Semantically Distributed Robust Optimization for Vision-and-Language Inference

Tejas Gokhale[*], **Abhishek Chaudhary**[*],
Pratyay Banerjee, **Chitta Baral**, and **Yezhou Yang**
Arizona State University
{tgokhale, achaud39, pbanerj6, chitta, yz.yang}@asu.edu

Abstract

Analysis of vision-and-language models has revealed their brittleness under linguistic phenomena such as paraphrasing, negation, textual entailment, and word substitutions with synonyms or antonyms. While data augmentation techniques have been designed to mitigate against these failure modes, methods that can integrate this knowledge into the training pipeline remain under-explored. In this paper, we present **SDRO**[†], a model-agnostic method that utilizes a set linguistic transformations in a distributed robust optimization setting, along with an ensembling technique to leverage these transformations during inference. Experiments on benchmark datasets with images (NLVR^2) and video (VIOLIN) demonstrate performance improvements as well as robustness to adversarial attacks. Experiments on binary VQA explore the generalizability of this method to other V&L tasks.

1 Introduction

"Does the text match the image?" – this simple question represents the Vision-and-Language Inference (VLI) task, as shown in Figure 1. Image-text matching forms the backbone for V&L pre-training (Sun et al., 2019; Tan and Bansal, 2019; Lu et al., 2019) and has resulted in improvements in downstream tasks such as visual question answering, image retrieval, referring expressions, and visual commonsense reasoning. While natural language inference (without visual inputs) has been extensively studied (Bowman et al., 2015; Williams et al., 2018; Khot et al., 2018; Demszky et al., 2018), VLI demands the additional capability of being grounded in the scene while understanding semantics. Although pre-trained language models (PLMs) (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020) have been useful for encoding text into vector embeddings, recent findings point to undesirably high cosine similarity of two random words (Ethayarajh, 2019), the struggle with negation (Kassner and Schütze, 2020; Ettinger, 2020), and semantically equivalent adversarial examples (Ribeiro et al., 2018). These findings call for robust training protocols to avoid propagation of these findings into VLI models.

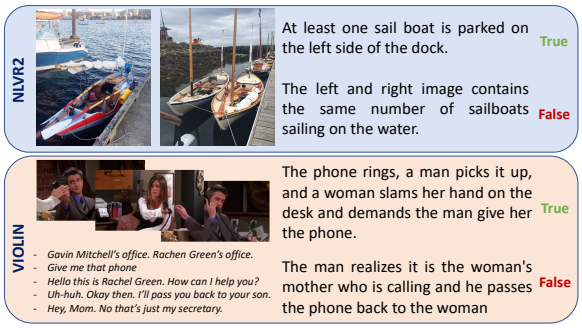

Figure 1: VLI models predict whether a sentence is `True` or `False`, given the visual input. *(Top)* sample from NLVR^2 with two images as input; *(bottom)* sample from VIOLIN with video and subtitles as input.

Adversarial training (AT) and distributed robust optimization (DRO) (Madry et al., 2018; Hu et al., 2018a; Sinha et al., 2018) have emerged as effective solutions to related problems in robust image classification, such as adversarial defense and domain generalization (Volpi et al., 2018). DRO assumes a perturbation set (typically an ℓ_p norm ball) around the training distribution, and minimizes the worst-case performance over this perturbation set. AT and DRO are popular for computer vision tasks, since the small perturbations of pixel intensities do not change the categorical meaning of the image.

However, in the case of text inputs, even small perturbations of their vector embeddings may result in absurd sentences or vectors that do not map to any word-token in vocabulary. The topology of the PLM embedding space is not well understood, especially with regard to what kind (and magnitude) of perturbations result in specific changes in

[*]Equal Contribution
[†]https://github.com/ASU-APG/VLI_SDRO

Findings of the Association for Computational Linguistics: ACL 2022, pages 1493 - 1513

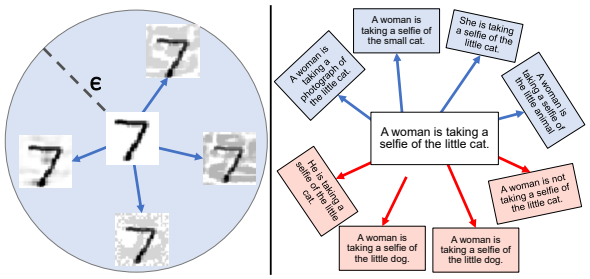

Figure 2: Comparison between *(left)* ϵ-bounded image perturbations and *(right)* linguistics-based *semantics-preserving* (blue) as well as *semantics-inverting* (red) transformations for sentences.

semantics, such as similar meanings (speak → talk) or opposite meanings (Heaven → Hell) without resulting in random or absurd words. Vector-based additive perturbations of text inputs thus restrict interpretability. However, in the domain of natural language, knowledge of logic, grammar, and semantics can be leveraged to transform sentences as shown in Table 1. Such linguistically-informed perturbations provide us control over the semantics of the resulting sentence and label, as shown in Figure 2.

We present a technique that modifies robust optimization by incorporating linguistically-informed transformations. Our approach: **S**emantically **D**istributed **R**obust **O**ptimization (SDRO) utilizes a pre-defined set of linguistic transformations (such as negation, word substitution, and paraphrasing) as the perturbation set instead of optimizing over the vector-space. We dub this set of transformations "SISP" i.e., semantics-inverting (SI) and semantics-preserving (SP) transformations. SDRO is *model-agnostic* since it can be applied to text inputs of any existing VLI model and *dataset agnostic* since it uses automated transformations without explicit knowledge of the text domain.

We apply SDRO to two VLI benchmark datasets: image-based NLVR^2 (Suhr et al., 2019) as well as video-based VIOLIN (Liu et al., 2020). To demonstrate the generalizability of SDRO to other V&L tasks, we also report results on the "yes/no" subset of VQA-v2 (Goyal et al., 2017). Our experiments show model-agnostic improvements in accuracy for all three benchmarks. While models trained with naive data augmentation using SISP suffer from a trade-off between robustness and accuracy, models that utilize SDRO improve along both metrics. SDRO also allows us to learn in low-resource settings, serving as a smart data augmentation tool – SDRO models trained only with 80% of the original dataset outperform existing state-of-the-art which utilizes the entire dataset.

Since SISP transforms do not require the ground truth label to either produce an SP or SI transformed sentence, we can also apply them to any new input sentences that are observed at test-time. Given a test input sentence, we generate its SISP versions, and obtain the prediction from our model for each SISP version. These predictions are ensembled using weighted averaging, giving equal weight to the prediction for the original sentence and the average predictions for all transformed sentences. We find that this ensembling of predictions of the SDRO model at test-time pushes the state-of-the-art further, thereby demonstrating the usefulness of semantic sentence transformations, both during training and testing.

2 Method

2.1 Preliminaries

Consider a training distribution P_{tr} consisting of inputs $\mathbf{x}$ and labels $\mathbf{y}$. For VLI, input $\mathbf{x}$ is multi-modal (visuals and text), with labels $\mathbf{y} \in \{\texttt{True}, \texttt{False}\}$. Under the empirical risk minimization (ERM), the following risk is minimized:

$$\mathcal{R}_{ERM} = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim P_{tr}} \ell(f(\mathbf{x};\theta),\mathbf{y}), \quad (1)$$

where ℓ is a suitable loss function such as cross-entropy loss for classification tasks. ERM provides generalization guarantees (Vapnik, 1991) for i.i.d. test samples, but not for out-of-distribution or adversarial examples.

Distributed Robust Optimization (DRO) (Hu et al., 2018b; Sagawa et al., 2020) searches for loss-maximizing perturbations of the input within an ϵ-divergence ball around P_{tr} and minimize the risk over such perturbed distributions.

$$\mathcal{R}_{DRO} = \mathop{sup}_{P:D(P,P_{tr})<\epsilon} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim P} \ell(f(\mathbf{x};\theta),\mathbf{y}). \quad (2)$$

The solution to Equation 2 guarantees robustness inside such ϵ-bounded distributions P. The inner maximization is typically solved using gradient-based methods (Madry et al., 2018) over additive perturbations δ such that $\mathbf{x}+\delta$ fools the classifier.

2.2 SDRO

For sentence inputs, additive perturbations are intangible and may result in ambiguity. An alterna-

	Category	Original	Transformed
SI	Noun-Antonym	The two women are driving on the street with the convertible top down.	The two ***men*** are driving on the street with the convertible top down.
	Verb-Antonym	There are children standing by the door.	There are children **sitting** by the door.
	Comparative-Antonym	There are more monitors in the image on the right than on the left.	There are ***few*** monitors in the image on the right than on the left.
	Number-Substitution	There are three bowls of dough with only one spatula.	There are ***eleven*** bowls of dough with only one spatula.
	Pronoun-Substitution	In one of the images, a woman is taking a selfie.	In one of the images, ***he*** is taking a selfie.
	Subject-Object Swap	The two women are driving on the street with the convertible top down.	The two ***top*** are driving on the street with the convertible ***women*** down.
	Negation	The closet doors on the right are mirrored.	The closet doors on the right are ***not*** mirrored
SP	Noun-Synonym	The right image shows three bottles of beer lined up.	The right ***picture*** shows three bottles of beer lined up.
	Verb-Synonym	Someone is using a kitchen utensil	Someone is ***utilizing*** a kitchen utensil.
	Comparative-Synonym	The bottle on the right is larger than the bottle on the left.	The bottle on the right is ***bigger*** than the bottle on the left.
	Number-Substitution	The two white swans are swimming in the canal gracefully.	The ***less than seven*** white swans are swimming in the canal gracefully.
	Pronoun-Substitution	In one of the images, a woman is taking a selfie.	In one of the images, ***she*** is taking a selfie.
	Paraphrasing	A man in a green shirt came on the porch and knocked on the door.	A man in a green shirt came ***up to*** the porch and knocked on the door.

Table 1: Examples illustrating the effect of each SISP transformation on input sentences.

tive approach is to consider *groups* $\mathcal{G}$ representing certain sub-populations or semantic categories within the data distribution. For text inputs in VLI, we consider the use of semantic sentence transformations as the perturbation mechanism – thus each transformation creates a sub-population or group of sentences. Examples of these transformations and their resulting effect on sentences is shown in Table 1. These transformations $g(x, y) = (\mathbf{x}_g, \mathbf{y}_g)$ are of two types: semantics-preserving (SP) if $\mathbf{y}_g = \mathbf{y}$, or semantics-inverting (SI) if $\mathbf{y}_g \neq \mathbf{y}$.

In this paper we propose the use of SI and SP transformations to create groups within the training data which can be leveraged by a robust optimization techniques to minimize worst group error. While previous work on adversarial training uses vector perturbations of sentence embeddings, our sentence-level transformations are interpretable as shown in Table 1). The ability of generating adversarial samples with inverted meanings is a key distinction between adversarial training (AT) and SDRO. While AT is restricted to SP perturbations inside an ϵ norm-ball, SDRO can impart larger linguistic perturbations (both SI and SP) beyond the norm-ball, by minimizing the worst-case expected risk over these groups:

$$\mathcal{R}_{SDRO} = \sup_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim g} \ell(f(\mathbf{x};\theta), \mathbf{y}). \quad (3)$$

Implementation. As a first step of SDRO, we randomly sample a subset $\mathcal{C}$ of the training dataset $\mathcal{D}$ s.t. $|\mathcal{C}|/|\mathcal{D}| = T$. We find adversarial samples after every epoch and create an augmented dataset $\mathcal{D}_{aug}$ which contains $(1-T)|\mathcal{D}|$ original samples and $T|\mathcal{D}|$ adversarial samples, thus retaining the size of the training dataset. We define ℓ_g as the classification loss for a transformed sample $(\mathbf{x}_g, \mathbf{y}_g)$:

$$\ell_g(\mathbf{x}, \mathbf{y}) \triangleq \ell(f(\mathbf{x}_g), \mathbf{y}_g),\ \forall g \in \mathcal{G}. \quad (4)$$

2.3 Variants of SDRO

We design two variants of SDRO: Sample-Wise (SW) and Group-Wise (GW) SDRO.

Sample-Wise SDRO is a greedy version of SDRO, in which, for every input $\mathbf{x}$, a transformation that maximally fools the classifier: $g^* = \text{argmax}_{g\in\mathcal{G}}\, \ell_g(\mathbf{x}, \mathbf{y})$, is added to the set of adversarial examples $\mathcal{D}_{adv}$. The model is then fine-tuned on the augmented dataset.

$$\mathcal{D}_{adv} = \{g^*(\mathbf{x}, \mathbf{y}) \colon (\mathbf{x}, \mathbf{y}) \in \mathcal{C}\}, \quad (5)$$

$$\mathcal{D}_{aug} = \mathcal{D}_{1:(1-T)|\mathcal{D}|} \cup \mathcal{D}_{adv} \quad (6)$$

However, this greedy approach is susceptible to the model's biases towards certain transformations. For instance, if negation and verb-antonym are universally hard for most sentences, i.e., result in the maximum classifier loss amongst all transformations g, then $\mathcal{D}_{adv}$ will be dominated by these groups, resulting in an unbalanced training set.

Group-Wise SDRO is devised to mitigate against the model becoming biased towards the "hardest" transformations. Using Equation 4, we calculate the transformation losses for each transformation of each sample in a training batch, yielding a set of classifier losses per "group" g:

$$L_g \colon \mathcal{C} \to \mathbb{R};\ L_g = \{\ell_g(\mathbf{x}, \mathbf{y}) \colon (\mathbf{x}, \mathbf{y}) \in \mathcal{C}\}. \quad (7)$$

We obtain the top-k losses per group g as:

$$L_G^k = \underset{\Lambda \subset L_G, |\Lambda|=k}{\text{argmax}} \sum_{\lambda \in \Lambda} \lambda, \text{ where } k = \left\lfloor \frac{|\mathcal{C}|}{|\mathcal{G}|} \right\rfloor. \quad (8)$$

Then $\mathcal{D}_{adv}$ is compiled as the union of per-group adversaries using Equation 8, and augmented to the training dataset using Equation 6.

Test-Time Ensembling of Predictions. Semantic transformations g allow us to obtain multiple "views" $\mathbf{x}_g = g(\mathbf{x})$ of the input, and the corresponding predictions $\hat{\mathbf{y}}_g = f(\mathbf{x}_g)$. We ensemble these predictions and the original prediction $\hat{\mathbf{y}} = f(\mathbf{x})$ with a simple weighted-average. Note that $\mathcal{G}$ contains both SP and SI transformations, $\mathcal{G}_{SP}$ and $\mathcal{G}_{SI}$. Since the expected label for $\mathcal{G}_{SI}$ is flipped, during ensembling we use the flipped probabilities $1-f(\mathbf{x}_g)$. The ensembled prediction is:

$$\hat{\mathbf{y}}_e = \alpha f(\mathbf{x}) + \frac{1-\alpha}{2} \sum_{g \in \mathcal{G}_{SP}} \frac{f(\mathbf{x}_g)}{|\mathcal{G}_{SP}|} + \frac{1-\alpha}{2} \sum_{g \in \mathcal{G}_{SI}} \frac{1-f(\mathbf{x}_g)}{|\mathcal{G}_{SI}|}. \quad (9)$$

Note that our method is a test-time ensemble and does not require training multiple models. This method is, in principle, similar to the ensembling strategy in image classification used by Chai et al. (2021) who train a generative model g to output different views of an image, and tune α over a validation set. In our work, g are semantic sentence transformations, and the value of α does not need to be tuned over a validation set – we find that the simple intuitive choice of α=0.5 (equal weight to the original sample and the SISP versions) improves performance. We find that:

- training models with SDRO using SISP transformations improves results on VLI tasks, and
- ensembling predictions of SDRO at test-time using Equation 9 further improves results.

3 SISP Sentence Transformations

This section describes the generation of semantics-preserving (SP) and semantics-inverting (SI) statements. **SISP** transforms are implemented using Spacy (Honnibal et al.). Dataset statistics and additional visualizations are in the Appendix.

Noun Synonym/Antonym: We extract nouns (subjects and objects) with dependency parsing, and find two nearest (synonyms) or farthest (antonyms) neighbors in the GloVe space (Pennington et al., 2014) using a threshold of 0.55.

Method	NLVR2			VIOLIN		
	Clean	SP	SI	Clean	SP	SI
Data-Aug	51.07	50.92	40.74	61.12	62.78	62.15
SW-SDRO	51.14	50.97	40.75	62.78	58.13	64.78
GW-SDRO	51.07	50.92	40.73	62.15	52.79	74.98

Table 2: Text-only evaluation of biases due to SISP transformations. 50% indicates no bias.

Verb Synonym/Antonym: We extract verbs using POS tagging and obtain their synonyms or antonyms. Verbs are lemmatized and inflected to the correct form using Lemminflect (Jascob, v0.2.1 (February 22, 2020).

Comparative Synonym/Antonym: Adjectival complements and modifiers are replaced with synonyms (*large* → *big*) or antonyms (*large* → *small*).

Number Substitution: Numerals are replaced by number-words (2 → *two*) or vice versa for SP transformations, or by their lower or upper bounds, (SP: *3* → *more than two*, SI: *two* → *less than two*).

Pronoun Substitution: Human-related nouns (such as *woman, boy, people*) are substituted by pronouns, while pronouns are substituted by generic descriptors (*something, someone, somebody, they*).

Negation: We use template-based negation (Gokhale et al., 2020b) with Subject-Verb Agreement (Wren and Martin, 2000). We add *'did not'* before a past-tense verb, *'do not'*, *'does not'*, or *'not'* before a base-form verb, gerund, or participle, or a *'not'* before an adposition or adjective.

Subject-Object Swap: Nominal or clausal subjects and direct or prepositional objects from the sentence are swapped for inverting semantics.

Paraphrasing: Input sentences are translated to Russian and then back-translated to English using neural machine translation (Ott et al., 2019).

3.1 Data Analysis

Quantification of Bias: Since SISP transforms are based on templates, they can potentially introduce spurious linguistic correlations in the dataset. For example, in NLVR^2 and VIOLIN datasets, negations and indefinite pronouns are infrequent. To quantify how this could impact models, we mask out the entire image and evaluate models (with VILLA as the backbone for NLVR^2 and HERO for VIOLIN). This acts as a 'text-only' evaluation,

with accuracies $\sim$50% implying lesser bias since models do not have access to visual information. Table 2 shows that SP transforms inflict lesser bias on models than SI transforms. The effect of bias is dataset-specific; SI makes the prediction of NLVR^2 samples harder than random (less than 50% accuracy) but easier for VIOLIN.

Transformation Fidelity: We employ human subjects to evaluate the quality of SISP-transformed sentences on (1) correctness of labels, (2) grammar, (3) semantics, and (4) visual grounding. We report a unified average 'transformation fidelity' (details are in Appendix). Fidelity is higher for SP samples than SI (90.50% v/s 79.51%), which resonates with the complexities of inversion of meaning (Russell, 1905) and leaves room for improvement in SI transformation.

4 Experiments

Datasets. For all datasets, given images/videos and natural language text as input, the system is expected to predict a binary class label. NLVR^2 (Suhr et al., 2019) contains $\sim 86K, 7K, 7K$ samples for training, development, and testing respectively. Each sample in NLVR^2 consists of a pair of images (from search engines) and a sentence (crowd-sourced). VIOLIN (Liu et al., 2020) contains video clips from popular TV shows and movies along with subtitles and crowd-sourced statements. VIOLIN contains $76K$, $9.5K$, $9.5K$ samples for training, validation and testing. VQA Yes/No consists of image-question-answer triplets from VQA-v2 dataset (Goyal et al., 2017). While VQA-v2 consists of multiple question and answer types, we focus on the subset of questions with binary *yes/no* answers ($\sim$38% of VQA-v2).

Evaluation Metrics. We use two evaluation metrics: (1) **Clean Accuracy**: accuracy on the i.i.d. benchmark test set, and (2) **SISP Accuracy**: average performance on SISP transformations of the test set. Since SISP transformations are automated and can be noisy (Sec 3.1), evaluation on the SISP test set can be considered a proxy for robustness.

4.1 Results

We compare SDRO with backbone models that use standard training data *(BASE)* and data-augmentation *(+data-aug)*. We train SDRO and backbones with the same hyperparameters. We apply test-time ensembling to the best SDRO model.

Model	Clean Acc.	SISP Acc.		
		SP	SI	Avg.
LXMERT$_{BASE}$	74.37	69.20	37.35	53.28
+ VILLA	75.98	69.94	39.09	56.15
+ data-aug	71.83	70.13	66.34	68.23
+ SW-SDRO	71.19	67.41	66.32	66.86
+ GW-SDRO	74.55	69.06	69.34	69.20
+ Test-Time Ensembling	74.75	–"–	–"–	–"–
UNITER$_{BASE}$	77.85	72.73	34.86	53.80
+ data-aug	76.65	70.34	81.04	75.69
+ SW-SDRO	78.43	69.71	67.50	68.61
+ GW-SDRO	77.55	67.93	81.66	74.79
+ Test-Time Ensembling	80.00	–"–	–"–	–"–
VILLA$_{BASE}$	78.39	73.15	34.15	53.65
+ data-aug	78.34	72.11	84.44	77.77
+ SW-SDRO	**79.23**	69.23	67.35	68.29
+ GW-SDRO	**79.41**	68.67	84.54	76.60
+ Test-Time Ensembling	**82.22**	–"–	–"–	–"–

Table 3: Results on the NLVR^2 public test set. ‡

Model	Clean Acc.	SISP Acc.		
		SP	SI	Avg.
VIOLIN$_{BASE}$	68.07	57.17	57.20	57.18
+ data-aug	61.58	67.64	67.70	67.67
+ SW-SDRO	62.81	62.84	62.68	62.76
+ GW-SDRO	63.71	64.58	63.16	63.87
+ Test-Time Ensembling	66.56	–"–	–"–	–"–
HERO$_{BASE}$	68.55	65.59	32.00	48.80
+ data-aug	65.21	59.20	81.81	70.51
+ SW-SDRO	**68.83**	58.97	77.83	68.41
+ GW-SDRO	68.19	56.20	82.92	69.57
+ Test-Time Ensembling	**69.90**	–"–	–"–	–"–

Table 4: Results on VIOLIN test set.‡

NLVR^2: We use Transformer-based models LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020b), and VILLA (Gan et al., 2020) as backbones for SDRO. VILLA (the current state-of-the-art for NLVR^2) uses standard adversarial training. The percentage of SISP-transformed samples is fixed at T=20%. Table 3 shows results on the NLVR^2 test set, with consistent model-agnostic improvements in clean accuracy over each baseline model and improved robustness on average. Both variants of SDRO improve over VILLA$_{BASE}$ by 0.84% and 1.02%, respectively. Test-time ensembling using Equation 9 leads to further gains, resulting in a new state-of-the-art accuracy of 82.22%, an improvement of 3.83% over VILLA$_{BASE}$. GW-SDRO results in the highest SI accuracy when used with each backbone model.

‡Notation: **bold**: $>$ SOTA; shaded: $>$ respective backbone model ($BASE$); underlined: best SI/SP accuracies.

Model	Clean Acc.	SISP Acc.		
		SP	SI	Avg.
UNITER$_{BASE}$	83.49	72.04	38.90	55.47
+ data-aug	82.53	77.03	93.70	85.36
+ SW-SDRO	83.92	75.82	88.92	81.48
+ GW-SDRO	84.05	76.95	93.41	85.18
+ Test-Time Ensembling	84.22	–"–	–"–	–"–
VILLA$_{BASE}$	84.82	74.15	37.40	55.77
+ data-aug	83.54	78.33	94.55	86.45
+ SW-SDRO	84.54	74.02	88.32	81.17
+ GW-SDRO	**85.12**	77.92	93.42	85.67
+ Test-Time Ensembling	**85.37**	–"–	–"–	–"–

Table 5: Results on the VQA yes/no subset.[‡] Not to be compared with VQA-v2 leaderboard since we use a smaller training set of *yes/no* questions.

VIOLIN: We consider VIOLIN$_{BASE}$ (Liu et al., 2020) and HERO (Li et al., 2020), the current state of the art, as baselines. VIOLIN$_{BASE}$ separately computes visual features using Faster-RCNN (Ren et al., 2015) and textual features using BERT (Devlin et al., 2019), and fuses them to be used as input to a classifier model. On the other-hand, HERO is a large-scale transformer-based pre-trained model which uses various V&L pre-training tasks to compute cross-modal features. We set T=40%. The results can be seen in Table 4. SW-SDRO model with the HERO backbone improves the state-of-the-art to 68.83%, and test-time ensembling further improves it to 69.90%. Interestingly, similar improvements in clean accuracy are not observed for VIOLIN$_{BASE}$, potentially because it does not use cross-modal pre-trained features.

VQA Yes/No: We use UNITER and VILLA as the backbone models, with T=20%. The motivation behind VQA experiments is to show that SISP transforms and SDRO can be extended to other V&L tasks. Table 5 shows that GW-SDRO is the best performing model in terms of clean accuracy, and is further improved by test-time ensembling.

5 Analysis

5.1 Visualization of Perturbations

In order to quantify the diverse and larger semantic transformations compared to additive perturbations, we study the tSNE (Van der Maaten and Hinton, 2008) embeddings of (i) original samples from NLVR^2 (P), (ii) their SISP-transformed versions (P_{SISP}), and (iii) their adversarially perturbed versions (P_{adv}). Input sentences are encoded using the UNITER text encoder for (i) and (ii), and the adversarial perturbation mechanism (Gan et al., 2020) for (iii). 3D tSNE embeddings are visualized in Figure 3; SISP transformed sentences (blue) are farther away than the perturbed versions. This shift is quantified by the KL-divergence (Kullback et al., 1951) between the distributions, with $D_{KL}(P_{SISP}||P) > D_{KL}(P_{adv}||P)$ implying that the diversity of SISP transformations is higher.

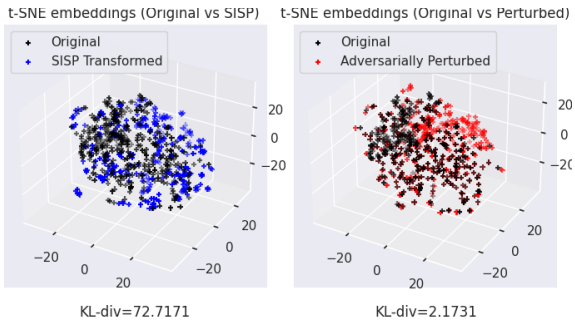

Figure 3: Comparison of original sentences (black) with *(left)* SISP-transformed sentences (blue) and *(right)* ϵ-bounded perturbations as a tSNE plot.

5.2 Comparison of Model Calibration

Figure 4 contains qualitative examples from NLVR^2 to compare output probabilities. We observe that SDRO models have higher clean accuracy, but lower confidence in the predictions than baseline and *data-aug* methods.

Reliability Diagrams. To validate this observation at scale, we use reliability diagrams to visualize model calibration (Niculescu-Mizil and Caruana, 2005), and plot model accuracy as a function of confidence (Guo et al., 2017). We use the softmax probability $\hat{p}$ of the predicted class as model confidence, split the range of probabilities into $M = 20$ equal-sized bins, and calculate bin accuracy $acc(B_m)$ and bin confidence $conf(B_m)$. If B_m is the set of all samples that fall in the m^{th} bin,

$$\text{acc}(B_m) \triangleq \frac{1}{|B_m|} \sum_{X_i \in B_m} \mathbb{1}(\hat{y}_i = y_i), \quad (10)$$

$$\text{conf}(B_m) \triangleq \frac{1}{|B_m|} \sum_{X_i \in B_m} \hat{p}_i. \quad (11)$$

A model with perfect calibration should have a reliability diagram such that $acc(B_m)$=$conf(B_m)$. We also report Expected Calibration Error (Naeini et al., 2015) over all n test samples:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|. \quad (12)$$

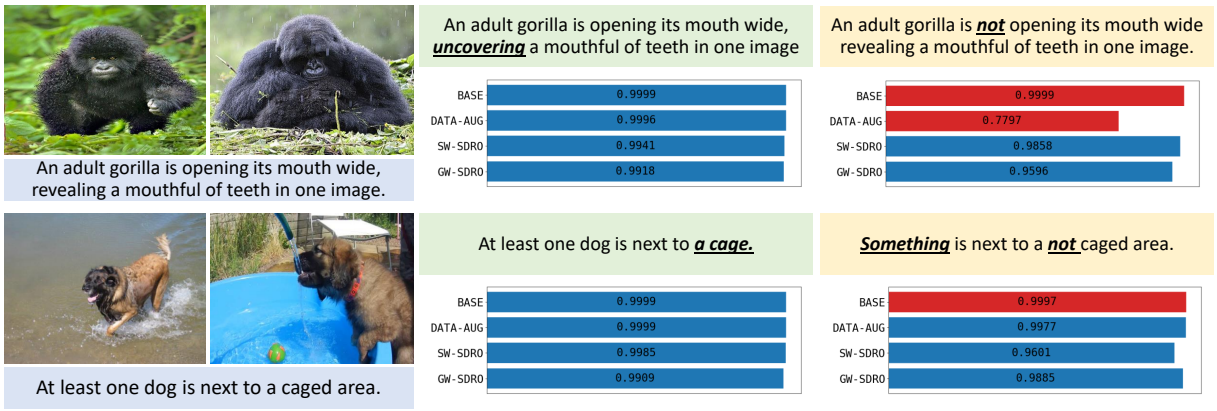

Figure 4: Qualitative examples showing test inputs from the NLVR^2 test set (left) with their respective SP (green) and SI (yellow) test samples. The predicted class (`True`/`False`) and the confidence of the predicted class is shown for baseline, data augmentation using SISP transforms, SW-SDRO and GW-SDRO. All models are built on the VILLA backbone. Wrong predictions are highlighted in red.

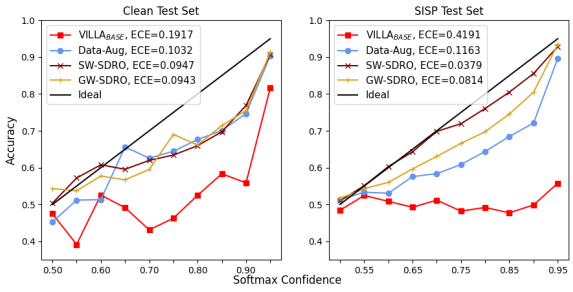

Figure 5: Comparison of reliability curves on the clean test set *(left)* and SISP test set *(right)*.

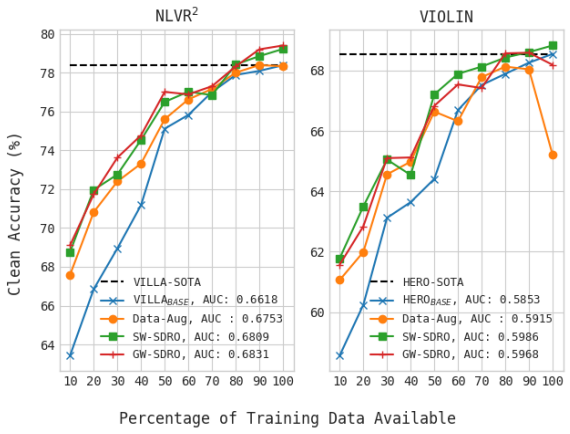

Figure 6: Effect of size of training data *(left)* NLVR^2, *(right)* VIOLIN. SDRO models are consistently better than baselines, even in low-data settings.

Reliability diagrams and corresponding ECE values for the VILLA trained with naive data augmentation and SDRO methods for NLVR^2 are shown in Figure 5. On both the clean test set and SISP test set, SDRO models have the lowest ECE. While the ECE for SDRO is marginally better than data augmentation for the clean test set, SDRO is better calibrated for the SISP test set, with SW-SDRO closest to ideal calibration among all evaluated models.

5.3 Size of Training Dataset

We evaluate models trained on small subsets of the original dataset, and compare their performance in Figure 6. SDRO models are significantly better at all sizes of training datasets as shown by accuracy and AUC (area under the curve). Notably, SDRO models trained with only 10% ($\sim 8.6K$) samples have performances similar to the baseline trained with 30% samples; SDRO models with 20% data are better than the baseline model with 40% data. While models trained with naive augmentation saturate below SOTA, at $\sim$80% data size, SDRO models cross the existing SOTA of 78.39%.

5.4 Proportion of Augmented Samples.

The final dataset has the same size as the original training set, but with T% transformed samples and $(100-T)\%$ original samples. The effect of this hyperparameter T is reported in Figure 7 as a percentage improvement of accuracy w.r.t. VILLA$_{BASE}$. An optimal value of T=20% leads to improvements in clean accuracy, but a larger proportion of augmented samples degrades performance. Similarly, higher T leads to higher robust accuracy, pointing to a trade-off between clean accuracy and robust accuracy at values of T higher than the optimal. This conforms with similar findings from Tsipras et al. (2019). While models trained with naive data-augmentation have better SISP accuracy than SDRO models as in Ta-

Model	SP only			SI Only			Both		
	Clean	SP	SI	Clean	SP	SI	Clean	SP	SI
Data-Aug	76.07	74.89	35.77	69.51	53.68	94.89	78.34	72.11	84.44
SW-SDRO	79.79	76.93	30.72	79.27	55.53	88.76	79.23	69.23	67.35
GW-SDRO	79.46	75.72	33.04	79.13	54.31	93.25	79.41	68.67	84.54

Table 6: Comparison of performance when only SP, only SI, or both types of transformations are performed.

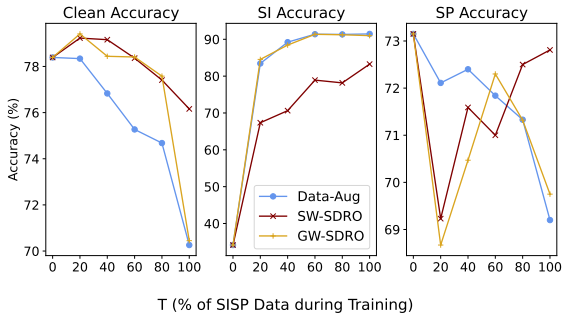

Figure 7: Plots showing the effect of the percentage of augmented samples on Clean, SP, and SI accuracies on NLVR^2, when using data-augmentation, and SDRO.

Model	SISP (Pos)			SISP (All)		
	Clean	SP	SI	Clean	SP	SI
Data-Aug	78.23	68.02	57.48	78.34	72.11	84.44
SW-SDRO	78.81	62.06	66.07	79.23	69.23	67.35
GW-SDRO	79.10	63.47	62.29	79.41	68.67	84.54

Table 7: Comparison of performance if only positive samples are used as inputs for SISP transformations

ble 3, they do so by sacrificing clean accuracy, while SDRO models improve along both dimensions compared to the baselines.

5.5 Ablation Studies

Contributions of SI and SP independently: We analyze which of the two categories (semantics-inverting (SI) or semantics-preserving (SP)) is the most effective by performing SDRO with only SI transforms, or with only SP transforms, and when using both. Table 6 shows that SDRO models trained only with SI suffer in terms of SP robustness and vice versa. However, there is still an increase in clean accuracy in both cases. This indicates that both SI and SP contribute towards improvements in robustness and clean accuracy.

Transformations of only `True` statements: Transforming *False* (negative) statements can lead to ambiguous and subjective meanings (Russell, 1905). We investigate if transforming only *True* (positive) statements is better than transforming both *True* and *False* statements. Table 7 shows that SISP transformations of both types of statements lead to higher clean accuracy and robustness.

	Model	CR	CS	CL	EDA	Emb	WN	Avg.
NLVR^2	VILLA	77.5	74.4	74.4	69.6	75.5	75.9	74.5
	+ SDRO	78.5	77.2	72.1	71.1	75.8	76.4	**75.2**
VIOLIN	HERO	66.1	63.0	68.6	60.9	63.8	63.4	64.3
	+ SDRO	68.7	65.0	69.0	61.3	65.5	64.6	**65.7**
VQA Yes/No	VILLA	80.5	75.7	84.9	74.6	78.6	76.4	78.5
	+ SDRO	86.0	84.5	84.1	87.0	84.3	84.0	**85.0**

Table 8: Performance evaluation on "text-attack" (Morris et al., 2020) versions of NLVR^2, VIOLIN, and VQA-Yes/No test sets.

5.6 Robustness to Text-Attacks

In this section, we test each model against text-based adversarial attacks – these attack samples are not seen by the models during training. Thus, this experiment seeks to verify if training with SDRO and SISP samples can also make VLI models robust against automated adversarial attack recipes. We utilize six common attack recipes implemented using the Text-Attack tool by Morris et al. (2020); these are – CLARE (CR) (Li et al., 2021a), character-swap (CS) (Pruthi et al., 2019), Checklist (CL) (Ribeiro et al., 2020), Easy Data Augmentation (Wei and Zou, 2019), counter-fitted embeddings (Emb.) (Alzantot et al., 2018), and WordNet-based swap (WN) (Ren et al., 2019). Table 8 shows results on each benchmark, using the best performing backbone for that benchmark and our SDRO model. On NLVR^2, VILLA+SDRO is better than VILLA for 4 out of 6 attack categories, and 0.7% on average. On VIOLIN, HERO+SDRO outperforms the baseline on all attack categories, leading to an average gain of 1.4%. On VQA-Yes/No, VILLA+SDRO outperforms the baseline on all attack categories, and 6.5% on average.

6 Related Work

Adversarial Training (AT) has been studied under a game-theoretic (Dalvi et al., 2004) and min-max setup (Madry et al., 2018). Volpi et al. (2018)

use AT to adversarially augment image classification datasets and show improved domain generalization for digit classification. Wong and Kolter (2020); Gokhale et al. (2021) modify AT for real-world adversaries beyond norm-bounded perturbations. AT has been used for text classification with LSTMs (Miyato et al., 2017) and for pre-training transformer-based models by adding label-preserving adversarial perturbations to embeddings of word tokens (Zhu et al., 2020; Jiang et al., 2020; Gan et al., 2020) Contrastive examples have been explored, collected from humans (Agrawal et al., 2018), negative mining (Shi et al., 2018), or synthetic generation (Agarwal et al., 2020; Chen et al., 2020a; Gokhale et al., 2020a; Teney et al., 2020).

Robustness in V&L has been explored for VQA, such as performance under prior probability shift (Agrawal et al., 2018) and domain adaptation (Chao et al., 2018; Xu et al., 2020), along with robustness for implied questions (Ribeiro et al., 2019) and novel compositions (Johnson et al., 2017; Agrawal et al., 2017), and robustness to logical connectives (including negation) Gokhale et al. (2020b). Teney et al. (2020) have shown that many V&L, image classification, and sentiment analysis models are sensitive to image editing. There has been a recent effort of model-in-the-loop dataset collection to guide humans to create harder VQA samples (Li et al., 2021b; Sheng et al., 2021).

Robustness in NLP: Generation of SP adversarial examples (Jia and Liang, 2017; Ribeiro et al., 2018; Iyyer et al., 2018; Alzantot et al., 2018), and approaches to defend against word substitution (Jia et al., 2019) have been explored in natural language processing tasks. Evaluation datasets have also been proposed for textual entailment that are manually crafted (Gardner et al., 2020) or template-based (McCoy et al., 2019; Glockner et al., 2018; Naik et al., 2018). Our method uses automated linguistically-informed SI and SP transforms for both training and inference.

7 Discussion

On Ensembling Coefficients. While designing our ensembling approach, we used $\alpha = 0.5$, i.e., equal contribution from the original output and the average of all outputs for transformed samples. This choice is generic and does not rely on dataset- or model-specific characteristics of SISP accuracy. While treating α as a hyperparameter and tuning it on validation datasets could lead to further gains, our intuitive choice of $\alpha = 0.5$ is effective by itself.

On SI Samples. Tables 3, 4, 5 show that existing models perform well on SP transforms, implying that equivalent semantics are captured in transformer-based models. However, these models fail on SI samples resulting in a close-to-random (50%) average SISP accuracy. While images perturbed with noise, blur, weather, or digital artifacts (Hendrycks and Dietterich, 2019) retain semantics (an image of a "cat" remains a cat after perturbation), minimal changes to text inputs, such as a single word changing from "sitting" to "standing" or "not sitting", inflict large changes in meaning. We hope that future work on design of V&L evaluation criterion along the SI axis, could benefit from our findings. While we generated SI and SP text for VLI tasks, the idea could be extended to design SISP transformations for images, by operating at object-level instead of pixel-level

On combination of AT and SDRO. We show that combining AT with SDRO can improve VLI performance and incorporate domain knowledge into the training process, such as semantic knowledge that often exists in natural language or linguistic rules. This is explicitly observed with VILLA, which is pre-trained and fine-tuned using standard adversarial training (Gan et al., 2020). When fine-tuned with SDRO, VILLA+SDRO further improves compared to UNITER+SDRO. The combination of standard adversarial training, (which accounts for local adversaries inside a ϵ norm-ball) and SDRO, (which accounts for linguistic adversaries and contrastive examples, typically outside the norm-ball as shown in Figure 2) could lead to improved generalization in many other V&L tasks.

On differentiability. Linguistic transformations are not differentiable and prohibit gradient-based solutions to the inner maximization in SDRO. However, most V&L tasks would benefit from the incorporation of semantic knowledge into the optimization framework. Through SDRO, we show that explicitly choosing the $argmax$ over a pre-defined set of transformations leads to model-agnostic improvements for binary classification tasks in V&L. More sophisticated methods may emerge in the future to address non-differentiability by leveraging proximal point or trust-region methods (Eckstein, 1993; Conn et al., 2000) or Interval Bound Propagation (Dvijotham et al., 2018), to incorporate semantic knowledge into adversarial training.

Acknowledgements

This work was funded in part by National Science Foundation grants 2132724, 1816039 and 1750082, DARPA SAIL-ON program (W911NF2020006), and DARPA CHESS program (FA875019C0003). The authors are grateful to the volunteers who worked on rating the fidelity of our proposed transformations. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9687–9695. IEEE.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. IEEE Computer Society.

Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. 2017. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset. *arXiv preprint arXiv:1704.08243*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. 2021. Ensembling with deep generative views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14997–15007.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Cross-dataset adaptation for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5716–5725. IEEE Computer Society.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020a. Counterfactual samples synthesizing for robust visual question answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10797–10806. IEEE.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *ECCV*.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. 2000. *Trust region methods*. SIAM.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. 2004. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

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.

Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A. Mann, and Pushmeet Kohli. 2018. A dual approach to scalable verification of deep networks. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 550–559. AUAI Press.

Jonathan Eckstein. 1993. Nonlinear proximal point algorithms using bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Tejas Gokhale, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J Thiagarajan, Chitta Baral, and Yezhou Yang. 2021. Attribute-guided adversarial training for robustness to natural perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7574–7582.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020a. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, Online. Association for Computational Linguistics.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020b. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*. Springer.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Dan Hendrycks and Thomas G. Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018a. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2034–2042. PMLR.

Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018b. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2034–2042. PMLR.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Brad Jascob. v0.2.1 (February 22, 2020). Lemminflect. a python module for english word lemmatization and inflection. `https://github.com/bjascob/LemmInflect`.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. 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 4129–4142, Hong Kong, China. Association for Computational Linguistics.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 2177–2190, Online. Association for Computational Linguistics.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5189–5197. AAAI Press.

S Kullback, RA Leibler, et al. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and William B 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.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 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 *International Conference on Computer Vision (ICCV)*.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10897–10907. IEEE.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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, Online. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2901–2907. AAAI Press.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

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.

Bertrand Russell. 1905. On denoting. *Mind*, 14(56):479–493.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. *arXiv preprint arXiv:2106.02280*.

Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning visually-grounded semantics from contrastive adversarial samples. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3715–3727, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. 2018. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. In *European conference on computer vision*. Springer.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations,*

ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

V Vapnik. 1991. Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, pages 831–838.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5339–5349.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Eric Wong and J Zico Kolter. 2020. Learning perturbation sets for robust machine learning. In *International Conference on Learning Representations*.

PC Wren and H Martin. 2000. English grammar and composition. *New Delhi: S Chand & Company Ltd*.

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2020. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376, Online. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

	Category	LXMERT	UNITER	VILLA
	Original	74.37	77.85	78.39
SI	Comparative Antonym	49.19	40.11	34.32
SI	Negation	35.19	36.92	35.39
SI	Noun Antonym	29.94	35.35	39.05
SI	Number Substitution	45.26	39.53	35.24
SI	Pronoun Substitution	47.76	34.79	29.78
SI	Subject-Object Swap	20.26	27.65	30.41
SI	Verb Antonym	27.86	29.72	34.89
SP	Comparative Synonym	61.35	65.58	66.86
SP	Paraphrasing	71.33	73.62	73.46
SP	Noun Synonym	71.24	75.32	75.78
SP	Number Substitution	70.68	74.33	74.37
SP	Pronoun Substitution	69.36	73.36	73.16
SP	Verb Synonym	71.26	74.16	75.24

Table 9: Evaluation of NLVR2 baselines on SISP test samples.

Appendix

In this supplementary material, we provide fine-grained results of our experiments, along with detailed analysis for VIOLIN and VQA-Yes/No similar to Section 5 in the main paper. We also provide visualizations of the SISP data creation process, statistics for SISP-transformed samples, and details of our human evaluation study.

A Fine-Grained Results

A.1 Baseline Performance on SISP

In Tables 9, 10, 11 we compare the performance of baseline models on all 13 categories of SISP transforms. All baseline models are below random performance on all three datasets for all SI categories, except for VIOLIN$_{BASE}$ (Liu et al., 2020). This is an interesting finding since VIOLIN$_{BASE}$ is the only model that is not a pretrained transformer-based model, but uses simple fusion of visual and textual modalities. In this paper, we've considered 3 benchmarks, and $3 + 2 + 3 = 8$ backbone models in total. Of these, only VIOLIN$_{BASE}$– a non-transformer model, retains above-random performance on SISP samples. Performance on SP categories is the best for VILLA (Gan et al., 2020) for NLVR^2 and VQA Yes/No, and HERO (Li et al., 2020) for VIOLIN.

A.2 SDRO Performance on SISP

In Tables 12, 13, 14 we compare performance for the state-of-the-art model VILLA, as well as mod-

	Category	VIOLIN$_{BASE}$	HERO$_{BASE}$
	Original	68.07	68.55
SI	Comparative Antonym	58.33	31.66
SI	Negation	57.75	34.73
SI	Noun Antonym	57.21	37.06
SI	Number Substitution	54.21	26.07
SI	Pronoun Substitution	57.66	24.64
SI	Subject-Object Swap	57.59	31.13
SI	Verb Antonym	57.68	38.77
SP	Comparative Synonym	57.92	67.87
SP	Paraphrasing	57.32	65.81
SP	Noun Synonym	57.67	67.15
SP	Number Substitution	54.87	58.88
SP	Pronoun Substitution	57.68	66.74
SP	Verb Synonym	57.53	67.09

Table 10: Evaluation of VIOLIN baselines on SISP test samples.

	Category	LXMERT	UNITER	VILLA
	Original	83.13	83.655	84.82
SI	Comparative Antonym	36.7	39.07	39.59
SI	Negation	29.59	31.93	29.59
SI	Noun Antonym	48.36	53.21	50.88
SI	Number Substitution	26.32	42.11	49.47
SI	Pronoun Substitution	21.28	24.05	24.36
SI	Subject-Object Swap	24.68	31.33	26.06
SI	Verb Antonym	35.88	50.63	41.86
SP	Comparative Synonym	67.72	71.28	74.11
SP	Paraphrasing	79.63	79.37	80.74
SP	Noun Synonym	74.09	73.37	74.61
SP	Number Substitution	72.32	57.89	62.11
SP	Pronoun Substitution	74.82	76.11	77.48
SP	Verb Synonym	73.76	74.22	75.82

Table 11: Evaluation of VQA Yes/No baselines on SISP test samples.

els trained with naive data augmentation and our SDRO methods.

B Analysis for VIOLIN

Proportion of Augmented Samples. We perform an analysis by varying T (proportion of augmented samples) and report performance in Figure 8 as a percentage improvement of accuracy w.r.t. HERO$_{BASE}$. It can be seen that there exists an optimal value of T (40%), which leads to improvements in clean accuracy, but higher values of T, i.e., a larger proportion of augmented samples degrades performance. Similarly, higher T leads to higher robust accuracy, but lower clean accuracy. Models trained with naive data-augmentation may be more robust on SP and SI test samples than SDRO models, but they do so by sacrificing clean accuracy, while SDRO models improve along both

	Category	BASE	Data-Aug	SW-SDRO	GW-SDRO
	Original	78.39	78.34	79.23	79.41
SI	Noun Antonym	39.05	85.79	63.13	76.64
SI	Negation	35.39	65.75	72.78	57.29
SI	Subject-Object Swap	30.41	87.13	60.19	89.06
SI	Verb Antonym	34.89	72.58	55.18	84.19
SI	Number Substitution	35.24	95.79	75.79	93.07
SI	Pronoun Substitution	29.78	98.44	81.31	98.35
SI	Comparative Antonym	34.32	78.62	63.11	93.17
SP	Pronoun Substitution	73.16	72.81	64.91	69.68
SP	Number Substitution	74.37	81.27	77.63	78.42
SP	Comparative Synonym	66.88	64.63	64.16	66.32
SP	Verb Synonym	75.24	69.88	65.78	59.83
SP	Paraphrasing	73.46	74.89	75.74	76.13
SP	Noun Synonym	75.78	69.15	67.67	61.64

Table 12: Evaluation of SDRO models (with VILLA bacbone) on NLVR^2 SISP test samples.

	Category	BASE	Data-Aug	SW-SDRO	GW-SDRO
	Original	68.55	65.21	68.83	68.19
SI	Noun Antonym	37.06	86.38	76.61	95.19
SI	Negation	34.73	53.18	58.31	61.14
SI	Subject-Object Swap	31.13	94.28	74.98	95.08
SI	Verb Antonym	38.77	81.96	77.05	94.95
SI	Number Substitution	26.07	76.32	80.03	71.04
SI	Pronoun Substitution	24.64	99.41	92.89	98.03
SI	Comparative Antonym	31.66	81.12	84.44	65.04
SP	Pronoun Substitution	66.74	62.29	60.76	69.67
SP	Number Substitution	58.88	56.62	57.14	51.22
SP	Comparative Synonym	67.87	58.31	57.64	49.67
SP	Verb Synonym	67.09	56.42	56.49	50.15
SP	Paraphrasing	65.81	63.59	65.22	66.03
SP	Noun Synonym	67.15	57.99	56.67	50.47

Table 13: Evaluation of SDRO models (with HERO backbone) on VIOLIN SISP Data

dimensions compared to the baselines.

Contributions of SI and SP independently Table 15 shows that SDRO models trained only with SI suffer in terms of SP robustness and vice versa. However, there is still an increase in clean accuracy in both cases, thus indicating the efficacy of both SP and SI transformations.

Transformations of only `True` statements Table 17 shows that training with transformations of both `True` and `False` helps both robustness and accuracy.

C Analysis for VQA-Yes/No

Proportion of Augmented Samples. We perform an analysis by varying T (proportion of augmented samples) and report performance in Figure 9 as a percentage improvement of accuracy w.r.t. VILLA$_{BASE}$. Higher T leads to higher robust accuracy, but lower clean accuracy.

Contributions of SI and SP independently Table 16 shows that SDRO models trained only with

	Category	BASE	Dataaug	SW-SDRO	GW-SDRO
	Original	84.82	83.54	84.88	85.19
SI	Noun Antonym	50.88	97.85	92.04	92.06
SI	Negation	29.59	80.81	82.36	81.39
SI	Subject-Object Swap	26.06	98.83	96.19	98.98
SI	Verb Antonym	41.86	97.71	88.17	98.6
SI	Number Substitution	49.47	94.74	78.95	92.63
SI	Pronoun Substitution	24.36	95.36	90.86	94.24
SI	Comparative Antonym	39.59	96.58	89.64	96.05
SP	Pronoun Substitution	77.48	77.41	75.88	76.98
SP	Number Substitution	62.11	77.89	56.84	81.05
SP	Comparative Synonym	74.11	80.72	78.63	80.25
SP	Verb Synonym	75.82	76.85	76.13	75.51
SP	Paraphrasing	80.74	80.57	81.31	81.49
SP	Noun Synonym	74.61	76.55	75.27	72.23

Table 14: Evaluation of SDRO models (with VILLA backbone) on VQA Yes/No SISP Data

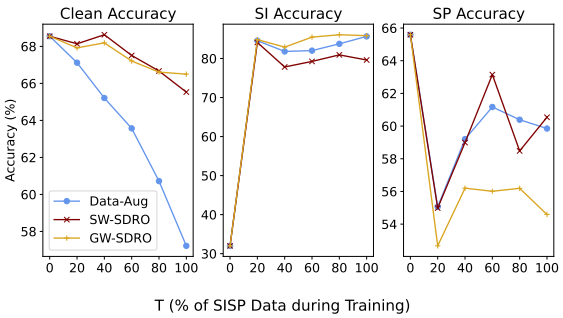

Figure 8: Plots showing the effect of the percentage of augmented samples on Clean, SP, and SI accuracies on VIOLIN, when using naive data-augmentation, SW-SDRO, and GW-SDRO.

SI suffer in terms of SP robustness and vice versa. However, there is still an increase in clean accuracy in both cases, thus indicating the efficacy of both SP and SI transformations. The increase in clean accuracy is greater for models trained with both SP and SI transformations.

Transformations of only `True` statements Table 18 shows that training with transformations of both `True` and `False` helps both robustness and accuracy.

D SISP Dataset

D.1 Statistics

In Tables 19, 20, 21, we show the number of SISP-transformed samples generated for the test sets of NLVR^2, VIOLIN and VQA Yes/No respectively. While we generate samples exhaustively for each category of transformation, during training these are sampled according to the proportion of augmented samples T, using three sampling strategies – naive data augmentation, SW-SDRO or GW-SDRO. On average, we obtain 5.69 SI samples and 5.65 SP

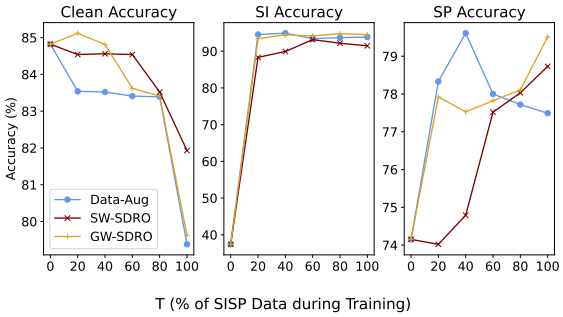

Figure 9: Plots showing the effect of the percentage of augmented samples on Clean, SP, and SI accuracies on VQA Yes/No, when using naive data-augmentation, SW-SDRO, and GW-SDRO.

samples per original sample for the NLVR^2 dataset, 11.14 SI samples and 10.83 SP samples for VIOLIN, and 2.75 SI samples and 3.5 SP sample for the VQA-Yes/No subset.

D.2 Data Generation

Figures 10 and 11 show flowcharts for our SISP transformation process for Semantics Preserving (SP) and Semantics Inverting (SI) respectively. For each image-sentence pair, the sentence is parsed using Spacy (Honnibal et al.) into tokens, dependencies, POS-tags, and noun chunks. Using this, each SISP function (for instance "Noun Synonym") generates insertions, deletions, substitutions, or paraphrasing as shown.

D.3 Transformation Fidelity

For each of the 13 SISP categories, we sampled 100 SISP-transformed examples from NLVR^2, thus giving us a total of 1300 samples. We employed 10 human subjects to evaluate the quality of SISP-transformed sentences. These human subjects were all proficient in English and at the time of the study were enrolled in graduate programs in an English-speaking country. The subjects were shown samples with the original images, sentences, and labels, as well as the new sentence and new label as shown in Figure 12. These subjects evaluated each sample with a binary (0/1) score, according to 4 metrics described below, along with an average "Transformation Fidelity":

1. Label Correctness (LC) – *Is the new label correct for the new sentence?*
2. Grammatical Correctness (GC) – *Does the sentence appear to be grammatically correct?*
3. Visual Grounding (VG) – *Does the sentence*

Model	SP only			SI Only			Both		
	Clean	SP	SI	Clean	SP	SI	Clean	SP	SI
Data-Aug	63.33	60.39	31.71	66.11	50.28	87.29	65.21	59.20	81.81
SW-SDRO	67.18	65.49	31.49	67.11	50.64	85.16	68.83	58.97	77.83
GW-SDRO	67.73	65.93	30.80	67.43	51.21	87.72	68.19	56.20	82.92

Table 15: Comparison of performance on the VIOLIN dataset when only SP, only SI, or both types of transformations are performed.

Model	SP only			SI Only			Both		
	Clean	SP	SI	Clean	SP	SI	Clean	SP	SI
Data-Aug	82	78.73	31.91	84.2	54.88	95.5	83.54	78.33	94.55
SW-SDRO	84.01	79.46	33.28	84.23	52.59	94.28	84.88	74.02	88.32
GW-SDRO	85.03	79.31	32.57	85.01	53.46	95.61	85.19	77.92	93.42

Table 16: Comparison of performance on the VQA Yes/No dataset when only SP, only SI, or both transformations are performed.

Model	SISP(Pos)			SISP(All)		
	Clean	SP	SI	Clean	SP	SI
Data-Aug	66.22	49.16	82.66	65.21	59.20	81.81
SW-SDRO	67.99	56.08	79.06	68.83	58.97	77.83
GW-SDRO	67.34	55.19	82.90	68.19	56.20	82.92

Table 17: Comparison of performance on VIOLIN dataset if only positive samples, i.e. samples with `True` labels are used as inputs for SISP transformations.

Model	SISP(Pos)			SISP(All)		
	Clean	SP	SI	Clean	SP	SI
Data-Aug	84.2	64.48	60.36	83.54	78.33	94.55
SW-SDRO	84.73	61.13	61.64	84.88	74.02	88.32
GW-SDRO	84.99	62.65	62.44	85.19	77.92	93.42

Table 18: Comparison of performance on VQA-Yes/No dataset if only positive samples, i.e. samples with `True` labels are used as inputs for SISP transformations, vs. transformations over both positive and negative samples.

	Category	Training	Test-P	Validation
	Original	86,373	6,967	6,982
SI	Comparative Antonym	14,177	1,244	1,172
	Negation	150,610	12,838	12,635
	Noun Antonym	148,959	12,719	12,635
	Number Substitution	83,080	7,468	7,113
	Pronoun Substitution	34,145	3,210	2,997
	Subject-Object Swap	30,533	2,944	2,787
	Verb Antonym	24,711	2,258	2,258
	Total SI	486215	42681	41714
SP	Comparative Synonym	13,302	1,066	1,163
	Paraphrasing	86,373	6,967	6,982
	Noun Synonym	212,904	18,570	18,968
	Number Substitution	60,582	5,194	4,994
	Pronoun Substitution	32,508	2,869	2,852
	Verb Synonym	78,103	7,314	6,919
	Total SP	483772	41980	41878

Table 19: Number of SISP-transformed samples generated per category for the NLVR2 dataset.

refer to at-least one visual entity from the image?

4. Semantic Correctness (SC) – *Is the sentence semantically sound and not absurd?*

The subjects were asked to view each sample and rate the new sentence and label on a binary scale for each of the four metrics. A snapshot of the interface used for the study as viewed by the human subjects is shown in Figure 12. Results are shown in Table 22 – split by the category of SISP transformation and in Table 23 – split by the ground-truth label of the original sample. Overall, our SISP transformed test set for NLVR^2 was rated at an average fidelity of 75.10%. It can be observed that on average, SP samples were rated to have higher average fidelity than SI samples, and False samples higher than True samples.

We also split the ratings (2 SISP categories and 2 labels: 2×2) and show results in Table 24. Overall, $SP(False)$ has the highest average fidelity, and $SI(False)$ has the lowest. LC (label correctness) for SI transformations of False statements is only 50%, probably because the inversion of a False statement using template-based methods may not always result in a True statement. On the other hand an SP transformation of a False statement remains False and got 100% LC. It is surprising

	Category	Training	Testing	Validation
	Original	76122	9600	9600
SI	Comparative Antonym	66,300	8,754	8,893
SI	Negation	249,836	31,634	31,923
SI	Noun Antonym	193,964	24,484	24,251
SI	Number Substitution	9,592	1,212	1,130
SI	Pronoun Substitution	156,466	19,785	20,100
SI	Subject-Object Swap	122,510	15,500	15,337
SI	Verb Antonym	49,802	6,358	6,356
	Total SI	848470	107727	107990
SP	Comparative Synonym	38,312	4,955	4,940
SP	Paraphrasing	76,122	9,600	9600
SP	Noun Synonym	418,285	52,857	52002
SP	Number Substitution	4,482	574	544
SP	Pronoun Substitution	91,125	11,464	11539
SP	Verb Synonym	196,826	25,044	25576
	Total SP	825152	104494	104201

Table 20: Number of SISP-transformed samples generated per category for the VIOLIN dataset.

	Category	Train	Trainval	Devval
	Original	92,761	38,374	5,323
SI	Comparative Antonym	18,839	8,044	1,139
SI	Negation	100,302	41,676	5,738
SI	Noun Antonym	82,885	34,543	4,835
SI	Number Substitution	1,505	730	95
SI	Pronoun Substitution	26,804	11,462	1,597
SI	Subject-Object Swap	11,793	4,999	683
SI	Verb Antonym	13,262	5,707	786
	Total SI	255390	107161	14873
SP	Comparative Synonym	21,259	9,037	1,271
SP	Paraphrasing	92,761	38,374	5,323
SP	Noun Synonym	119,301	49,977	6,850
SP	Number Substitution	1,443	678	95
SP	Pronoun Substitution	44,435	19,025	2,620
SP	Verb Synonym	45,612	19,384	2,622
	Total SP	324811	136475	18781

Table 21: Number of SISP-transformed samples generated per category for the VQA Yes-No dataset.

to observe that LC for SP transformations of True statements is low. $SP(True)$ received the highest GC and VG ratings, but low SC and LC ratings. VG ratings for all categories were consistently high.

Category	Fidelity Metrics				
	LC	GC	VG	SC	Avg.
SP	73.33	80.00	96.67	70.00	80.00
SI	71.15	67.31	96.15	57.69	73.08
All	71.95	71.95	96.34	62.20	75.10

Table 22: Human validation of our SISP transforms split according to the category of transformation.

GT Label	Fidelity Metrics				
	LC	GC	VG	SC	Avg.
`True`	70.69	72.41	98.28	56.90	74.59
`False`	75.00	70.83	91.67	75.00	78.13
All	71.95	71.95	96.34	62.20	75.10

Table 23: Human validation of our SISP transforms split according to the GT label of the original sample.

GT Label	Fidelity Metrics				
	LC	GC	VG	SC	Avg.
SP(`True`)	55.55	83.33	100.0	66.67	76.39
SI(`True`)	77.50	67.50	97.50	52.50	73.75
SP(`False`)	100.0	75.00	91.67	75.00	85.42
SI(`False`)	50.00	66.67	91.67	75.00	70.83
All	71.95	71.95	96.34	62.20	75.10

Table 24: Human validation of our SISP transforms split according to the GT label of the original sample.

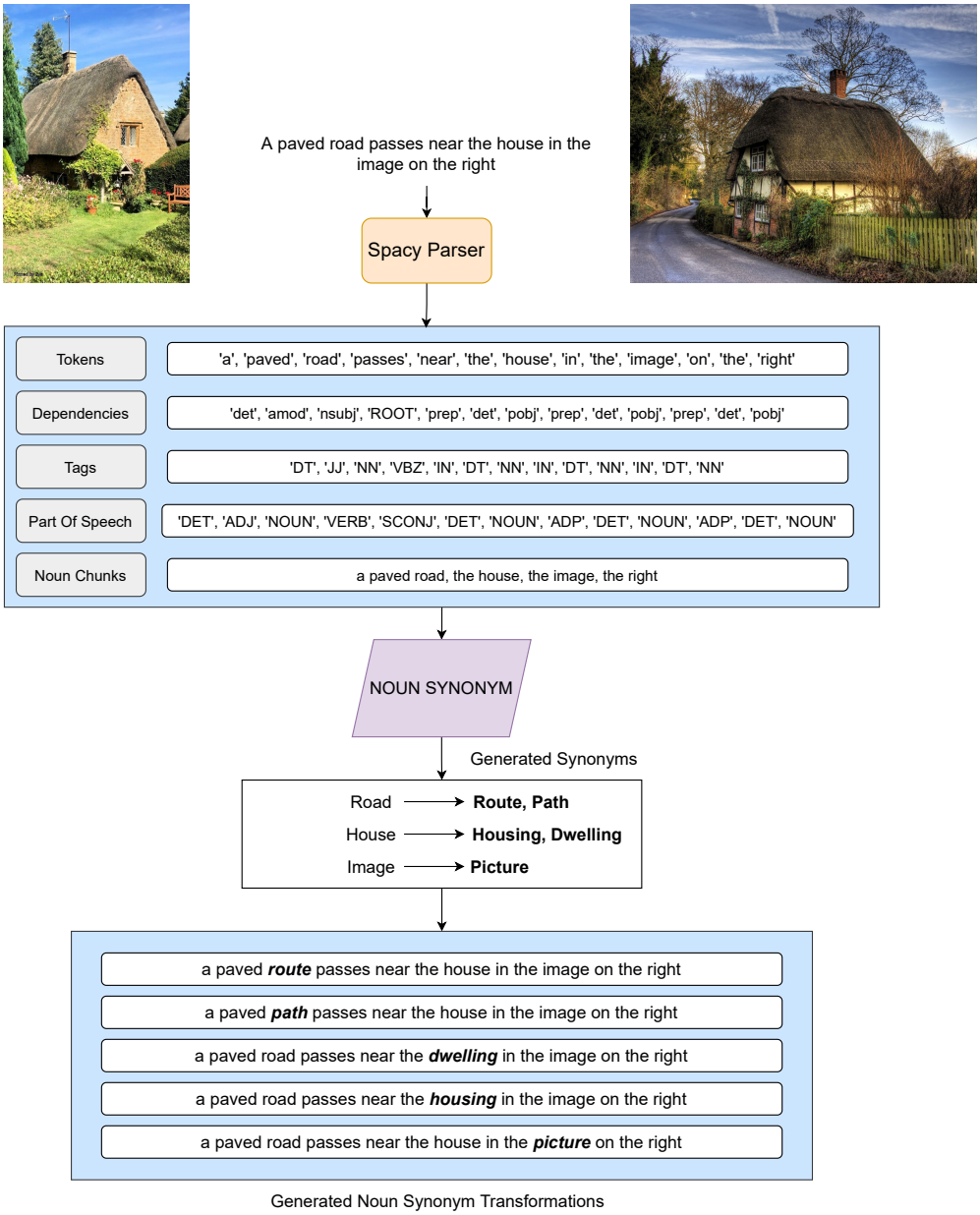

Figure 10: Illustration of the work-flow for generating SISP-transformed versions of input sentences. A Semantics-Preserving (SP) transformation is shown above.

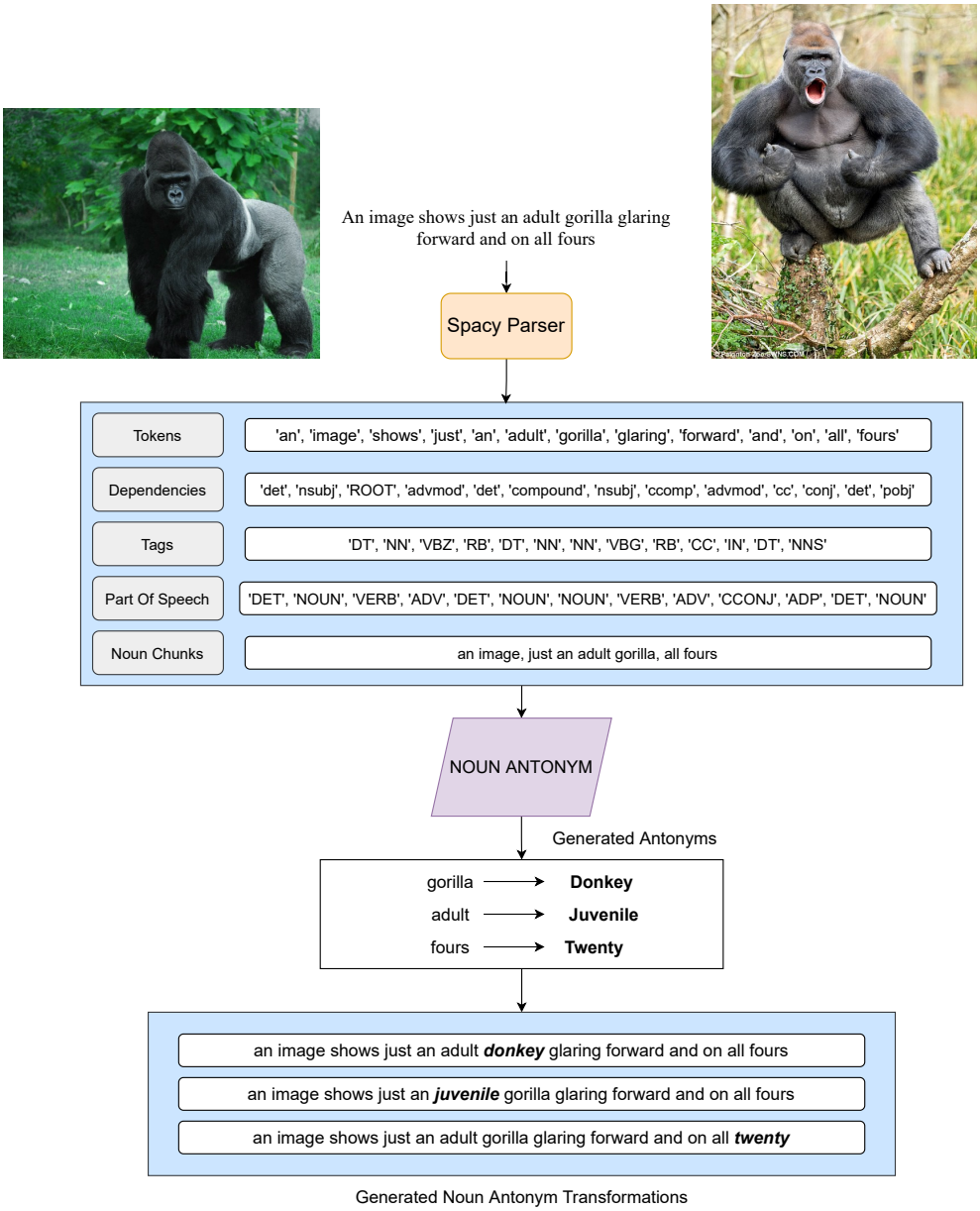

Figure 11: Illustration of the work-flow for generating SISP-transformed versions of input sentences. A Semantics-Inverting (SI) transformation is shown above.

3	sp_noun_synonym			In the right image, one horse is pulling a four-wheeled cart with two passengers to the right.	False	in the right image , one steed is pulling a four - wheeled cart with two passengers to the right .	False
4	si_negation			Three wine bottles with gold foil tops are stacked on a red mat.	True	three wine bottles with gold foil tops are not stacked on a red mat .	False
19	sp_noun_synonym			There are at least 4 gorillas sitting in the greenery.	True	there are at least 4 gorillas sitting in the green .	True
47	sp_noun_synonym			All of the pelicans are on shore and none of them are extending their wings.	True	all of the pelicans are on seaboard and none of them are extending their wings .	True
61	si_noun_antonym			One of the monitors is silver in color.	True	one of the monitors is purple in color .	False
62	sp_comparative_synonym			All paper rolls are upright, and one image shows a paper towel roll on an upright stand.	True	all paper rolls are vertical , and one image shows a paper towel roll on an upright stand .	True

Figure 12: Snapshot of a SISP example being evaluated by human subjects. Columns from left to right: sample-ID, SISP-tag, Left Image, Right Image, Original Sentence, Original Label, New Sentence, New Label.