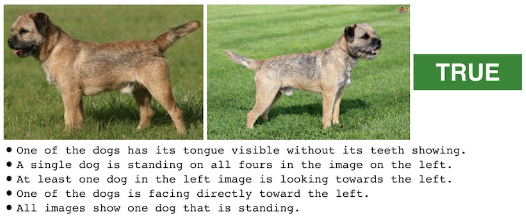
MODEL CITIZENS: FINAL REPORT

Angela Qu[*] **Changwook Shim**[*] **Nikhil Reddy**[*] **Tyler Ho**[*] **Yuling Wang**[*]
{angelaqu, cshim, nthambal, tylerho, yulingwa}@andrew.cmu.edu

ABSTRACT

We propose an embedding-level adversarial training approach using FGSM to enhance the robustness of ViLT on the NLVR2 visual reasoning benchmark. By directly perturbing visual and textual embeddings before transformer encoding, our method improves both accuracy and resilience to adversarial and natural perturbations without significant computational overhead. Empirical evaluation confirms consistent improvements over baseline models on NLVR2.

1 [2 POINTS] INTRODUCTION AND PROBLEM DEFINITION (1-1.25 PAGES)



- One of the dogs has its tongue visible without its teeth showing.
- A single dog is standing on all fours in the image on the left.
- At least one dog in the left image is looking towards the left.
- One of the dogs is facing directly toward the left.
- All images show one dog that is standing.

Figure 1: Five sample captions (all true) for a pair of images

Natural language visual reasoning requires models to deeply understand the relationship between textual descriptions and complex visual scenes. The NLVR2 benchmark Suhr et al. (2019) challenges models to determine whether a statement correctly describes a pair of images, requiring alignment of linguistic clauses with visual elements. Recent transformer-based approaches, such as ViLT Kim et al. (2021), have demonstrated that vision-and-language encoders operating directly on patch embeddings and token embeddings without heavy object-detection backbones can achieve competitive performance with remarkable efficiency. However, like many deep models, ViLT remains vulnerable to perturbations, which can undermine its reasoning capabilities and generalization to out-of-distribution examples.

In this paper, we propose to apply FGSM proposed by Goodfellow et al. (2015) directly to the embedding layer of ViLT, perturbing both patch embeddings and token embeddings before they enter the transformer encoder. By integrating adversarial noise at this earliest fusion point, our method encourages the model to learn more stable cross-modal alignments and to resist subtle input variations. We evaluate our approach on NLVR2 and show that embedding-level adversarial training yields consistent improvements in accuracy and robustness. Our contributions are two fold:

1. Embedding-Level Adversarial Framework: We adapt FGSM to perturb visual and textual embedding layers in ViLT, creating multimodal adversarial examples at the fusion interface.
2. Empirical Analysis on NLVR2: We demonstrate that embedding layer level adversarial training improves accuracy on natural and adversarial perturbations, without having to rerun full fine-tuning for the task.

[*] Everyone Contributed Equally – Alphabetical order

2 [5 POINTS] RELATED WORK AND BACKGROUND (5 PAPERS PER PERSON)

2.1 RELATED DATASETS

Vision-language models have largely relied on massive, automatically collected image-text corpora for pretraining and fine-tuning. Many models that are competitive across a wide variety of visual reasoning tasks rely on a combination of multiple such datasets. For example, ViLT (introduced in greater detail in the **Prior Work** subsection) uses the following datasets for pretraining:

- Microsoft COCO (Lin et al., 2015): Consists of over 330k images with fives crowd-sourced captions each
- Visual Genome (Krishna et al., 2016): Consists of over 100k images with dense annotations of objects, with each image annotated with an average of 21 objects, 18 attributes, and 18 pairwise relationships.
- SBU Captions (Ordonez et al., 2011): Consists of over 1 million images sourced from Flickr along with alternative captions.
- Conceptual Captions (Sharma et al., 2018): Consists of around 3.3M image-caption pairs that represent a wider variety of images and caption styles than MS COCO.

ViLT is evaluated across 4 downstream vision-and-language downstream tasks, requiring finetuning on the following datasets:

- VQA v2.0 (Goyal et al., 2017): Consists of over 1M question-image pairs. The dataset is designed so that each questions is associated with a pair of similar images that result in two different answers. It is used to evaluation classification.
- Flickr30k Entities (Plummer et al., 2016): Augments the 158k captions from the Flickr30k dataset with 244k coreference chains and 276k manually annotated bounding boxes. It is used to evaluation classification.
- MS COCO (Lin et al., 2015): Introduced previously. It is used to evaluate retrieval.
- NLVR2 (Suhr et al., 2019): It is used to evaluate retrieval.

NLVR2 is uniquely challenging among vision-language datasets. It consists of pairs real-life photographs drawn from diverse web sources along with crowd-sourced free-form English captions describing relationships across the two images. This causes the images to be variable and noisy, and the captions to be linguistically diverse and unconstrained. Each caption is paired with four image pairs where two result in the caption being true while two result in the caption being false. In total, there are over 100k image-pair-caption examples.

NLVR2 marked a step up in difficulty from its direct predecessor, NLVR (Suhr et al., 2017), which consisted of synthetic images instead of real-life photographs. The highly related CLEVR dataset (Johnson et al., 2016) consists of over 100k synthetic 3D scenes, and 700k automatically generated captions. This automatic generation of captions sets it apart from NLVR and NLVR2. Each training example in CLEVR consists of only 1 image, but instead evaluating the truth of the caption, the CLEVR task is multi-way question answering. Other competitive models such as UNITER are pretrained on the CLEVR dataset before being fine-tuned on NLVR2.

2.2 UNIMODAL BASELINES

In order to assess the effectiveness of different unimodal approaches on the NLVR2 task, we implemented and evaluated four baseline models. These baselines draw upon established architectures in natural language processing and computer vision, allowing us to evaluate how invidual modalities perform when reasoning over complex multimodal data.

CNN O'Shea & Nash (2015) : Convolutional Neural Networks (CNNs) have been foundational in image processing. In our implementation, we used MobileNetV2, a modern lightweight CNN architecture, for extracting visual features from the NLVR2 image pairs. We used the pretrained IMAGENET1K V1 weights and fine-tuned the model on the NLVR2 dataset.

BERT Devlin et al. (2019) : Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model that captures deep contextual relationships in text through masked language modeling and next sentence prediction. For our text-only baseline, we used TinyBERT, a compressed version optimized for efficiency. We fine-tuned it on NLVR2 to predict the truth value of captions based on textual understanding alone.

GPT-4 OpenAI et al. (2024) : GPT-4 is a generative transformer model that achieves strong performance across a wide range of natural language tasks. For this baseline, we used GPT-4o-mini, prompting it to predict a truth value given a natural language caption.

RNN Sherstinsky (2020) : Recurrent Neural Networks (RNNs) have been widely used for sequence modeling tasks. In our experiments, we trained an LSTM on caption text to predict the truth value, using it as a text-only baseline for the NLVR2 task.

2.3 Prior Work

Vision-and-language reasoning task, exemplified by NLVR2, has driven significant progress in multimodal learning research. Several key models have laid the foundation.

UNITER Chen et al. (2020) introduced a transformer framework for pretraining on multiple vision-language tasks, achieving strong results on NLVR2 by learning fine-grained alignments between word tokens and image regions.

LXMERT Tan & Bansal (2019) similarly adopted a transformer approach, with separate encoders for language and vision, before fusing information via cross-modal layers. It emphasized reasoning and improved the previous best result by 22% on NLVR2.

VisualBERT Li et al. (2019) proposed an early fusion of image features, which uses pre-extracted object features from a vision model and directly feeds into the Transformer. Experiments on vision-and-language tasks, including NLVR2, show that VisualBERT outperforms its rivals while being significantly simpler.

More recently, CLIP Radford et al. (2021) shifted the paradigm by pretraining on large-scale image-text pairs with contrastive learning. It trains an image encoder and a text encoder to predict the correct pairings of training examples. While CLIP was not specifically optimized for structured reasoning tasks such as NLVR2, its robustness has motivated continued exploration in vision-and-language reasoning.

ViLT Kim et al. (2021) achieved further improvements by processing raw image patches alongside the tokenized text. ViLT achieves competent performance on vision-and-language tasks, including NLVR2, without using region features or deep convolutional visual embedders. Its architecture provides a clean and simple framework to explore vision-and-language reasoning tasks.

VLMO Bao et al. (2022) proposed a unified dual/fusion encoder using a Mixture-of-Modality-Experts (MOME) Transformer, allowing flexibility for classification or retrieval tasks. It also introduced stagewise pre-training leveraging unimodal data alongside paired data to achieve strong performance on tasks like NLVR2.

XFM Zhang et al. (2023) aimed for a general foundation model across modalities using language, vision, and fusion encoders trained with techniques like stopping language encoder gradients during vision-language training. Its training also guided vision encoder learning via Masked Image Modeling using vision-language signals, enhancing both uni-modal and multi-modal capabilities.

Finally, PaliGemma Beyer et al. (2024) represents a new generation of large multimodal models, combining scalable vision encoders with autoregressive language models. Although originally designed for broad vision-language tasks, its architecture opens new possibilities for structured reasoning tasks such as NLVR2.

2.4 Relevant techniques

In this work, we build upon a range of adversarial and noise-injection methods developed for vision, language, and vision-and-language (V&L) models. We briefly summarize the core ideas behind each relevant technique.

The idea of adversarial paradigms trace back to the minimax framework of Generative Adversarial Networks proposed by Goodfellow et al. (2014), where a generator and discriminator compete to produce realistic samples and robust classification boundaries. Simple stochastic noise-injection schemes such as injecting Gaussian or uniform noise into CNN training by Akbiyik (2023) demonstrates that even unstructured perturbations can regularize models by smoothing decision surfaces. In the other modality, Rajeswar et al. (2017)'s adversarial generation of natural language highlights the challenges of discrete text perturbations and motivates attacks in a continuous space instead.

Inspired by Goodfellow et al. (2015)'s Fast Gradient Sign Method(FGSM)—which generates adversarial examples via a single-step gradient-sign perturbation of the input, subsequent work has extended gradient-based attacks and noise-injection strategies to vision-and-language models. For instance, Gan et al. (2020) integrated FGSM-style perturbations into both visual region features and word embeddings during large-scale pre-training and fine-tuning, yielding substantial gains on VQA and image–text retrieval benchmarks. More recently, Jang et al. (2024) proposed Replace-then-Perturb, a targeted two-stage attack that first swaps salient visual objects and then applies adversarial noise to the modified scene, underscoring the importance of jointly reasoning over perturbed visual and textual cues.

Building on these insights, we apply FGSM directly at the embedding layer of ViLT to bolster its robustness and improve visual reasoning performance on the NLVR2 benchmark.

3 [1 POINTS] TASK SETUP AND DATA

NLVR2 contains 107,292 examples of captions and image pairs, including 29,680 unique sentences and 127,502 unique images. Its total storage size is ∼13GB. The data is organized in .json files, where captions are explicitly typed out, but images are provided in the form of urls.

The following is the breakdown of data splits:

	Unique sentences	Examples
Train	23,671	86,373
Development	2,018	6,982
Test-Public	1,995	6,967
Test-Unreleased	1,996	6,970
Total	29,680	107,292

The images can be requested using a Google form, or we can use a provided script to download the images. All provided scripts to download and clean data need a Linux-based OS to run (no Windows)

Our experiments are all contained in the following GitHub repository: `https://github.com/Tyrest/Model-Citizens`.

4 [1 POINTS] BASELINES

We implemented ten baselines for NLVR2. Unimodal baselines include Tiny-BERT (text-only), GPT-4o (text-only), and CNN (image-only) to capture modality-specific biases. Simple multimodal baselines include BERT + CNN, CNN + RNN, and VisualBERT to explore basic cross-modal interactions. Competitive baselines include finetuned versions of ViLT, LXMERT, and UNITER, as well as base GPT-4o for zero-shot evaluation, integrating visual and textual embeddings. These models assess individual modalities, simple multimodal reasoning, and state-of-the-art performance.

4.1 UNIMODAL BASELINES

1. Tiny-BERT(text-only): For our text-only baseline, we employed a TinyBERT, a compressed version of BERT as text feature extractor. Specifically, we used the pretrained huawei-noah/TinyBERT_General_4L_312D model and fine-tuned it on the NLVR2 dataset.
2. GPT 4o mini (text-only): For this baseline, we prompt GPT 4 to generate a truth value given a caption. It is unknown how GPT 4 is trained but it has shown very strong performance on a variety of language-centric tasks. NLVR2 might even be in the dataset so it is possible that the model could have memorized some truth values.
3. CNN(image-only): We employed MobileNetV2 as our feature extractor. MobileNetV2 is a lightweight convolutional neural network designed for efficient image classification and feature extraction. We used a pretrained MobileNetV2 (IMAGENET1K_V1) model as our feature extractor and fine-tuned it on the NLVR2 dataset.

4.2 SIMPLE MULTIMODAL BASELINES

1. BERT + CNN: We integrated both BERT and CNN described in the unimodal baselines as feature extractors for textual and visual data, respectively. The extracted features from both models are then combined using a Multi-Layer Perceptron to generate the final prediction.
2. CNN + RNN: The key insight behind this model is its ability to separately process visual and textual information before integrating them for reasoning. However, performance may be constrained by the need for explicit multimodal fusion mechanisms.
3. VisualBERT: VisualBert integrates textual and visual embeddings within a BERT-based architecture. The visual embeddings are passed to the multi-layer Transformer along with the original set of text embeddings, allowing the model to implicitly discover useful alignments between both sets of inputs, and build up a new joint representation.

4.3 COMPETITIVE BASELINES

1. ViLT(fine-tuned): The key insight behind this model is its ability to jointly process textual and visual information in a transformer-based architecture without relying on an external object detector, enabling direct image-text reasoning.
2. LXMERT: The key insight is that leveraging cross-modal attention between visual and textual modalities enables effective reasoning over complex, multimodal relationships.
3. UNITER: This model performs very well on a variety of V+L tasks including visual question answering. It can be further fine-tuned on the NLVR2 dataset to have a very competitive performance.
4. GPT-4o: The key insight is whether a state-of-the-art model that is not specifically fine-tuned for the NLVR2 task can perform well. While it offers strong language understanding, its performance may vary due to the lack of task-specific training and reliance solely on textual reasoning. In practice, performance is strong but less impressive when taking into account its size.

5 [3 POINTS] PROPOSED MODEL (>1 PAGE)

5.1 MODEL DIAGRAM

We propose an adversarial training method based on the ViLT model to fine-tune on the NLVR2 dataset, aiming to enhance the model's robustness. Specifically, we feed the two images into the model separately, including the caption for both. The resulting two output embeddings are then pooled and a simple linear layer is used to compute the confidence for true and false. We introduced perturbations to the weights of both the word embedding layer and the image embedding layers during training, adding the gradients of both backward passes (one with noise and one without) to determine the final gradient. The overall model architecture is illustrated in Figure 2.

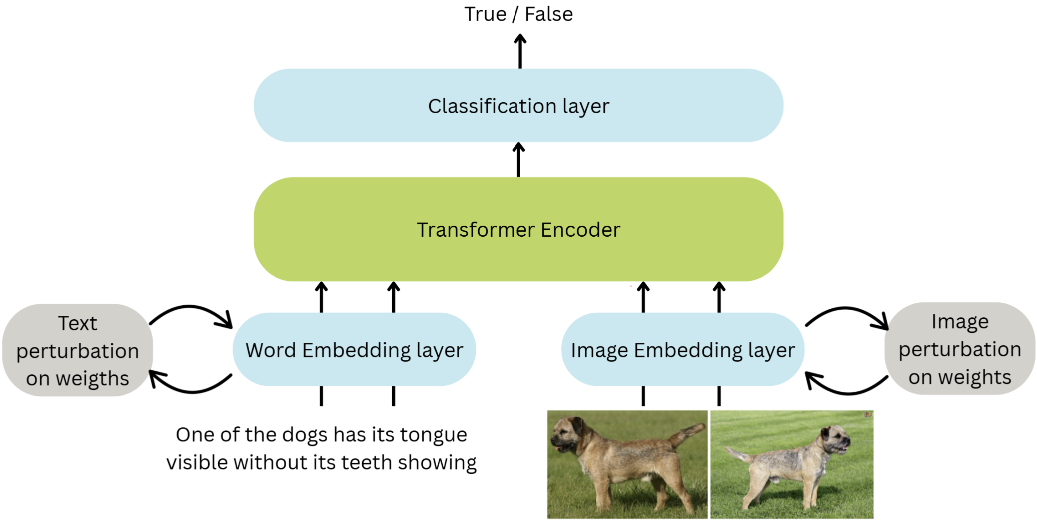

True / False
Classification layer
Transformer Encoder
Text perturbation on weigths
Word Embedding layer
Image Embedding layer
Image perturbation on weights
One of the dogs has its tongue visible without its teeth showing

Figure 2: Model architecture

5.2 LOSS FUNCTIONS

When fine-tuning our pretrained ViLT model for NLVR2, the loss function used is binary cross entropy loss:

$$\mathcal{L}_{\text{NLVR2}} = -\big[y\log(p_1) + (1-y)\log(p_2)\big]$$

where $[p_1, p_2]$ is the output of the model's final softmax layer.

5.3 CHANGES TO TRAINING DATA

We had originally experimented with augmenting the training data through image perturbation using FGSM, but ultimately found that it did not have a positive effect on performance. So, for our final experiments, we used the training data as is.

5.4 TRAINING SETUP

The data was already predivided into 4 sets: a training set, a development (validation) set, a public test set, and a private test set. The images were resized to a constant size of 384x384 pixels before being loaded into the model.

We modify the standard ViLT fine-tuning algorithm to also include perturbing the weights of specific layers before updating. Specifically, for any weight w that we perturb, we use the following procedure to update w at each step of the algorithm:

```
$w_0 \leftarrow w$
$g_0 \leftarrow \nabla\mathcal{L}(w = w_0)$
$w_1 \leftarrow w_0 + \varepsilon * \text{sign}(g_0)$
$g_1 \leftarrow \nabla\mathcal{L}(w = w_1)$
$w \leftarrow w - \alpha * (g_0 + g_1)$
```

We evaluate the model after each pass-through of the dataset (one epoch) on the development set.

5.5 Hyperparameters and their effects

The primary hyperparameter that our model tunes on is the ε parameter in FGSM. Specifically, we explore different combinations of ε values for image and text to see what combination of perturbations to the two modalities is optimal. This parameter enumerates how big of a perturbation we make in the direction of the gradient. A lower value of ε will lead to a more stable training process, although it might not make the model robust enough to more general noise. A higher value of ε means that we challenge the model to be robust to stronger noise, although this might have the unintended consequence of harming the model's ability to generate strong input embeddings.

6 [1 POINTS] RESULTS (1 PAGE)

Replace columns with the correct metrics for your task (extrinsic). Include multiple versions of your final model. You do not need to run on the test set but are encouraged to try if you have nice results on Dev.

Methods	Dev Accuracy ↑	Test-P Accuracy ↑	Test-U Accuracy ↑
Unimodal 1: BERT(text-only)	50.9	51.1	51.4
Unimodal 2: GPT-4o(text-only)	50.9	51.1	51.4
Unimodal 3: CNN(image-only)	50.9	51.1	51.4
Simple Multimodal 1: BERT+CNN	49.1	50.1	50.6
Simple Multimodal 2: CNN+RNN	53.4	52.4	53.2
Simple Multimodal 3: VisualBERT	67.4	67.0	67.3
Previous Approach 1: LXMERT	74.9	74.5	76.2
Previous Approach 2: UNITER	78.4	79.5	80.4
Previous Approach 3: ViLT	73.9	74.1	73.9
Proposed: AdViLT (both) (v1)	75.3	74.3	75.4
Proposed: AdViLT (text) (v2)	75.3	74.7	75.0
Proposed: AdViLT (image) (final)	75.2	74.8	75.6

7 [3 POINTS] ANALYSIS (2 PAGES)

7.1 INTRINSIC METRICS

For our intrinsic metrics, we want to evaluate whether our models can capture meaningful differences in the representations of inputs with true and false labels, and whether our attention-based models can properly identify important sections of the inputs. To that end, we choose the following three intrinsic metrics:

Cosine similarity For models with multi-modal inputs, We take the average of the embeddings of the pair of images in the input data, and calculate the cosine similarity between that and the embedding of the caption in the input data. Cosine similarity is useful because it emphasizes the directional alignment of the features in our inputs and assesses how well the model aligns semantic content across modalities.

Silhouette score For each input, we cluster the embedding in the penultimate layer based on its true label and then calculate the silhouette score for each data point. Silhouette score measures how similar an embedding is to its own cluster relative to other clusters. In this case, it is a useful metric because it measures how well our model can separate true labels from false labels, which provides more information than accuracy does.

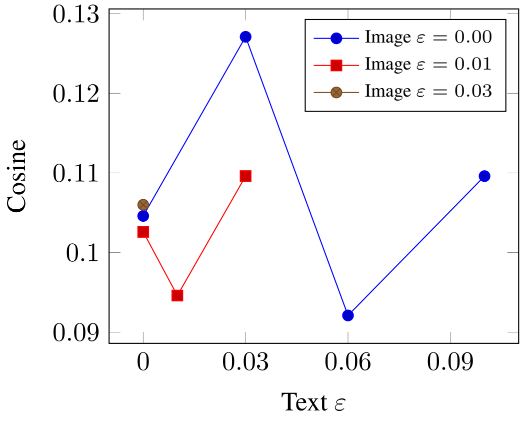
Attention entropy For models that have attention mechanisms, we calculated the attention entropy of the model. Attention entropy measures how widely an attention head distributes its attention across the tokens in a sequence. A lower entropy score means that the model is more precise in determining which input tokens are important in determining the output. In this case, it is a useful metric because it measures how well our model can determine the relevant areas of the text/images to focus attention on.

		Dev	
Methods	Cosine Similarity ↑	Silhouette score ↑	Attention entropy ↓
BERT Devlin et al. (2019)	-	0.0002	2.4800
CNN	0.8321	0.0001	-
LSTM	-	0.0002	-
BERT+CNN Devlin et al. (2019)	-0.0780	0.0001	-
RNN+CNN	0.0132	0.0001	-
CLIP Radford et al. (2021)	0.2725	0.0176	3.3320
PaliGemma Beyer et al. (2024)	0.4224	0.1036	1.7004
ViLT Pretrained Kim et al. (2021)	0.1190	0.0314	3.3451
ViLT Fine-tuned Kim et al. (2021)	0.1171	0.1101	3.2095
AdViLT	0.1271	0.0536	3.0650

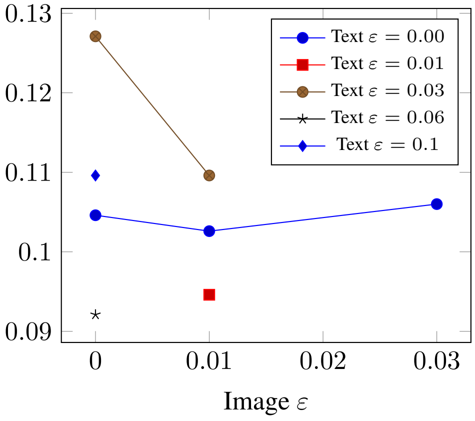
Table 1: A complete table of intrinsic metrics

The following charts show the values of our implicit metrics across various values of text ε and image ε.

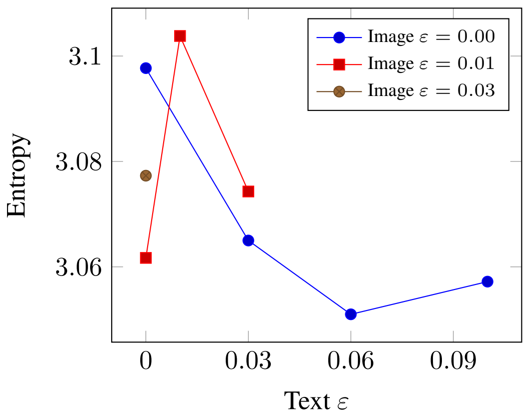
Cosine Similarity vs. Text ε



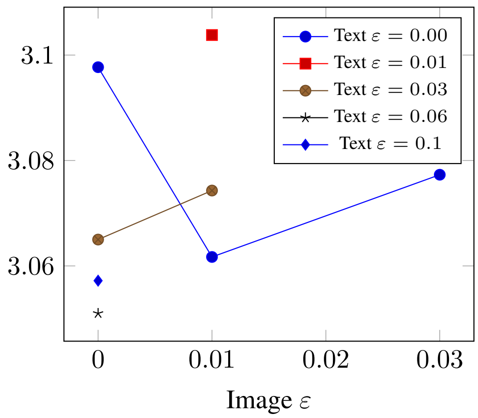
Cosine Similarity vs. Image ε



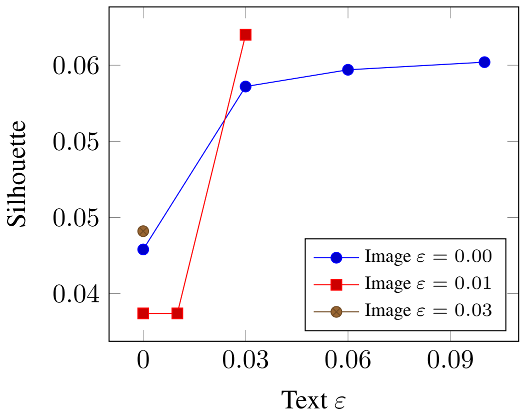
Attention Entropy vs. Text ε



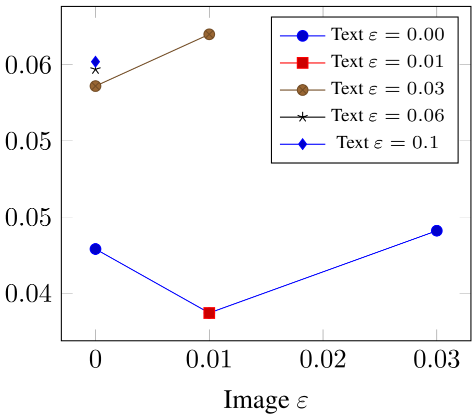
Attention Entropy vs. Image ε



Silhouette Score vs. Text ε



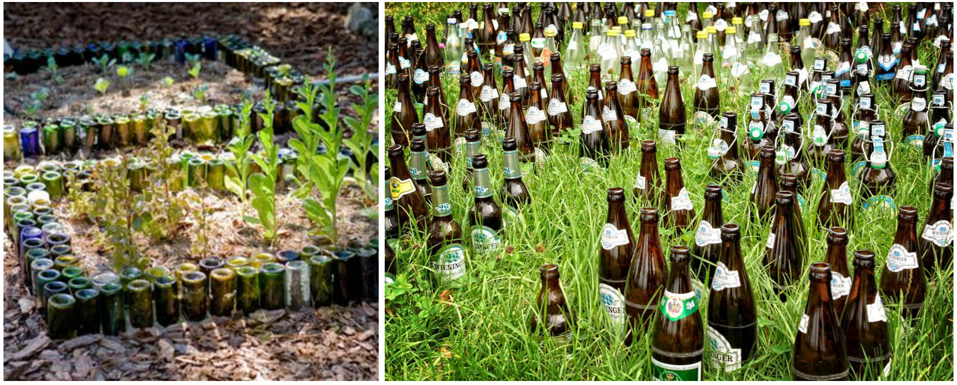
Silhouette Score vs. Image ε



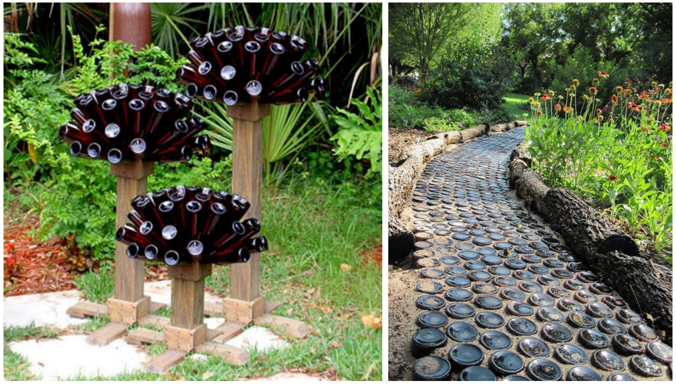
7.2 Qualitative Analysis and Examples (full page tables – multiple pages for most projects)

We tested the performance of our 10 models on the following 5 inputs:

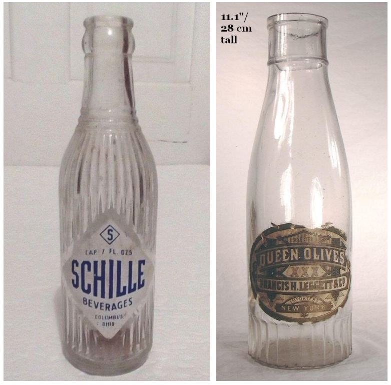
1. The right image shows a curving walkway of dark glass circles embedded in dirt and flanked by foliage.



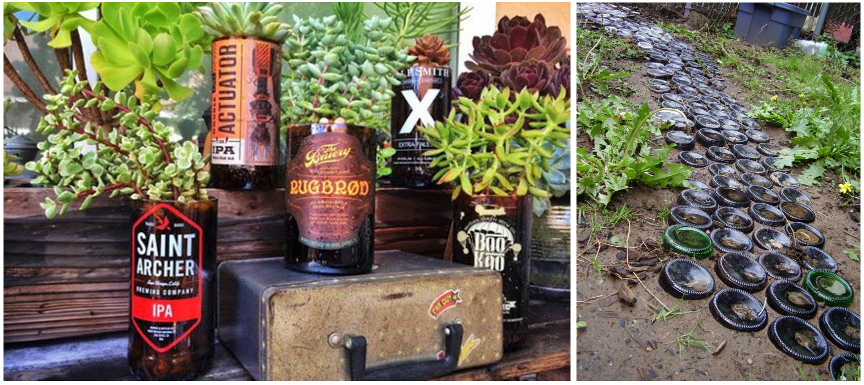
2. The right image shows a curving walkway of dark glass circles embedded in dirt and flanked by foliage.



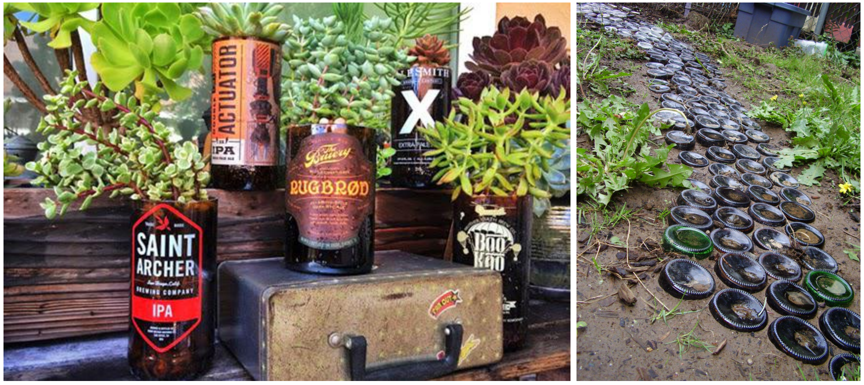
3. The right image shows a curving walkway of dark glass circles embedded in dirt and flanked by foliage.



4. The right image shows a curving walkway of dark glass circles embedded in dirt and flanked by foliage.



5. IN at least one image there are at least four bottle rows that together make a walking path.



Methods	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
BERT Devlin et al. (2019)	True	True	True	True	True
CNN	True	True	True	True	True
LSTM	True	True	True	True	True
BERT+CNN Devlin et al. (2019)	False	False	False	False	False
RNN+CNN	False	True	True	True	True
CLIP Radford et al. (2021)	False	False	False	False	False
PaliGemma Beyer et al. (2024)	False	True	False	True	True
ViLT Pretrained Kim et al. (2021)	-	-	-	-	-
ViLT Fine-tuned Kim et al. (2021)	False	True	False	True	True
AdViLT	False	True	False	True	True
Ground Truth	False	True	False	True	True

Table 2: Model outputs

Methods	Reason for failure
BERT Devlin et al. (2019)	The NLVR2 task is impossible without access to images.
CNN	The NLVR2 task is impossible without access to captions.
LSTM	The NLVR2 task is impossible without access to images.
BERT+CNN Devlin et al. (2019)	Concatenation of independent embeddings doesn't capture the cross-modal interactions required by NLVR2.
RNN+CNN	Concatenation of independent embeddings doesn't capture the cross-modal interactions required by NLVR2.
CLIP Radford et al. (2021)	CLIP primarily measures the similarity between the sentence and images based on learned associations and it doesn't perform explicit reasoning.
PaliGemma Beyer et al. (2024)	The model outputs matched ground truth for the five, but for other wrong predictions PaliGemma still likely lacks the vision-language reasoning capabilities to solve them.
ViLT Pretrained Kim et al. (2021)	The pre-trained VilT model is incapable of performing classification tasks and exhibits lower performance compared to the fine-tuned VilT model.
ViLT Fine-tuned Kim et al. (2021)	The model outputs matched ground truth for the five. However, ViLT's performance remains suboptimal due to its lack of a powerful visual feature extractor, high dependency on large-scale pretraining, and limited effectiveness of masked visual modeling.
AdViLT	Applying single-step FGSM perturbations directly to the embedding weights ignores other forms of perturbations that might be present in real-world data, and doesn't take into account other components of the model that are more closely aligned with reasoning. It could also weaken the model's ability to reason across both modalities leading to loss of performance on visual reasoning task.

Table 3: Failure Conditions

8 [2 POINTS] FUTURE WORK AND LIMITATIONS (1 PAGE)

Our initial experiments revealed several limitations.

Because FGSM is a one-step, gradient-sign attack, it can only attack in the nearest adversarial directions and may fail to explore other vulnerabilities in the model's representation. As a result, any robustness improvements we observe against FGSM might not generalize to real-world noise patterns present in real-life photographs and crowd-sourced captions, and instead be the result of overfitting.

Additionally, by focusing exclusively on perturbing input embeddings, we ignore the other components of the model that are more closely aligned with the visual reasoning component of the NLVR2 dataset. For instance, the primary weakness of existing models might be in their intermediate representations or self-attention masks. As such, any attacks that aren't directly related to input embeddings would be able to bypass our FGSM defense.

Finally, perturbing the embedding space could have unintended negative impacts on the model's ability to learn strong embeddings. For one, our FGSM implementation treats image and text embeddings identically even though both have different statistical properties and are likely sensitive to different kinds of perturbations. Furthermore, we risk perturbing embedding vectors beyond the space of valid image-text pairs.

If given more time, we would try several approaches to address these weaknesses and strengthen our model. For one, instead of using a one-step adversarial attack like FGSM, we could implement multi-step adversarial attacks such as PGD or AdvGAN to craft more challenging perturbations, potentially improving our model's robustness against a wider variety of attacks.

Additionally, instead of only perturbing the input embedding space, we could explore hybrid perturbation methodologies that combine embedding-space attacks with adversarial attacks on input images and text. This could also help our model be resilient to other kinds of noise in its inputs.

9 [1 POINTS] ETHICAL CONCERNS AND CONSIDERATIONS (UNINTENTIONAL, MALICIOUS, AND DUAL-USE)

In our approach to improving a model's performance on the NLVR2 dataset through controlled noise injection in images, we acknowledge several unintentional ethical risks that may arise:

Model Interpretability: Altering input data through augmentation introduces an additional layer of complexity in the model pipeline. This can pose challenges to the interpretability and transparency of model decisions.

Bias Amplification: Adding noise to images, if not done carefully, may affect certain visual features like skin tone, backgrounds, or clothing more than others. This could cause the model to learn incorrect patterns and perform poorly on underrepresented groups, potentially reinforcing existing biases in the data.

Misinterpretation of Visual Data: By introducing noise into visual inputs, there is a possibility of obscuring critical semantic or contextual details in the images. This may result in the model overfitting to noisy patterns while neglecting subtle but essential cues necessary for visual reasoning.

REFERENCES

M. Eren Akbiyik. Data augmentation in training cnns: Injecting noise to images, 2023. URL https://arxiv.org/abs/2307.06855.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, 2022. URL https://arxiv.org/abs/2111.02358.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL https://arxiv.org/abs/2407.07726.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020. URL https://arxiv.org/abs/1909.11740.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning, 2020. URL https://arxiv.org/abs/2006.06195.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

Jonggyu Jang, Hyeonsu Lyu, Jungyeon Koh, and Hyun Jong Yang. Replace-then-perturb: Targeted adversarial attacks with visual reasoning for vision-language models, 2024. URL https://arxiv.org/abs/2411.00898.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. URL https://arxiv.org/abs/2102.03334.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019. URL https://arxiv.org/abs/1908.03557.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL `https://proceedings.neurips.cc/paper_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf`.

Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015. URL `https://arxiv.org/abs/1511.08458`.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL `https://arxiv.org/abs/1505.04870`.

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, 2021. URL `https://arxiv.org/abs/2103.00020`.

Sai Rajeswar, Sandeep Subramanian, Francis Dutil, Christopher Pal, and Aaron Courville. Adversarial generation of natural language, 2017. URL `https://arxiv.org/abs/1705.10929`.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL `https://aclanthology.org/P18-1238/`.

Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, March 2020. ISSN 0167-2789. doi: 10.1016/j.physd.2019.132306. URL `http://dx.doi.org/10.1016/j.physd.2019.132306`.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2034. URL `https://aclanthology.org/P17-2034/`.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019. URL `https://arxiv.org/abs/1811.00491`.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019. URL `https://arxiv.org/abs/1908.07490`.

Xinsong Zhang, Yan Zeng, Jipeng Zhang, and Hang Li. Toward building general foundation models for language, vision, and vision-language understanding tasks, 2023. URL `https://arxiv.org/abs/2301.05065`.

A Appendix

Code Repository: `https://github.com/Tyrest/Model-Citizens`