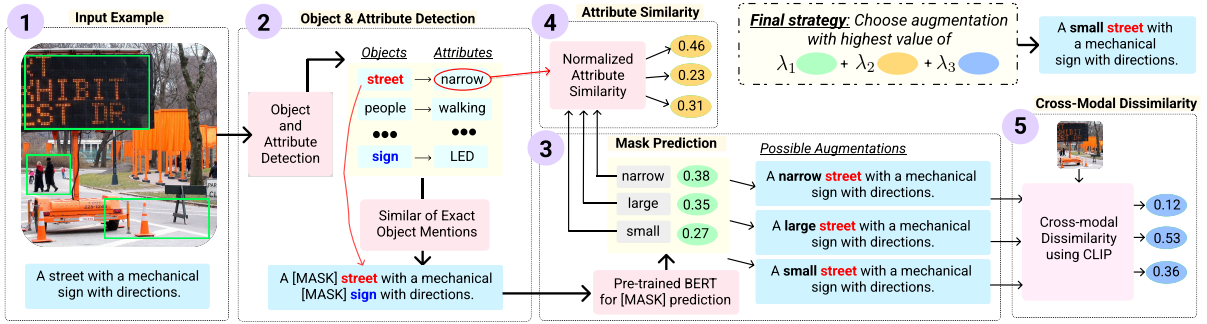
Figure 2: **Schematic depicting cross-modal attributes insertion**. For each input example *(1)*, we find objects that are depicted in the image and also mentioned in the text *(2)*. Masked-language modeling is used to predict potential words that could describe common objects, leading to candidate augmentations *(3)*. However, the final strategy takes into consideration the similarity of predicted words with the detected attribute for the object *(4)* and the cross-modal dissimilarity between augmented text and the original image *(5)* using λ_i hyper-parameters. The augmentation strategy is also presented as an algorithm in Appendix Alg. 1.

However, these methods lack in two critical ways: *(i)* automated text augmentations often compromise the semantic meaning to notable extents (Wang et al., 2021), and (ii) they only rely on the information contained in the text modality. In this work, we introduce augmentations in the textual part of multimodal data using TextAttack methods and consider them as *baseline* augmentations. Then to overcome the flaws mentioned, we propose an approach that leverages the information in the visual modality to insert visual attributes in the textual modality (i.e., cross-modal attribute insertions).

Robustness of Multimodal Models: Previous studies independently introduce unimodal perturbations in the visual or textual part of the input to study multimodal robustness. This could involve introducing imperceptible adversarial noise in the images and independently modifying the text using the augmentation approaches discussed earlier (Li et al., 2021a; Ribeiro et al., 2020; Wei and Zou, 2019). For instance, Chen et al. (2020) synthesize counterfactual samples of multimodal data using language models to modify the text. To ensure the preservation of the semantic meaning in the augmented text, Sheng et al. (2021) and Li et al. (2021b) employ humans to perturb the textual questions to fool the state-of-the-art models for Visual Question Answering (Antol et al., 2015). In a step towards using cross-modal interactions in image-text data to generate realistic variations, Verma et al. (2022) proposed adding relevant information from the image to expand the original textual description, and assess the robustness of multimodal classifiers. Our work proposes a different approach to leverage cross-modal associations to augment multimodal

data. Instead of expanding the original text, we *insert attributes* of objects in the image that are also mentioned in the corresponding text to modify the original text. Additionally, our work considers more multimodal tasks by studying text-to-image retrieval and cross-modal entailment.

3 Cross-Modal Attribute Insertions

Our approach for augmenting text in multimodal data involves identifying objects in the image that are also mentioned in the text, and inserting words similar to their attributes in the text at relevant places. An overview of our approach (XMAI) is depicted in Figure 2.

We denote paired multimodal units as $(\mathcal{I}, \mathcal{T})$, where $\mathcal{I}$ represents the input image and $\mathcal{T}$ is the text corresponding to that image. Our goal is to transform $\mathcal{T}$ into $\mathcal{T}'$ such that the text includes relevant information from $\mathcal{I}$ while effectively highlighting the target model's vulnerabilities. Our method to infuse object attributes in the text can be broken into four parts: *(a)* object and attribute detection in $\mathcal{I}$, *(b)* BERT-based [MASK] prediction in $\mathcal{T}$ while ensuring *(c)* similarity of inserted tokens with detected object attributes, and *(d)* enforcing dissimilarity between modified text $\mathcal{T}'$ and $\mathcal{I}$ to obtain robustness estimates of multimodal models.

Object and Attribute Detection: For each image $\mathcal{I}$ we use a pre-trained bottom-up attention model (Anderson et al., 2018; Yu et al., 2020) to extract objects and their associated attributes. The bottom-up attention model identifies objects and corresponding attributes with a one-to-one mapping. We use these objects and attributes to modify $\mathcal{T}$, by intro-

ducing masks (i.e., the [MASK] token), in front of the mentions of the objects in $\mathcal{T}$. However, a strict matching criterion would ignore similar objects or alternatively named objects in $\mathcal{T}$. To address this, whenever the text does not have any direct object matches we use a Parts of Speech (PoS) tagger to identify nouns that could represent objects in the image. These identified nouns are compared to objects using cosine similarity between the word embeddings. If the cosine similarity between a noun $\mathcal{T}$ and a detected object in $\mathcal{I}$ is above some threshold, t, then a [MASK] token is placed before that noun. Overall, this step ensures that insertions are made only for objects in $\mathcal{T}$ that are seen in the corresponding image $\mathcal{I}$ to obtain $\mathcal{T}'$.

Mask Prediction: Next, we aim to fill in the [MASK] tokens with contextually relevant object attributes. To do so, we use the pre-trained language model BERT (Devlin et al., 2019). We sample top-k predictions from the BERT model based on probability scores that also meet the following criteria: the predicted word should not be a stop word and should not exist in the 3-hop neighborhood of the current [MASK]. Furthermore, since $\mathcal{T}$ may contain more than one [MASK] token, we carry out this process sequentially for each [MASK] to utilize newly introduced contexts. Following this process, we obtain k candidate insertions that are contextually relevant for each of the identified objects in $\mathcal{T}$ that also exists in $\mathcal{I}$. In the next step, to maintain cross-modal relevance, we consider the similarity of these candidate attributes with the attributes detected in $\mathcal{I}$.

Attribute Similarity: To better select a word for a specific mask that aligns well with information in $\mathcal{I}$, we only consider predicted tokens similar to the attributes of the associated object detected in $\mathcal{I}$. In order to do so, we utilize embedding-based similarities between each predicted token and the detected attribute string. The image attributes can describe the shape, size, color, or other characteristics (like '*floral* dress') of detected objects.

Cross-Modal Dissimilarity for Estimating Robustness: Finally, to estimate the robustness of multimodal models, we explicitly include a component that encourages dissimilarity in the embeddings of the image $\mathcal{I}$ and the modified text $\mathcal{T}'$. For each possible modified text $\mathcal{T}'$, we compute the cosine distance between its embedding obtained using the CLIP model (Radford et al., 2021) and that of the corresponding image's CLIP embedding.

While the mask prediction and attribute similarity steps ensure that the attribute insertions are semantically meaningful and maintain cross-modal relevance, the cross-modal dissimilarity between $\mathcal{T}'$ and $\mathcal{I}$ ensures that we leverage the vulnerabilities in the encoding mechanism of multimodal models. We use CLIP as the encoder for this step as it is a strong representative of the state-of-the-art vision-language models.

Text Augmentation Strategy: We now choose the final augmentation of $\mathcal{T}$ by combining the above four components —— object and attribute detection, mask prediction, attribute similarity, and cross-modal dissimilarity for estimating robustness. After placing [MASK] tokens in front of the identified objects mentions or similar nouns in $\mathcal{T}$, we consider the top-k BERT predictions for each of the [MASK] words, denoted by $w_i \, \forall \, i \in \{1, \ldots, k\}$. We take the predicted probability scores of these k words and normalize them to sum to one, denoting each by p_i. The attribute similarity step computes the similarities for w_i with the corresponding attribute, which are then normalized to sum to one and denoted by s_i. Finally, we create k augmentations of $\mathcal{T}$, each denoted by $\mathcal{T}'_i$, and compute the cosine distance of their CLIP embeddings with that of the corresponding image $\mathcal{I}$. The distances are also normalized to sum to one and denoted by d_i. Mathematically, the cumulative score for a predicted word w_i is given as,

$$\mathcal{S}_{w_i} = \lambda_1 \cdot p_i + \lambda_2 \cdot s_i + \lambda_3 \cdot d_i \qquad (1)$$

where, λ_1, λ_2, and λ_3 are hyper-parameters that control the contribution of mask prediction using BERT, attribute similarity, and cross-modal dissimilarity, respectively. The word w_i with the maximum score $\mathcal{S}$ is the word that is inserted in the place of the [MASK]. For text with multiple [MASK] tokens, we repeat this process iteratively in the order of their occurrence in $\mathcal{T}$.

By design, our cross-modal attribute insertion approach is modular as different components serve complementary functions toward the final objective of introducing semantically meaningful augmentations. It is also controllable using hyper-parameters λ_1, λ_2, λ_3, k, and t. Finally, our approach is training-free and, therefore, can be applied to investigate several tasks and models.

4 Experiments

We study the effect of cross-modal attribute insertions on two multimodal tasks: text-to-image retrieval (i.e., retrieving images for a textual description) and cross-modal entailment (i.e., predicting the relationship between textual hypothesis and visual premise).

4.1 Text → Image Retrieval

Task: Given a set of text and image pairs as input, the goal is to retrieve the associated image for each text. The retrieval occurs for each text over a set of images, in our case we use a subset of 1000 text-image pairs, with the objective being to rank the original/ground-truth image the highest.

Axiom for Retrieval: Given an image $\mathcal{I}$ in the search repository and two search queries Q_1 and Q_2, such that Q_1 contains more specific details of objects than Q_2, $\mathcal{I}$ should be retrieved at the same or higher rank for query Q_1 than for Q_2.

Dataset: For this task, we use the MSCOCO dataset (Lin et al., 2014), which contains images-caption pairs. Specifically, we use the image-caption pairs from 2017's validation split. This subset of the data contains $5,000$ unique images and $25,010$ captions, where each image can have multiple captions. To assess robustness, we perform augmentations on $25,010$ captions one by one while ranking all the images for each caption.

Model Under Investigation: For this task we consider the CLIP model (`ViT-B/32`) (Radford et al., 2021). CLIP is pretrained on 400 million image-caption pairs using contrastive learning, resulting in image and text encoders that produce unimodal embeddings that lie in a common latent space. The CLIP model has demonstrated great generalizability to various downstream tasks, including zero-shot and few-shot image classification and cross-modal retrieval. We obtain the CLIP embeddings for each image and caption in the MSCOCO dataset and rank all the images for a given caption based on their cosine similarities. We then contrast the ranking performance of the CLIP model using the original and augmented captions as textual queries.

4.2 Cross-Modal Entailment

Task: Cross-modal entailment aims to determine whether the relationship between a visual premise and a textual hypothesis is 'entailment,' 'contradiction,' 'neutral.' Specifically, 'entailment' is observed when the textual hypothesis is logically implied (true) by the image while 'contradiction' indicates that the textual hypothesis is not implied (false) by the visual premise. Finally, 'neutral' represents an inconclusive or uncertain relationship between the hypothesis and the premise.

Axiom for Entailment: If the relationship between a visual premise and a textual hypothesis is 'entailment,' it should not change to 'contradictory' if the textual hypothesis is enriched with the information from the visual modality.

Dataset: We perform this task on SNLI-VE (Xie et al., 2019), a visual entailment dataset. We use the test set of the dataset, comprising $17,859$ image (premise) & text (hypothesis) pairs. For robustness assessment, we augment all text hypotheses while keeping the visual premise the same.

Model Under Investigation: We investigate the pre-trained METER model (Dou et al., 2022), which consists of vision-and-language transformers that are trained end-to-end. The model's comprises CLIP's vision encoder and RoBERTa (Liu et al., 2019) as the text encoder. The model pretraining objectives consist of masked language modeling and image-text-matching on four datasets: MSCOCO, Conceptual Captions (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011), and Visual Genome (Krishna et al., 2017). In addition, METER is fine-tuned on SNLI-VE in order to achieve competitive performance. We contrast the performance of the model on the unmodified SNLI-VE dataset with the performance on the augmented version of SNLI-VE dataset.

4.3 Baselines for Perturbations

We compare our cross-modal attribute insertion approach (XMAI) with competitive baselines that are capable of introducing perturbations based on text-only information. We utilize the TextAttack (Morris et al., 2020)[1] framework for implementing all the baseline perturbation strategies.

Deletion: A perturbation strategy that randomly removes words from the text.

EDA (Wei and Zou, 2019): This approach combines random deletion, random swapping, random insertion, and synonym replacement to modify each caption. We keep all parameters as default and set the percentage of words to swap to 20%.

CheckList (Ribeiro et al., 2020): Developed to generate a diverse set of evaluation examples, CheckList works by coalescing name replacement,

---

[1] https://github.com/QData/TextAttack

location replacement, number alteration, and word contraction/extension.

CLARE (Li et al., 2021a): This perturbation strategy uses language models to replace, insert, and merge tokens in captions. We use TextAttack's default fast implementation of CLARE.

4.4 XMAI Implementation Details

We choose $k = 3$ for the number of top-k predicted BERT words for each `[MASK]` token and `flair/pos-english-fast` for PoS tagging of text. Next, to compare the nouns in the text with the objects identified in the image, we use word embeddings produced by a Transformer-based model (`bert-base-nli-mean-tokens` on HuggingFace (Wolf et al., 2020)). We set the threshold, t, for cosine similarity between nouns in $\mathcal{T}$ and objects in $\mathcal{I}$ to be 0.7. For `[MASK]` filling, we use the `bert-base-cased` model on HuggingFace and the list of stopwords is adopted from NLTK. [2] To compute the similarity between attributes detected in $\mathcal{I}$ and BERT predictions, we employ SpaCy's pretrained tok2vec model (`en_core_web_md`), which contains 300-dimensional embeddings for $\sim 500k$ words (Honnibal et al., 2020). Lastly, the pre-trained CLIP model (`ViT-B/32`) is used to compute image and text embeddings in a common latent space. For our main experiments, we set the values of λ_i as $\lambda_1 = 1$, $\lambda_2 = 5$, and $\lambda_3 = 5$.

4.5 Evaluation Metrics

We measure the impact of perturbations in the text on the capabilities of multimodal models using task-specific metrics. We quantify text-to-image retrieval performance using mean reciprocal rank (MRR). For cross-modal entailment, we report standard classification metrics (accuracy, precision, recall, and F_1 score).

While the effectiveness of perturbations is important for highlighting model vulnerabilities, it is also imperative to measure the relevance of augmented text $\mathcal{T}'$ with original text $\mathcal{T}$ and image $\mathcal{I}$. To this end, we compute mean cosine similarity $Sim_{\mathcal{T}-\mathcal{T}'}$ between original and modified texts (i.e., $\mathcal{T}$ & $\mathcal{T}'$, respectively) using a sentence Transformer model (`all-mpnet-base-v2`) (Reimers and Gurevych, 2019). Similarly, we report BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), using NLTK to further compare the

---

[2] https://www.nltk.org/

texts (considering n-grams of size upto 4 in the case of BLEU). Additionally, we compute the mean cosine similarity $Sim_{\mathcal{I}-\mathcal{T}'}$ using CLIP (`ViT-B/32`) embeddings.

5 Results and Analysis

Recall that our primary goal is to use XMAI to obtain a complementary and novel set of text augmentations that can highlight the vulnerabilities of multimodal models. To this end, we contrast the performance of the models under investigation on the original and the modified examples, and quantify the relevance of the modified text with respect to the original text and image. We recruit human annotators to compare the quality of the augmentations generated using our approach with (i) the ones generated using the most competitive baseline, and (ii) the original text.

Robustness of Multimodal Models: Table 1 shows that for the text $\rightarrow$ image retrieval task, our cross-modal attribute insertions cause the greatest drop in observed MRR; the MRR drops from an original value of 0.632 to 0.536. Similarly, Table 2 shows that for the cross-modal entailment task our approach performs second only to CLARE — an observation that is consistent across all the metrics, accuracy, precision, recall, and F_1. It is worth noting that our approach is the only one that uses the information from the image modality to introduce textual perturbations and hence the resultant perturbed examples are characteristically different than the ones that are produced using baseline methods like CLARE. We will revisit this using qualitative examples. Overall, results demonstrate that state-of-the-art vision-and-language learning approaches for text-to-image retrieval and cross-modal entailment tasks are not robust to our proposed augmentations.

Relevance of Augmentations: Tables 1 and 2 show that XMAI produces augmentations $\mathcal{T}'$ that maintain high-relevance with the original text $\mathcal{T}$ and the image $\mathcal{I}$, in terms of $Sim_{\mathcal{T}-\mathcal{T}'}$ and $Sim_{\mathcal{I}-\mathcal{T}'}$. It is interesting to note that the BLEU scores for augmentations generated by XMAI are notably lower than that for the baseline augmentations. On the contrary, METEOR scores show that XMAI's augmentations are "closer" to the original texts compared to most baselines. XMAI's poor BLEU scores can be largely attributed to BLEU's tendency to penalize novel insertions severely compared to removals or replacements, as it is a precision-

Approach	MRR $\downarrow$	$\mathcal{S}im_{\mathcal{T}-\mathcal{T'}}\uparrow$	$\mathcal{S}im_{\mathcal{I}-\mathcal{T'}}\uparrow$	BLEU $\uparrow$	METEOR $\uparrow$
Original	0.632	1.000	0.294	1.000	1.000
Deletion	0.581	<u>0.948</u>	<u>0.288</u>	<u>0.758</u>	0.910
EDA	0.570	0.914	0.287	0.647	0.931
Checklist	0.630	**0.996**	**0.294**	**0.982**	**0.993**
CLARE	<u>0.567</u>	0.899	0.285	0.749	0.947
XMAI (Ours)	**0.536**	0.924	0.283	0.623	<u>0.969</u>

Table 1: **Results on the text-to-image retrieval task.** The effectiveness of augmentation in highlighting the model's vulnerability is noted by the drop in MRR with respect to the original MRR score. $\mathcal{S}im_{\mathcal{T}-\mathcal{T'}}$, BLEU, and METEOR capture the relevance of augmented text with the original text and $\mathcal{S}im_{\mathcal{I}-\mathcal{T'}}$ captures the relevance of the augmented text with the original text. Best results are in **bold**; second best are <u>underlined</u>.

Approach	Acc. $\downarrow$	Precision $\downarrow$	Recall $\downarrow$	F$_1$ $\downarrow$	$\mathcal{S}im_{\mathcal{T}-\mathcal{T'}}\uparrow$	$\mathcal{S}im_{\mathcal{I}-\mathcal{T'}}\uparrow$	BLEU $\uparrow$	METEOR $\uparrow$
Original	0.792	0.790	0.792	0.791	1.000	0.246	1.000	0.998
Deletion	0.742	0.742	0.742	0.740	<u>0.892</u>	0.240	<u>0.632</u>	0.853
EDA	0.719	0.719	0.719	0.718	0.878	<u>0.241</u>	0.541	0.900
Checklist	0.791	0.790	0.791	0.790	**0.973**	**0.246**	**0.993**	**0.986**
CLARE	**0.632**	**0.652**	**0.631**	**0.618**	0.804	0.238	0.592	0.911
XMAI (Ours)	<u>0.643</u>	<u>0.682</u>	<u>0.643</u>	<u>0.625</u>	0.873	0.235	0.621	<u>0.963</u>

Table 2: **Results on the cross-modal entailment task.** Augmentations that cause a greater drop in classification metrics are more effective at highlighting the lack of multimodal robustness, while the similarity metrics capture their relevance with the original example. The best results are in **bold** and the second best results are <u>underlined</u>.
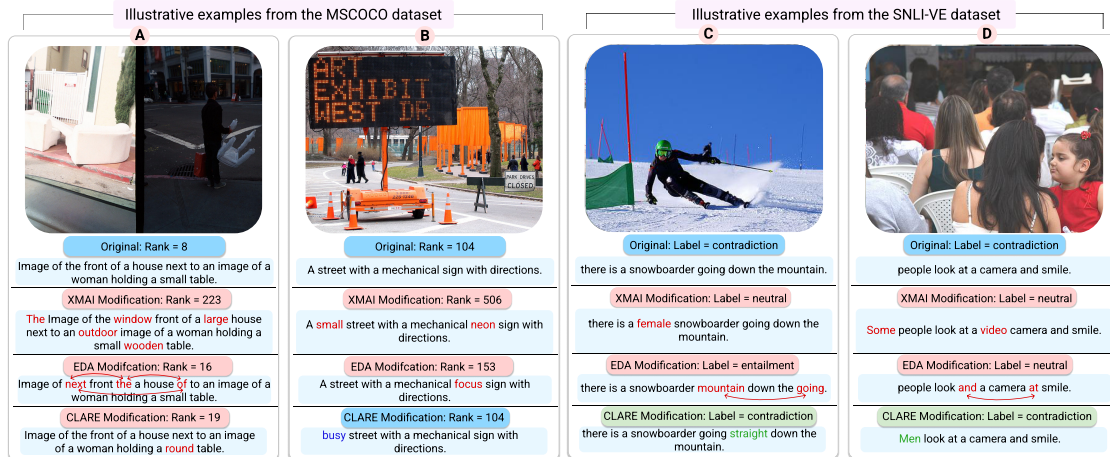

Figure 3: Qualitative examples comparing the augmentations produced by our XMAI method to both EDA and CLARE on both MSCOCO and SNLI-VE tasks. Red text represents a drop in rank or misclassification, green text indicates improvement in rank or correct classification, and blue marks when a change has no impact on rank. Lastly, arrows and the words at either end of each arrow indicate swapping by EDA.

based metric (Banerjee and Lavie, 2005).[3] In Appendix A.3 (Table 4), we further note that, on average, XMAI inserts 1.660 ($\pm$0.947) new words in MSCOCO captions and 1.269 ($\pm$0.768) new words in SNLI-VE hypotheses. This is considerably higher than the rate of insertions made by other methods, especially Checklist, where an obvious consequence of making a fewer number of augmentations is better text similarity across a corpus.

We thus attribute the poor performance of XMAI in terms of BLEU scores to BLEU's inability to handle insertions appropriately. This is further substantiated by the human assessments.

5.1 Human Assessment of Augmentations

We recruit annotators using Amazon Mechanical Turk (AMT) to answer the following two questions: *(i)* do cross-modal attribute insertions lead to better text augmentations than the most competitive baseline (i.e., CLARE), and *(ii)* are cross-modal

---

[3]Novel insertions in $\mathcal{T'}$ mean more 'false positives' with respect to $\mathcal{T}$, indicating lower precision and BLEU scores.

attribute insertions as good as the original text accompanying the image. Please see Appendix A.1 for details on the recruitment filters and compensation on AMT.

XMAI versus CLARE: We randomly sampled 100 examples from the validation set of the MSCOCO dataset and showed the modified captions using CLARE and XMAI. 5 annotators annotated each example. We asked annotators to indicate their agreement to the following question after seeing two captions for a given image using a 5-point Likert scale (1: Strongly disagree, ..., 5: Strongly agree): `Caption 2 is a better description of the shown image than Caption 1 in terms of its quality and accuracy`. `Caption 1` and `Caption 2` were randomly flipped between CLARE and XMAI to avoid any position bias. Furthermore, to ensure quality annotations, we randomly inserted some "attention check" examples that instructed annotators to ignore all previous instructions and mark specified responses on the Likert scale. We discarded responses from annotators who marked the attention-check examples incorrectly and re-collected the annotations.

For 63% of the examples, a majority of annotators (i.e., at least 3 out of 5) preferred the captions modified using XMAI over CLARE. The captions modified using CLARE were preferred for 26% examples. The rest were either marked as 'equivalent' (i.e., 3: Neither disagree nor agree) or had ambiguous majority votes. Overall, the results demonstrate that the annotators preferred the captions modified using XMAI over the ones modified using CLARE, in terms of their accuracy in describing the image and their quality. We next assess how XMAI modified captions compare against the original captions.

XMAI versus Original: We randomly sampled 100 examples from the validation set of the MSCOCO dataset and randomly chose 50 of them to be modified using XMAI while leaving the other 50 unmodified. We first primed the annotators to view 5 original image caption pairs, noting them as reference examples.[4] We then asked the annotators to view a *list* of image-caption pairs and evaluate the caption quality using the following prompt: *Rate the caption quality for the given image based on the reference examples shown earlier.* A response of 1 on the 5-point Likert scale indicated 'extremely poor quality' whereas that of 5 indicated

'extremely good quality.' The shown list comprised randomly shuffled original image-caption pairs and modified image-caption pairs using XMAI, and a few attention-check examples. Each example received annotations from 5 different annotators.

The unmodified captions received an average score of 4.12 (± 0.37) whereas that for the modified caption using XMAI was 4.07 (± 0.33). The observed inter-rater agreement was strong, with a Krippendorf's α of 0.78. Additionally, a two-sided t-test with unequal variances assumption failed to reject the null hypothesis ($p > 0.05$) that the average Likert scores for the original and modified captions are from different distributions. In sum, the perceived quality of the modified captions using XMAI is not statistically different from that of the original captions.

Computational Efficiency: In Appendix Fig. 7 we demonstrate that our approach for inserting cross-modal attributes is $14.8\times$ and $9.4\times$ faster than the most competitive baseline approach (i.e., CLARE) on MSCOCO and SNLI-VE, respectively. Combined with the fact that XMAI augmentations are perceived to be of better quality than CLARE augmentations and are effective at highlighting model vulnerabilities, the increased computational efficiency allows for more rapid and realistic model validation. In the following section, we demonstrate via qualitative examples that, being the only approach that leverages cross-modal interactions in multimodal data, the augmentations produced by XMAI are novel to those produced by the text-only baselines.

5.2 Qualitative Analysis

In Figure 3 we show illustrative examples of the insertions introduced using our approach and contrast them with existing text-only perturbations. We observe that our cross-modal insertions lead to a complementary set of augmentations that are not covered by text-only approaches.

We note that our method does not remove any information present in the original caption/hypothesis. This prevents our method from drastically changing the original semantics, which has been a known shortcoming of text-only perturbations (Wang et al., 2021). In Figure 3(A), we note that EDA produces a grammatically incoherent augmentation ("*Image of next front the a house of to...*") and CLARE inserts an inaccurate attribute ("*round table*"). Whereas, our approach only in-

---

[4]These 5 reference examples were not included in the subset of 100 examples selected for annotations.
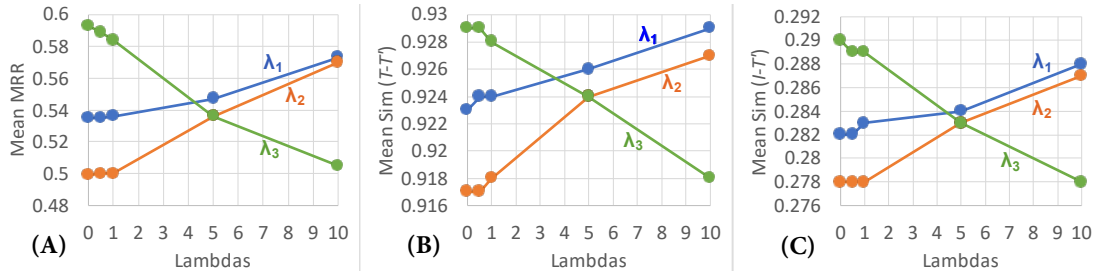
Figure 4: **Varying λ_i to isolate their effect on the text-to-image retrieval task.** Ablations on independent effects of lambda values, where the default lambdas are: $\lambda_1 = 1$, $\lambda_2 = 5$, and $\lambda_3 = 5$. Each line plot represents changing the specified λ while keeping the others as default. We observe the variation in task-specific performance as well as the similarity metrics.

serts relevant attributes to the original text ("*The Image of the window front of a large house next to an outdoor image of a woman holding a small wooden table.*"). In Figure 3(B&D) we see that XMAI modifies the text using the information in the corresponding images — for instance, our approach identifies the neon LEDs and inserts 'neon' in front of 'sign.' However, EDA and CLARE introduce inaccurate details. XMAI is also capable of multiple meaningful insertions. Our work is the first to enable cross-modal insertion capabilities to obtain meaningful augmentations of multimodal (image + text) data.

5.3 Ablations for λ_i Sensitivity

In Figure 4, we visualize the change in retrieval performance with respect to independent changes in λ_1, λ_2, and λ_3. In other words, we vary a given λ_i while keeping other hyper-parameters at their aforementioned values. We find that increasing λ_1 and λ_2 improves the relevance of augmentations but reduces their effectiveness in highlighting vulnerabilities. Intuitively, these components increase the likelihood that our approach picks insertions with high BERT prediction scores (controlled by λ_1) and similarities with the identified image attribute (controlled by λ_2). On the other hand, increasing λ_3, which controls the contributions of the robustness assessment component, generates less relevant augmentations that are highly effective. This observation also aligns with our goal of exploiting the lack of robust encoding mechanisms to highlight model vulnerabilities.

Overall, these results demonstrate that the individual components of our approach play significant roles, and can be controlled using the λ_i hyper-parameters. Similar trends are observed for the cross-modal entailment task; see Appendix Fig. 6.

We discuss the ablations pertaining to the similarity threshold for matching image objects and text nouns in Appendix A.2.

6 Conclusion

A *robust* understanding of vision-and-language data is crucial for powering several applications. We propose cross-modal attribute insertions – i.e., adding the visual attributes to text, as a new variation that is likely in multimodal data and to which multimodal models should be robust. Our approach produces novel augmentations that are complementary to existing methods that model text-only data, and are preferred over them by human annotators. Using our augmentations we effectively highlight the vulnerabilities of state-of-the-art multimodal models for text-to-image retrieval and cross-modal entailment. In the future, we aim to empirically study the effect of including XMAI augmented data in task-specific training sets and expand to a broader set of multimodal tasks and metrics.

7 Limitations and Broader Perspective

Limitations and bias of pre-trained models: Our work uses detected objects and their attributes in the images to introduce novel insertions in the corresponding text. To this end, it is important to address the limitations of the state-of-the-art object and attribute detection methods. The undesired artifacts of these methods could be categorized as inaccurate or biased. The detected objects could be incorrect, but since we only consider objects that are also mentioned in the text, the effect of incorrect object detections is non-existent in our augmentations. However, we notice that some of the detected attributes in images and BERT predictions reflect stereotypical associations and have been documented in prior works (Li and Xu, 2021;

Kaneko and Bollegala, 2022). We acknowledge that the current state of deep learning research is limited, and the consequential shortcomings are reflected in our augmentations to some extent.

Broader social impact: The authors do not foresee any negative social impacts of this work. We believe our cross-modal augmentations will enable an exhaustive evaluation of the robustness of vision-and-language models, leading to more reliable multimodal systems. We release the code for our experiments to aid reproducibility and enable future research on this topic.

Annotations, IRB approval, and datasets: The annotators for evaluations done in this study were recruited via Amazon Mechanical Turk. We specifically recruited 'Master' annotators located in the United States; and paid them at an hourly rate of 12 USD for their annotations. The human evaluation experiments were approved by the Institutional Review Board (IRB) at the authors' institution. The datasets used in this study are publicly available and were curated by previous research. We abide by their terms of use.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809.

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

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Ivan Evtimov, Russel Howes, Brian Dolhansky, Hamed Firooz, and Cristian Canton Ferrer. 2020. Adversarial evaluation of multimodal models under realistic gray box assumption. *arXiv preprint arXiv:2011.12902*.

Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, and Chunhong Pan. 2016. Cross-modal retrieval via deep and bidirectional representation learning. *IEEE Transactions on Multimedia*, 18(7):1363–1377.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. GENIE: A leaderboard for human-in-the-loop evaluation of text generation. *CoRR*, abs/2101.06561.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021a. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021b. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2042–2051.

Zhiheng Li and Chenliang Xu. 2021. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14970–14979.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13516–13524.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *The 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2022. Are multimodal models robust to image and text perturbations? *arXiv preprint arXiv:2212.08044*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213, Dublin, Ireland. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359.

Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. 2021. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599.

Gaurav Verma, Vishwa Vinay, Ryan Rossi, and Srijan Kumar. 2022. Robustness of fusion-based multimodal classifiers to cross-modal content dilutions.

In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 360–374, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Advances in Neural Information Processing Systems*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Zhou Yu, Jing Li, Tongan Luo, and Jun Yu. 2020. A pytorch implementation of bottom-up-attention. https://github.com/MILVLG/bottom-up-attention.pytorch.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

A   Appendix

A.1   Human Evaluation Details

For our annotation tasks, we recruited annotators using Amazon Mechanical Turk. We set the criteria to 'Master' annotators who had at least 99% approval rate and were located in the United States. The rewards were set by assuming an hourly rate of 12 USD for all the annotators. In addition, the annotators were informed that the aggregate statistics of their annotations would be used and shared as part of academic research.

Previous research has demonstrated the role of providing priming examples in obtaining high-quality annotations (Khashabi et al., 2021). Therefore, we showed unmodified examples from the MSCOCO corpus to help annotators establish a reference for quality and accuracy. For both the crowd-sourced evaluations, we inserted some "attention-check" examples to ensure the annotators read the text carefully before responding. This was done by explicitly asking the annotators to mark a randomly-chosen score on the Likert scale regardless of the actual content. We discard the annotations from annotators who did not correctly respond to all the attention-check examples.

A.2   Further Ablations

Extended Variations in t: To illustrate the effect of changing the threshold, t, we plot our described MSCOCO metrics with respect to variations in t in Figure 5. We see that as the criterion for matching nouns in the text and objects in the image is made more stringent, the quality and relevance of the augmentations improve further. However, the effectiveness of the resulting augmentations in highlighting model vulnerabilities decreases. It is worth noting as the model becomes more selective in inserting attributes (due to fewer matched nouns and objects), we witness a stark increase in BLUE scores. Variations in t effectively capture the trade-off between maintaining the relevance of augmentations and effectively highlighting model vulnerabilities. Even though it is possible to construct augmentations that will be more effective in making the multimodal models perform poorly, it would sacrifice the relevance and quality of resulting augmentations. Our main results demonstrate that with $t = 0.7$, we obtain high-quality and human-preferred augmentations that are also effective in highlighting vulnerabilities.

Variations in k: We perform another ablation by increasing the value of k for the top-k predictions made by pre-trained BERT model. Table 3 shows that increasing the search space for possible insertion tokens leads to a notable drop in the retrieval
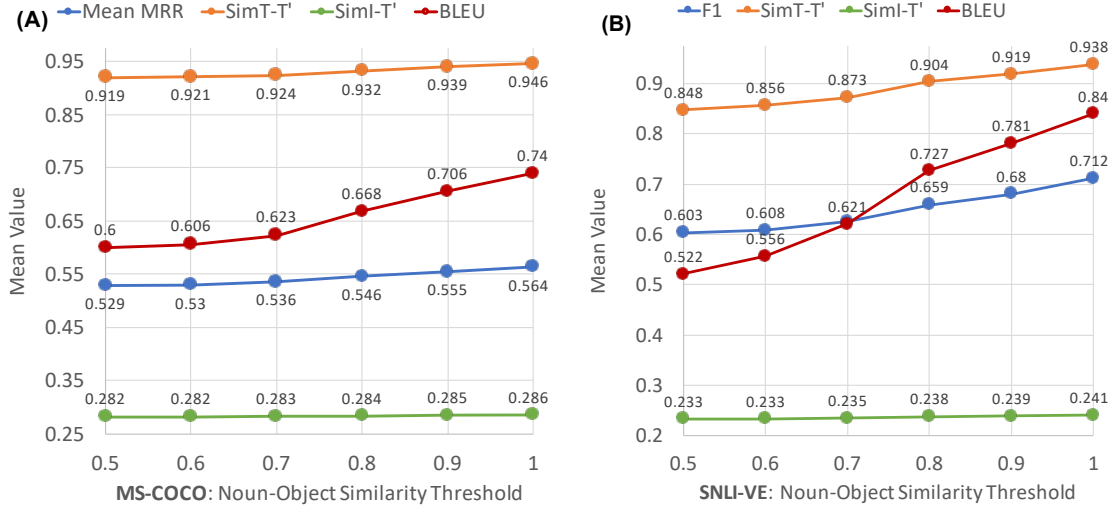
(a) Ablation on the noun-object similarity threshold t for the text-to-image retrieval task (MSCOCO).

(b) Ablation on the noun-object similarity threshold t for the cross-modal entailment task (SNLI-VE).

Figure 5: Ablations on threshold, t, across task-specific metrics for both tasks.

k	MRR $\downarrow$	$Sim_{\mathcal{T}-\mathcal{T}'}$ $\uparrow$	$Sim_{\mathcal{I}-\mathcal{T}'}$ $\uparrow$	BLEU $\uparrow$	METEOR $\uparrow$	Time (s) $\downarrow$
3	0.536	0.924	0.283	0.623	0.969	0.196
5	0.506	0.919	0.278	0.623	0.970	0.311
10	0.476	0.914	0.274	0.623	0.970	0.474

Table 3: Ablations by varying the number of BERT predictions considered (i.e., k) for the text-to-image retrieval task on MSCOCO. The reported time is in seconds (per caption).

Method	MSCOCO Captions	SNLI-VE Hypotheses
	μ ($\pm\sigma$)	μ ($\pm\sigma$)
Deletion	0.000 ($\pm$ 0.022)	0.000 ($\pm$ 0.015)
EDA	0.679 ($\pm$ 0.784)	0.562 ($\pm$ 0.633)
Checklist	0.077 ($\pm$ 0.284)	0.087 ($\pm$ 0.300)
CLARE	0.982 ($\pm$ 0.161)	0.990 ($\pm$ 0.132)
XMAI (Ours)	1.660 ($\pm$ 0.947)	1.269 ($\pm$ 0.768)

Table 4: Mean (and standard deviation) of the number of novel insertions/replacements in modified MSCOCO captions and SNLI-VE hypotheses with respect to their original counterparts.

Method	MSCOCO Captions		SNLI-VE Hypotheses	
	r/i	*any*	*r/i*	*any*
Deletion	12	25,010	4	17,857
EDA	12,492	24,980	8,876	17,786
Checklist	1,817	1,879	1,464	1,492
CLARE	24,456	24,972	17,619	17,817
XMAI (Ours)	24,244	24,970	16,275	16,512

Table 5: Number of examples that are modified in MSCOCO and SNLI-VE. As mentioned in the Experiments section, we consider 25,010 captions for MSCOCO and 17,859 hypotheses for SNLI-VE. We further split this by considering modifications as only novel replacements/insertions (*r/i*) or any difference between original and modified (*any*).

performance over resulting augmentations. However, the relevance values with the original image and text drop too. Increasing the search space allows the model to explore potential insertions that

Ablation	λ_1	λ_2	λ_3	k	t	MRR $\downarrow$	$\mathcal{Sim}_{\mathcal{T}-\mathcal{T}'}\uparrow$	$\mathcal{Sim}_{\mathcal{I}-\mathcal{T}'}\uparrow$	BLEU $\uparrow$	METEOR $\uparrow$
Original	1	5	5	3	0.7	0.536	0.924	0.283	0.623	0.969
λ_2	0	1	0	3	0.7	0.593	0.929	0.290	0.623	0.970
	0	-1	0	3	0.7	0.588	0.922	0.290	0.623	0.968
λ_3	0	0	1	3	0.7	0.499	0.917	0.277	0.623	0.969
	0	0	-1	3	0.7	0.672	0.932	0.301	0.623	0.969

Table 6: Effect of variations in λ_2, λ_3, and k for the text-to-image retrieval on MSCOCO.
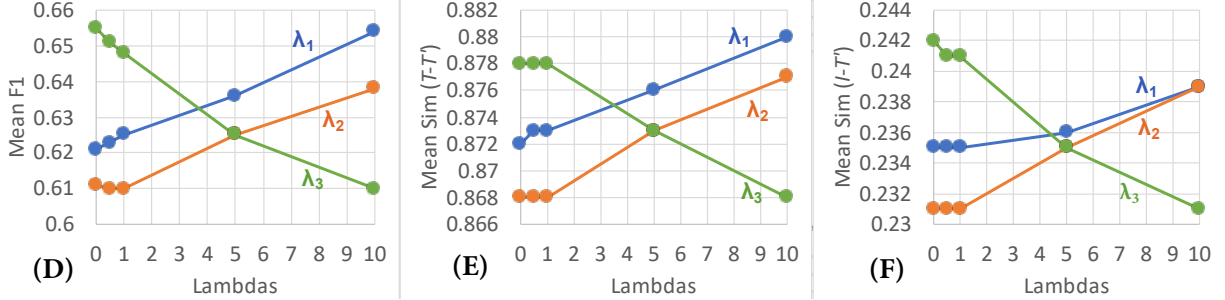

(D) **(E)** **(F)**

Figure 6: **Varying λ_i to isolate their effect on the cross-mdodal entailment task.** Ablations on independent effects of lambda values, where the default lambdas are: $\lambda_1 = 1$, $\lambda_2 = 5$, and $\lambda_3 = 5$. Each line plot represents changing the specified λ while keeping the others as default. We observe the variation in task-specific performance as well as the similarity metrics.

could produce highly dissimilar cross-modal representations, thereby helping the adversarial component of our framework, but with the compromise being the relevancy of the augmentations. We also note that exploring more insertion possibilities increases the per-caption augmentation time taken by XMAI.

Varying λ_2 and λ_3: To comprehensively understand the effect of variations in λ_i, specifically λ_2 and λ_3, we set each to 1 or -1 (one at a time) while setting the other two lambdas to 0. Results in Table 3 empirically show that both the attribute similarity and robustness assessment components are essential and can serve dual purposes; i.e., negative values of λ_2 and λ_3 allow the associated components to serve the opposite purpose. In contrast to their original objective, when the associated λ values are set to negative values, attribute similarity decreases performance and robustness assessment increases it.

A.3 Number of Insertions

We report the number of insertions or replacements each augmentation method makes to the original text as well as the number of texts modified for each dataset. The results are reported in Table 4 and 5. We find that our XMAI approach introduces more novel words in the augmentation than any other approach, while also augmenting nearly

the same amount of captions as most competitive baseline approaches. This observation, combined with the fact that human annotators prefer XMAI augmentations over baseline augmentations, shows that cross-modal insertions can be used to introduce new information without causing semantic deterioration of the text. Additionally, these results allow us to attribute the low BLEU scores observed in Table 1 to the higher number of insertions that XMAI makes. Note that BLEU computation is precision-based and hence penalizes novel insertions more severely. It is interesting to note that even though 'Deletion' is expected to have no insertions or replacements, we found that in very few cases, due to the adopted implementation as well as the noise in the text could result in fragmented or fused words that were being considered novel compared to the original text and therefore counted as an insertion/replacement.

A.4 Augmentation Time

In Figure 7 we show the average time to augment the input text for each of the methods. The results are plotted using a logarithmic scale to ensure a clearer depiction across methods that vary considerably in terms of computational time.

Simpler approaches such as Deletion, EDA, and CheckList can modify tens or hundreds of samples each second. Intuitively, the reliance on simple
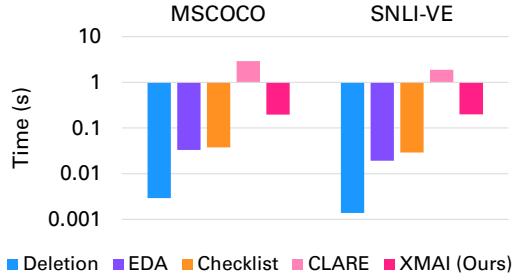
Figure 7: Comparing the per-caption augmentation time of various methods across the two tasks. Results are shown in logarithmic scale due to the disparity between computational times across different methods.

rules makes these approaches very efficient. On the other hand, context-aware methods such as CLARE and XMAI are slower due to their reliance on large language models and involved selection processes. However, XMAI can augment the text a magnitude faster than CLARE, even after using the fast implementation of CLARE from TextAttack.

We don't consider the time to compute objects and attributes for XMAI for two reasons. First, the cost to perform this step was < 1 hour for our datasets and the relationship between methods remains the same. Secondly, objects and attributes only need to be computed once, so there is no additive cost for augmentation unless changes are made to the detection component.

A.5 Compute Resources

Our experiments were split between a single Tesla V100 for object detection (~ 1 hour) and NVIDIA Tesla T4 GPUs for our augmentation (~ 3 hours).

Algorithm 1: Algorithmic block describing the text augmentation method for XMAI. For details reference back to and follow along with Section 3.

---

1   `// ▷ Cross-Modal Attribute Insertions`

    Input   : An image-text pair denoted by $(\mathcal{I}, \mathcal{T})$

    Output : Augmented text $\mathcal{T}'$.

2   `// ▷ Object and Attribute Detection`

3   Detect objects and attributes in $\mathcal{I}$

4   Introduce masks into $\mathcal{T}$ where direct matches exist

5   If no direct matches, use word similarity b/w detected objects in $\mathcal{I}$ & nouns in $\mathcal{T}$

6   `// ▷ Mask Prediction`

7   **for** $i = 1, ..., N_{[MASK]}$ **do**

8   Use BERT to obtain top-k predictions for current mask

9   For current mask, maintain probability score vector p

10   `// ▷ Attribute Similarity`

11   **for** $j = 1, ..., k$ **do**

12   Compute maximum attribute similarity between relevant object attributes and the current predicted word, s_j

13   **end for**

14   `// ▷ Cross-Modal Dissimilarity for Estimating Robustness`

15   Create k candidate augmentations

16   Compute CLIP dissimilarity for each candidate augmentation, $d_1, ..., d_k$

17   `// ▷ Text Augmentation Strategy`

18   Compute the final score vector, $\mathcal{S}_w$

19   Insert word with maximum score in $\mathcal{S}_w$ in place of current `[MASK]`

20   **end for**

21   Output text $\mathcal{T}'$ with insertions