

Generalization Differences between End-to-End and Neuro-Symbolic Vision-Language Reasoning Systems

Wang Zhu Jesse Thomason Robin Jia
University of Southern California, Los Angeles, CA, USA
{wangzhu, jessetho, robinjia}@usc.edu

Abstract

For vision-and-language (VL) reasoning tasks, both fully connectionist, end-to-end methods and hybrid, neuro-symbolic methods have achieved high in-distribution performance. In which out-of-distribution settings does each paradigm excel? We investigate this question on both single-image and multi-image visual question-answering through four types of generalization tests: a novel segment-combine test for multi-image queries, contrast set, compositional generalization, and cross-benchmark transfer. Vision-and-language end-to-end (VLE2E) trained systems exhibit sizeable performance drops across all these tests. Neuro-symbolic (NS) methods suffer even more on cross-benchmark transfer from GQA to VQA, but they show smaller accuracy drops on the other generalization tests and their performance quickly improves by few-shot training. Overall, our results demonstrate the complementary benefits of these two paradigms, and emphasize the importance of using a diverse suite of generalization tests to fully characterize model robustness to distribution shift.

1 Introduction

Widely used multi-modal pretrained models (Chen et al., 2020; Lu et al., 2019; Li et al., 2019) have exhibited great performance when fine-tuned on downstream vision-and-language tasks like VQA (Antol et al., 2015) and GQA (Hudson and Manning, 2019a). These models often generalize poorly to out-of-distribution (OOD) data, suggesting shortcomings in the VLE2E pipeline. Neuro-symbolic methods (Wu et al., 2017; Yi et al., 2018) try to address this issue by disentangling grounding and reasoning mechanisms in multi-modal systems. NS methods generate grounded visual representations, parse the language into executable programs for reasoning, and execute the programs on the visual representations. Previous work (Hudson and

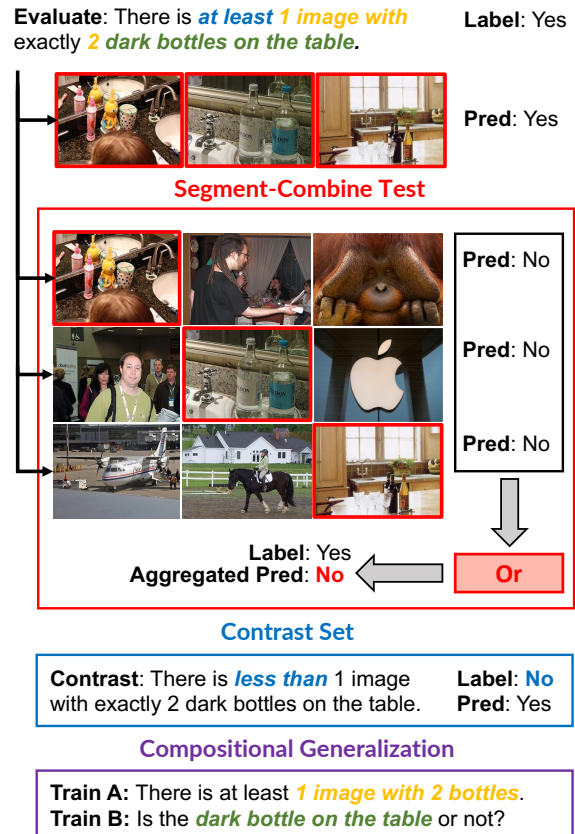


Figure 1: We build **segment-combine tests**, **contrast sets** and **compositional generalization** splits for multi-image question answering in the COVR dataset. The above question requires counting both within and across images. **Segment-Combine Test:** the multi-image query enables considering each image in isolation, pairing them with random unrelated images and feeding to the model, doing an OR operation of per-image answers. **Contrast Set:** language perturbation by replacing phrases in query with synonyms or antonyms. **Compositional Generalization:** the evaluated query is a compositional variant of questions Train A and Train B, involving reasoning on both **counting** and **relations**.

Manning, 2018; Mao et al., 2019) has shown the effectiveness of neuro-symbolic methods for OOD compositional generalization on single-image VL reasoning tasks. However, we still lack a compre-

hensive understanding of the generalization differences between these two paradigms under various setups. Given recent work suggesting that OOD accuracy often strongly correlates with in-distribution accuracy (Miller et al., 2020, 2021), we might expect VLE2E and NS systems to often have similar generalization abilities. But do they?

In this work, we conduct the first comprehensive comparison of generalization behavior between VLE2E and NS systems for VL reasoning tasks. Our study spans single-image and multi-image settings with natural images and includes four distinct types of generalization tests, three of which are shown in Figure 1. We introduce a novel **segment-combine test** for multi-image settings that requires models to make consistent predictions when some input images are replaced with irrelevant ones. We evaluate on **contrast sets** (Gardner et al., 2020), including new contrast sets we construct for COVR that test understanding of quantifiers. We also measure **compositional generalization** as defined by compositional splits from COVR (Bogin et al., 2021) and **cross-benchmark transfer** between VQA and GQA. We also develop improved NS systems for GQA by handling mismatches between program and scene graph object descriptors, and for COVR by refining the original logical language.

Overall, we find that VLE2E and NS systems exhibit distinct and complementary generalization patterns. The NS systems are more robust than the VLE2E systems in the first three testing situations. The VLE2E systems exhibit overstability to meaning-altering perturbations, suggesting they overfit to spurious correlations in the training data and do not learn precise reasoning skills. We further find that the semantic parsing module of NS systems can quickly improve on generalization tests given a few training examples, whereas VL models do not adapt as quickly. On the other hand, while VLE2E systems lose more than 10% in accuracy on transfer between VQA and GQA, the NS methods perform even worse. Taken together, our findings underscore the need for a diverse suite of generalization tests to fully compare different modeling paradigms. The different behavior of these two systems could guide the community to design more robust VL reasoning systems. We release our code for generating test data, and we encourage future VL models to be evaluated on these tests.¹

¹We release our code and test data at https://github.com/Bill11235813/gendiff_vlsys

2 Related Work

We first survey related work on vision-language reasoning models and OOD evaluation tests.

VL OOD Generalization. Many efforts have been made to evaluate the generalization ability of VLE2E systems and task-specific methods on compositionality (Johnson et al., 2017; Thrush et al., 2022a), language perturbations (Ribeiro et al., 2019) and visual perturbations (Jimenez et al., 2022). Li et al. (2020) showed VLE2E systems exhibit better robustness than task-specific methods. We are the first to comprehensively compare the generalization differences between VLE2E and NS systems across different OOD tests.

VL Pretrained Models. Large-scale, VL pretrained models for question-answering can be single-stream—encoding vision and language features together with a single transformer—such as VisualBERT (Li et al., 2019) and VinVL (Zhang et al., 2021), or dual-stream—encoding vision and language with separate transformers and applying cross-modal transformers later—such as ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019). We evaluate on both single- and dual-stream VL pretrained models.

Neuro-Symbolic Methods. NS-VQA (Wu et al., 2017) disentangled vision and language processing for VL reasoning tasks on simulated images. However, it requires the datasets to include annotations of logical forms to describe language. To reduce the supervision signal from program annotations, NS-CL (Mao et al., 2019) jointly learned concept embeddings and latent programs, and extended to natural images. NSM (Hudson and Manning, 2019b) learned graph-level reasoning and showcased the compositional reasoning abilities of NS methods. To be applicable to both single- and multi-image setups, we choose the same pipeline as in the original NS-VQA. We use the scene graph as the structural representation, and test on multiple language models for semantic parsing.

Single- and Multi-Image VL Reasoning Tasks. For VL reasoning, there are many datasets that focus on single images, such as CLEVR (Johnson et al., 2017), VQA, and GQA, as well as many other datasets that involve multi-image reasoning, such as NLVR (Suhr et al., 2017), NLVR2 (Suhr et al., 2019), COVR (Bogin et al., 2021), and

Winoground (Thrush et al., 2022b). We experiment with two single-image datasets, VQA and GQA, and one multi-image dataset, COVR, all of which use natural images.

3 Models

Next, we formally define the VL reasoning tasks and VLE2E and NS methods we study. We also discuss a new NS system for COVR and associated changes to the original COVR logical forms.

3.1 Vision-Language Reasoning

In a VL reasoning task, each example consists of a triple (q, \mathcal{I}, y) , where q is a natural language query, \mathcal{I} is a set of queried images and y is the corresponding answer of the query. The number of queried images is $|\mathcal{I}|$, e.g., $|\mathcal{I}| = 1$ for a single-image query. Given query q and image set \mathcal{I} , a VL system f predicts an answer $\hat{y} = f(q, \mathcal{I})$. Models are trained on $\mathcal{D}_{\text{train}}$ and evaluated on $\mathcal{D}_{\text{test}}$.

3.2 Modified VLE2E System

For a VLE2E system, f is a single neural network that is trained end-to-end. Since current VL pre-trained models are trained to process single images, we modify the VLE2E pipeline for multi-image settings following Bogin et al. (2021). Given a multi-image query (q, \mathcal{I}) and a pretrained model, for each image $I \in \mathcal{I}$, we feed the pair (q, I) to the pretrained model to get an image-text representation. We concatenate these $|\mathcal{I}|$ image-text representations and prepend a [CLS] token to construct a sequence of length $|\mathcal{I}| + 1$. We then input this generated sequence into a two-layer transformer, and take the produced embedding of the [CLS] token as the representation of the entire multi-image query. Finally, we feed the representation into an MLP classifier to predict y . All modules including the pretrained model are fine-tuned. We experiment with 4 different VL pretrained models: the single-stream VisualBERT and VinVL and the dual-stream LXMERT and ViLBERT.

3.3 Modified NS System

A NS system separately processes vision and language with two trainable modules ϕ and ψ . The image set is represented as $\phi(\mathcal{I})$, and the query semantics is represented as a functional program $\psi(q)$. A pre-defined executor executes $\psi(q)$ on $\phi(\mathcal{I})$ to predict the answer \hat{y} . To apply NS-VQA-like pipelines to real-world images, we use scene graphs as the structured representation $\phi(\mathcal{I})$.

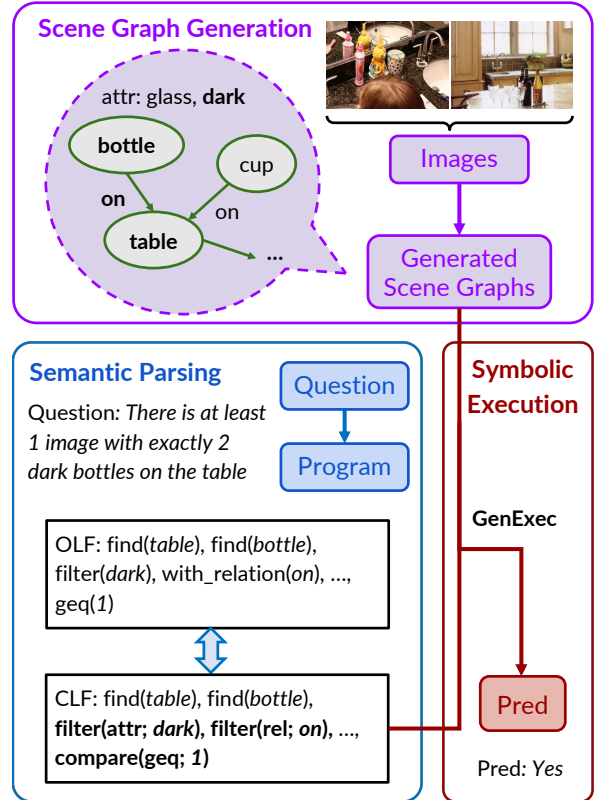


Figure 2: The process of the multi-image query with the modified neuro-symbolic methods. A language model (blue) maps the question to a functional program in our compositional logical forms (CLF) format; differences with the original logical forms (OLF) are shown in **bold**. A scene graph generator (purple) processes each image into a separate scene graph; queried information shown in **bold**. The program is executed on all the scene graphs together to produce an answer (red).

We use a pre-trained scene graph generator that can be fine-tuned on task-specific scene graph data, depending on the dataset (see §5.1 for details). We fine-tune large language models to map queries q to functional programs $\psi(q)$ (i.e., semantic parsing). We experiment with 3 language models: (1) T5 (Raffel et al., 2020), (2) BART (Lewis et al., 2020) and (3) GPT-2 (Radford et al., 2019).

Now, we describe dataset-specific work needed to build a full NS pipeline for GQA and COVR. Both datasets provide logical forms for each question, but these forms require modification to be compatible with NS systems.

Single-Image Queries. In GQA, functional programs align with objects in scene graphs via object IDs. For example, a program may refer to object “bird(775)”, while the corresponding node for object 775 in the scene graph could have the name parrot.

Since object IDs are not predictable by a semantic parsing model given q , we remove them from the annotated programs. Thus, we need to ground object references like “bird” to likely coreferents like the parrot node. We construct a dictionary that maps each object type mentioned in a program in $\mathcal{D}_{\text{train}}$ (e.g., “bird”) to the set of all scene graph object types that such a mention matched to (e.g., parrot). We use this dictionary to match objects between programs and scene graphs when executing programs at test time. Mismatches between object names in the program and scene graph occur in 9.5% of validation examples, but using this dictionary resolves 99.6% of the mismatches.

Multi-Image Queries. Like GQA, multi-image queries in COVR are annotated with executable programs and ground truth scene graphs of images. The program annotation incorporates quantifier operations, which enables NS execution of the multi-image queries without changing the pipeline of the NS-VQA methods (Yi et al., 2018). Figure 2 provides an overview of our multi-image NS pipeline.

In the compositional splits for COVR, models must generalize to some unseen compounds (e.g., phrases) consisting of seen tokens (e.g., words). For example, models could be tested on “Is the child sitting on a branch or a swing?” after seeing “What is the child sitting on?”, “Is the child sitting on a swing?” and “Is the child sitting on a branch?” at training time. However, the annotated logical forms in COVR for the above test query include an unseen unit operation `choose_name` (used to choose either “branch” or “swing”), which is not possible to generate as it was not seen at training time. To at least make compositional generalization possible, we design a set of compositional logical forms as an intermediate representation (Herzig et al., 2021) based on the existing programs in COVR. For the operation `choose_name(branch, swing)`, we take the prefix “choose” as the operation name and leave the postfix “name” as an argument, the new operation is `choose(name, branch, swing)`. By doing so, it becomes possible to generate this operation once we see a `choose(attr, ·)` and a `query(name, ·)` operation. We try to keep a minimum set of operations by redesigning non-composable operations and eliminating redundant operators. By doing so, we reduce the size of the operation set from 33 to 17. We cover the details of the modified programs in Appendix A. We denote the new logical forms as compositional logi-

cal forms (CLF) in contrast to the original logical forms (OLF), and evaluate the NS system based on these programs for the generalization tests.

Evaluation Metrics. We use 3 different evaluation metrics for the NS system. Our main evaluation metric is GENEXEC, the accuracy with the program execution on the generated scene graphs. To measure the effect of errors during scene graph generation errors, we also measure GTEXEC, the accuracy with the program execution on the ground truth scene graphs. Finally, we also measure EXACT, the exact match accuracy of the programs generated by semantic parsing; this penalizes “spuriously correct” parses that execute to the right answer but compute the wrong function.

4 Evaluation Methods

We evaluate VLE2E and NS systems on four generalization tests. We create a new multi-image perturbation test called the Segment-Combine Test, and create new contrast sets for COVR by perturbing quantifiers. We also test models on compositional generalization and cross-benchmark transfer.

Segment-Combine Test. We introduce the segment-combine test to test model generalization on multi-image perturbations. For a multi-image query (q, \mathcal{I}) where $\mathcal{I} = (I_1, \dots, I_{|\mathcal{I}|})$, we first perform a segmentation phase. We make $|\mathcal{I}|$ queries, where the k -th query uses the original question q and an image set formed by the union of the original image I_k plus $|\mathcal{I}| - 1$ random images *unrelated* to q . We feed these to the model to get $|\mathcal{I}|$ predictions. Next, in the combination phase, we apply an aggregation function (e.g., SUM or OR) based on the question type to fuse these predictions (Figure 1). A robust model should return the same answer on the segment-combine test and the original example.

We run the segment-combine test on COVR, sampling random images from all images in the COVR validation set. To confirm that we only sample images unrelated to the original image set (i.e., will not change the answer after fusion), we execute ground truth programs on ground truth scene graphs for each query in the segmentation phase, and find that the accuracy is 100%.

We focus on two templates in COVR for which there is an appropriate fusion function. For the template COUNTGROUPBY (e.g., “How many images have 2 bottles?”), the fusion function is SUM. That is to say, the answer on the original input

Template	Transfer Setting	VLE2E		NS		
		VisualBERT	ViLBERT	CLF w/ BART	CLF w/ GPT-2	CLF w/ T5
COUNTGROUPBY	Original Query	55.8	52.4	48.8 [99.0]	49.0 [99.7]	49.0 [99.9]
	Segment-Combine	44.6	40.1	48.7 [98.9]	48.9 [99.6]	49.0 [99.7]
VERIFYCOUNTGROUPBY	Original Query	72.1	73.6	70.7 [99.5]	70.8 [99.6]	70.9 [99.7]
	Segment-Combine	52.5	55.7	70.7 [99.3]	70.8 [99.6]	70.8 [99.7]

Table 1: Segment-Combine Test results on COVR. VLE2E models fail on both counting questions (COUNTGROUPBY) and binary questions (VERIFYCOUNTGROUPBY), while all NS models are robust. Oracle GTEXEC results for NS models are in brackets.

should be equal to the sum of the $|\mathcal{I}|$ answers from the segmentation phase. For the template VERIFYCOUNTGROUPBY, the fusion function is logical OR, as shown in Figure 1.

Contrast Sets. For VL reasoning, we define a contrast set (Gardner et al., 2020) of an example $(q, \mathcal{I}, y) \in \mathcal{D}_{\text{test}}$ to be a set of similar examples (q', \mathcal{I}, y') , where q' is similar to q and y' may or may not be the same as y , depending on q' . q' could be constructed by replacing specific words or phrases in q with synonyms or antonyms, or by substituting objects with other objects. Given n contrast set examples $(q'_1, \mathcal{I}, y'_1), \dots, (q'_n, \mathcal{I}, y'_n)$, we primarily evaluate models on the average accuracy on these n examples. We also measure the average *local coherency* as $\frac{1}{n} \sum_{i=1}^n (\hat{y}_i = \hat{y}'_i)$, which measures how much the model ignores perturbations.

We use the single-image contrast sets created by Bitton et al. (2021) for GQA. Their contrast sets involve object substitutions from scene graphs and mainly test the robustness of VLE2E systems for grounding objects.

For multi-image COVR, we design new contrast sets that target perturbations involving cross-image reasoning for multi-image queries. We replace quantifiers in the testing data with phrases of the equivalent and opposite meanings and change the labels accordingly. We focus on examples generated by 4 templates, where quantifiers (e.g., *at least*, *all*) play the role of introducing cross-image reasoning: one counting question template COUNTGROUPBY, and three binary question templates, VERIFYCOUNTGROUPBY, VERIFYCOUNT (e.g., “At least 2 bottles on the table?”) and QUANTIFIER (e.g., “No bottles are on the table?”). We test meaning-preserving perturbations such as replacing *at least* with *no less than* on counting and binary questions. We also test meaning-altering perturbations such as replacing *no* with *some* on binary questions and flipping the answer. We do not

apply meaning-altering perturbations to counting questions as it is non-trivial to determine what the new answer y' should be.

Compositional Generalization. In this setting, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ are from the same benchmark, but the queries in $\mathcal{D}_{\text{test}}$ are compositional variants of those in $\mathcal{D}_{\text{train}}$. For example, $\mathcal{D}_{\text{test}}$ examples may contain two phrases that were seen independently in $\mathcal{D}_{\text{train}}$ but never together. We test on the compositional generalization splits as defined in COVR, which are constructed by holding out a question template or holding out the examples where multiple query properties co-occur during training.

Cross-Benchmark Transfer. In this setting, $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ are from different benchmarks. We choose one of VQA and GQA as $\mathcal{D}_{\text{train}}$ and the other as $\mathcal{D}_{\text{test}}$.

5 Experiments

We present our experimental setup and results on four types of generalization tests below. The results indicate the complementary robustness of VLE2E and NS systems in OOD settings.

5.1 Experimental Setup

We use VQA (Antol et al., 2015) and GQA (Hudson and Manning, 2019a) as our single-image QA dataset and COVR (Bogin et al., 2021) as our multi-image QA dataset. VQA has three types of questions: binary, (yes/no), counting (answer is a number) and open-ended (answer can be any term from a vocabulary).

For the cross-benchmark transfer between VQA and GQA, as VQA and GQA have different sets of labels, we filter both validation sets to only include labels that appear in both datasets. Note that VQA has no program and scene graph annotations, so we can only train the NS methods on GQA.

Eval data	VLE2E			NS		
	LXMERT	ViLBERT	VinVL	w/ BART	w/ GPT-2	w/ T5
GQA-Val	83.9	83.5	89.1	65.6 [78.1]	71.8 [80.4]	74.3 [85.1]
GQA-Val-Contrast	66.5	68.2	73.3	64.8 [78.8]	71.9 [82.3]	74.1 [85.0]

Table 2: Contrast Set results on GQA. VLE2E shows $\sim 15\%$ of performance drop even if the contrast set is an easy object substitution, while NS models are highly robust. Note that GQA-Val is a subset of the validation set of GQA used to create the contrast set GQA-Val-Contrast. Oracle GTEXEC results for NS models are in brackets.

For model training, we use the fine-tuning setups described in the respective papers for each model. We give further details about hyperparameter selection in Appendix B. For NS methods, we generate scene graphs with the unbiased scene graph generation method Causal-TDE (Tang et al., 2020), which uses Faster R-CNN (Ren et al., 2015) as the backbone for object detection.

5.2 Segment-Combine Test

We test VLE2E and NS systems with segment-combine test and list their accuracy in Table 1.

VLE2E models fail on the segment-combine test. Both VisualBERT and ViLBERT fail on the segment-combine test, but NS models achieve accuracy close to the original query. The performance drop of VLE2E models is 11-12% on counting questions (VERIFYCOUNTGROUPBY) and 18-20% on binary questions (COUNTGROUPBY), as shown in Table 1. Though NS models with generated scene graphs show 1-7% lower accuracy than VLE2E models on the original multi-image queries, they achieve 4-18% higher accuracy on the segment-combine evaluation data.

VLE2E models learn multi-image spurious correlations. We notice VisualBERT’s performance on the segment-combine test for binary questions (52.5%) is close to random guessing. Thus, we extract the prediction from VisualBERT on the segment-combine test. For binary questions, 93% of the prediction are *no*. For counting questions with 6 labels, 38% of the predictions are *0*. As COVR queries are created by sampling related and distracting images, VLE2E models tend to predict *no* or *0* for queries with more irrelevant images, which is a spurious correlation between queried images learned during fine-tuning. By contrast, semantic parsing produces the right program to execute with EXACT score above 98.5% for all NS models, not just spuriously correct programs which are accidentally correct during execution.

5.3 Contrast Sets

We test on the augmented GQA contrast set from Bitton et al. (2021) for single-image queries, and compare the performance on the corresponding portion of GQA validation data. We also test VLE2E and NS systems on our generated contrast set on COVR involving cross-image reasoning.

VLE2E models show weak object grounding.

For perturbations that only involve object substitution, LXMERT, ViLBERT, and VinVL show a performance drop of 15-17%, as shown in Table 2. This drop implies the VLE2E training is not robust even on object grounding. Although the NS methods are worse than VLE2E systems on in-domain test data, they are highly robust on language-side object substitutions. Our NS pipeline with T5 outperforms the best VLE2E method by 0.8 points on the contrast set, despite being 14.8 points worse on the in-domain test data. This finding indicates the benefits on robustness of having a separate object grounding module.

VLE2E suffers on meaning-altering perturbations.

For perturbations involving cross-image reasoning, both VisualBERT and ViLBERT perform worse on meaning-altering perturbations than meaning-preserving perturbations, as shown in Table 3. For meaning-preserving perturbations, we observe no major accuracy drop on the counting questions, and a performance drop of 10-20% on the binary questions. On meaning-altering perturbations, replacing *at least* and *all* causes a more drastic performance drop of 40-65% for both VisualBERT and ViLBERT, while exchanging *no* and *some* only leads to 10% drop. Our hypothesis is VLE2E systems cannot generalize well to logical operations that are rare in the fine-tuning data: the opposites of *at least* and *all* rarely or never appear in the training data, whereas the opposites of *no* and *some* (i.e., *some* and *no*, respectively) are common. The local coherency is 96.3% for *at least* \rightarrow *less than* and 80.2% for *all* \rightarrow *either none*

Template	Transfer Setups	OOD FL	VLE2E		NS		
			VisualBERT	ViLBERT	CLF w/ BART	CLF w/ GPT-2	CLF w/ T5
COUNTGROUPBY	<i>at least</i> \rightarrow <i>at least</i>		53.3	53.0	48.9 [99.0]	49.0 [100.0]	49.0 [100.0]
	<i>at least</i> \rightarrow <i>no less than</i>	✓	54.0	52.9	23.7 [43.6]	21.7 [38.8]	21.5 [39.4]
VERIFYCOUNT	<i>at least</i> \rightarrow <i>at least</i>		87.6	84.3	75.1 [99.1]	75.6 [100.0]	75.6 [100.0]
	<i>at least</i> \rightarrow <i>no less than</i>	✓	68.3	67.0	49.9 [69.8]	48.2 [67.4]	48.4 [67.8]
	<i>at least</i> \rightarrow <i>less than</i>	✓ ✓	21.1	24.5	52.3 [64.0]	53.0 [65.5]	52.9 [65.6]
VERIFYCOUNT-GROUPBY	<i>at least</i> \rightarrow <i>at least</i>		74.8	70.5	71.4 [99.3]	71.6 [99.5]	71.4 [99.5]
	<i>at least</i> \rightarrow <i>no less than</i>	✓	65.8	63.2	56.2 [77.5]	55.0 [76.0]	55.6 [76.2]
	<i>at least</i> \rightarrow <i>less than</i>	✓ ✓	28.5	32.1	50.8 [51.8]	49.9 [50.4]	51.1 [50.6]
QUANTIFIER	<i>no</i> \rightarrow <i>no</i> , <i>some</i> \rightarrow <i>some</i>		86.5	90.2	75.2 [97.4]	75.5 [97.9]	75.5 [97.8]
	<i>no</i> \leftrightarrow <i>some</i>	✓ ✓	74.9	80.6	75.1 [97.3]	75.7 [97.9]	75.6 [97.8]
	<i>no</i> \rightarrow <i>no</i>		94.4	92.1	77.3 [96.7]	77.6 [97.0]	77.4 [96.9]
	<i>no</i> \rightarrow <i>at least one</i>	✓ ✓	83.9	78.6	53.7 [61.2]	54.5 [61.4]	53.9 [61.1]
	<i>some</i> \rightarrow <i>some</i>		80.3	88.7	75.4 [98.3]	75.6 [98.7]	75.6 [98.6]
	<i>some</i> \rightarrow <i>none of the</i>	✓ ✓	67.8	68.2	71.9 [92.1]	75.3 [97.3]	75.5 [97.8]
	<i>all</i> \rightarrow <i>all</i>		80.2	79.9	77.0 [98.2]	78.0 [98.9]	78.1 [99.1]
	<i>all</i> \rightarrow <i>either none or only some</i>	✓ ✓	36.2	39.2	54.0 [66.4]	55.1 [67.6]	57.5 [69.8]

Table 3: Contrast Set results on COVR. OOD: OOD test; FL: Flip labels. VLE2E has drastic performance drops on some of the meaning-altering perturbations, while NS shows equally performance drops regardless of meaning changes. Oracle GTEXEC results for the NS models are in brackets.

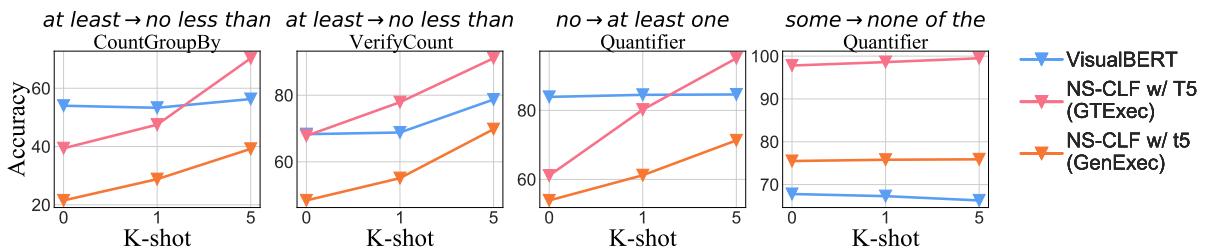


Figure 3: Few Shot Training on the COVR Contrast Set. The NS model with T5 as the semantic parsing module quickly improves performance with 5 new training examples. VisualBERT does not improve as much.

or only some, which implies the VLE2E systems do not pay enough attention to quantifiers whose opposites were not seen during fine-tuning.

NS performance has no correlation with meaning change. The NS methods, instead, show similar performance drop for both the meaning-preserving and the meaning-altering perturbations. The accuracy is higher than VLE2E models on most meaning-altering perturbations, but lower on the meaning-preserving ones, especially on counting questions. In some meaning-altering cases, the oracle accuracy is even close to 100%, which shows that the semantic parser is very robust in those situations.

NS recovers quickly with few-shot training. We add 1 to 5 examples from a contrast set to the full training dataset and re-train the model for few-shot learning. Figure 3 shows NS methods learn

quickly and adapt to the new example types, while VisualBERT learns slowly under few-shot training. With gold scene graphs, the NS accuracy increases even more quickly, suggesting that some improvements are hidden by the fact that our generated scene graphs are imperfect. Note that for the NS systems, we only adapt the language modeling part, as the contrast sets only affect language. Thus, we can also conclude language-only models adapt faster than VL neural models.

5.4 Compositional Generalization

We test VLE2E and NS systems with COVR compositional generalization sets and list their accuracy in Table 1. For the NS systems, we compare our compositional logical forms (CLF) to the original logical forms (OLF) from COVR.

CLF improves generalization. Comparing the last two columns in Table 4, it is clear that the

Split	Random Guess	In-distribution		Compositional Generalization			
		Text only	VLE2E	NS	VLE2E	NS	
		VisualBERT	CLF w/ T5	VisualBERT	CLF w/ T5	OLF w/ T5	
TPL-CHOOSEOBJECT	52.0	62.6	62.0	1.6	46.2 [50.9]	0.0	[0.0]
TPL-VERIFYQUANTATTR	50.4	76.9	70.5	71.2	44.8 [48.4]	0.0	[0.0]
TPL-VERIFYATTR	49.6	75.4	67.6	0.0	24.9 [34.4]	0.0	[0.0]
HAS-COUNT & HAS-ATTR	41.2	62.6	72.4	58.7	70.9 [99.1]	68.1	[93.4]
HAS-COUNT & RM/V/C	40.5	82.2	76.1	74.1	76.1 [100.0]	75.0	[98.4]
HAS-SAMEATTR-COLOR	49.8	71.2	67.7	66.0	67.7 [100.0]	67.7	[100.0]

Table 4: Compositional Generalization on COVR. The upper panel shows splits with held-out templates. The lower panel shows splits with held-out property combinations. In-domain random guessing accuracy is from a VisualBERT text only model. CLF improves the accuracy of NS on compositional generalization, and outperforms VisualBERT on most of the tests. Oracle GTEXEC results for the NS models are in brackets.

new CLF logical forms improve generalization to new combinations of query properties compared to original logical forms, and make generalization to new templates possible.

NS has lower in-domain but higher compositional generalization performance. In Table 4, the in-domain accuracies of the NS system are always lower than those of the VLE2E systems. However, on most of the compositional splits, the performance of VisualBERT is worse than the NS method. The only exception is VERIFYQUANTATTR, where there is a complex operation comparing whether two lists of objects have some same attributes. Following our hypothesis of VLE2E is better at questions with phrases occurring frequently in training, we compute the cosine similarity of the text embedding in ViLBERT, and find examples in the template VERIFYQUANTATTR are semantically close to examples in the template SPECIFICSAMEATTR. Examples for templates VERIFYQUANTATTR and SPECIFICSAMEATTR are “*Do all cats that are on a floor have the same color?*” and “*Does the dog that is in grass and the dog that is in water have the same color?*”, respectively. However, these two templates have different logical forms in both CLF and OLF, making it easier for VLE2E systems to generalize but harder for NS systems.

5.5 Cross-Benchmark Transfer

The cross-benchmark test aims to explore transferability between benchmarks of the same visual question-answering task. We evaluate the transfer between GQA and VQA because they share similar types of queries.

VLE2E is more transferable than NS. In Table 5, LXMERT has an 8-15% accuracy drop for

transfer from each dataset to the other. However, the NS method with T5 as the semantic parser has even worse performance on transfer. Using the scene graph generator and semantic parser trained on GQA, the accuracy of the NS method drops by over 70% on open questions.

Failure of NS is mainly due to scene graph generation error To understand the reasons for the failures of the NS system, we conduct manual analysis on 40 VQA examples. We observe that more than 75% of the VQA programs are correctly generated with the semantic parser trained on GQA. However, they often do not execute to the correct answer because (1) semantically similar objects have different node names in the generated scene graphs; (2) some objects are harder to detect due to visual domain shift. For example, for a generated program like [“operation”: “select”, “argument”: “mattress”], we may not find an object named “mattress” in the generated scene graph, where it could be named “bed”. To quantify this issue, we compute the missing object ratio, the percentage of programs that throw errors during execution because objects mentioned in the program are not found in the scene graph. The high missing object ratio in Table 5 suggests that the scene graph generation module trained on GQA cannot correctly match objects mentioned in the programs for VQA images.

NS occasionally requires new primitives. Another possible reason for the NS system’s cross-benchmark failure from GQA to VQA would be that some question types in VQA require new primitive operations. In our manual analysis, less than 10% of the VQA programs require the addition of new primitive operations, demonstrating that this

Transfer Setups		VLE2E	NS
		LXMERT	w/ T5
VQA→VQA	Binary	97.4	-
	Open	90.0	-
GQA→VQA	Binary	86.4	52.0 (78.4)
	Open	78.7	8.3 (80.2)
GQA→GQA	Binary	90.6	78.6 (3.8)
	Open	81.8	60.1 (5.3)
VQA→GQA	Binary	82.5	-
	Open	66.2	-

Table 5: The accuracy of LXMERT versus our NS method trained on X and deployed on the *validation* set of Y ($X \rightarrow Y$) for VQA and GQA. NS methods are not able to train on VQA due to lack of scene graph and program annotations (marked with dashes). NS methods show bad transfer performance, especially on the open-ended questions, mainly due to the high ratio of programs that cannot find the queried objects from the generated scene graphs (marked in parentheses). VLE2E has less accuracy drop compared to NS.

is not the primary reason for NS struggles. Most of these questions involve commonsense reasoning, such as asking why some event happens in the image (e.g., “*Why is the man on the street?*” where the answer is “*homeless*”). We also note that we only evaluate on the binary and open questions of VQA but exclude the counting questions, which are about 13% of the dataset. GQA has no counting questions, so the semantic parser trained on GQA cannot generate counting operations.

How much manual adaption is required to transfer NS systems to a new benchmark? NS systems require manual adaptation for different datasets. From the transfer between GQA and VQA, we show little manual adaption is required on the language side of NS systems to transfer between benchmarks of the same task. With some entity matching mechanism between semantically similar objects and a stronger scene graph generation module that generalizes well between datasets, NS might be possible to transfer well.

6 Discussion and Conclusion

In conclusion, VLE2E training systems do not learn precise reasoning, which inhibits their generalization ability under small perturbations to either language or vision. Though the in-domain results of NS systems are usually slightly worse than VLE2E systems, the NS methods are more robust on most of the generalization tests we de-

velop here. Even when the performance of NS methods drops on some OOD data, they can still quickly recover by few-shot training. Nonetheless, VLE2E systems still achieve better performance on cross-benchmark transfer, while NS methods struggle when test questions require novel program constructs or scene graph object types.

Our work highlights the importance of evaluating on a diverse set of metrics besides in-distribution accuracy, in line with recent work on improving leaderboards (Ethayarajh and Jurafsky, 2020; Ma et al., 2021). Our analysis suggests that we should not expect in-domain and out-of-domain accuracies to be strongly correlated when evaluating very different types of models, such as VLE2E and NS models, in contrast with Miller et al. (2020, 2021). Finally, we hope our observation that end-to-end and neuro-symbolic systems have complementary generalization advantages will inspire the community to design more robust VL reasoning systems that share the benefits of both approaches.

Acknowledgements

This work was supported in part by the NSF (RI AWD-00001042).

Limitations

Most of our experiments focus on datasets with synthetic language annotations. In particular, GQA and COVR both use synthetic language, while VQA has human-written questions. Existing VL reasoning datasets with natural language questions do not have annotated functional programs and scene graphs. Since we use NS systems must be trained on annotated programs, we cannot easily extend our work to these other datasets. One possible solution would be to adapt other single-image NS methods (e.g., NSM (Hudson and Manning, 2019b)) that do not require program and scene graphs annotation to the multi-image setup.

Our evaluation requires a custom modification of the semantic parsing language on GQA and COVR. To apply similar evaluations to other datasets, if their program annotations are not directly applicable to our NS system, practitioners might need to make similar task-specific modifications.

Finally, all of our experiments are on English-only data, which requires limited morphological reasoning reasonable semantic parsing. The results and conclusions might not be applicable to other language with richer morphology.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- Yonatan Bitton, Gabriel Stanovsky, Roy Schwartz, and Michael Elhadad. 2021. Automatic generation of contrast sets from scene graphs: Probing the compositional consistency of GQA. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Ben Bogin, Shivanshu Gupta, Matt Gardner, and Jonathan Berant. 2021. COVR: A test-bed for visually grounded compositional generalization with real images. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*.
- Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- 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*.
- Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. 2021. Unlocking compositional generalization in pre-trained models using intermediate representations. *arXiv*.
- Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations (ICLR)*.
- Drew A. Hudson and Christopher D. Manning. 2019a. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Computer Vision and Pattern Recognition (CVPR)*.
- Drew A. Hudson and Christopher D. Manning. 2019b. Learning by abstraction: The neural state machine. In *Neural Information and Processing Systems (NeurIPS)*.
- Carlos E. Jimenez, Olga Russakovsky, and Karthik Narasimhan. 2022. CARETS: A consistency and robustness evaluative test suite for VQA. In *Association for Computational Linguistics (ACL)*.
- 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 *Computer Vision and Pattern Recognition (CVPR)*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Association for Computational Linguistics (ACL)*.
- Linjie Li, Zhe Gan, and Jingjing Liu. 2020. A closer look at the robustness of vision-and-language pre-trained models. *arXiv*.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems (NeurIPS)*.
- Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking. In *Neural Information Processing Systems (NeurIPS)*.
- Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations (ICLR)*.
- John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning (ICML)*.
- John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning (ICML)*.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *arXiv*.

- 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 (JMLR)*, 21.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NeurIPS)*.
- Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Association for Computational Linguistics (ACL)*.
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Association for Computational Linguistics (ACL)*.
- 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 *Association for Computational Linguistics (ACL)*.
- Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.
- Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Computer Vision and Pattern Recognition (CVPR)*.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022a. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Computer Vision and Pattern Recognition (CVPR)*.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022b. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. 2017. Neural scene de-rendering. In *Computer Vision and Pattern Recognition (CVPR)*.
- Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Computer Vision and Pattern Recognition (CVPR)*.
- Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. Neural-Symbolic VQA: Disentangling reasoning from vision and language understanding. In *Neural Information and Processing Systems (NeurIPS)*.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Computer Vision and Pattern Recognition (CVPR)*.

A Compositional Logical Forms

We create compositional logical forms as an intermediate representation of the original logical forms. We explain how to design new operations and default value grammar checker below.

A.1 Operation Modifications

As shown in Table 7, we refactor the quantifier operations to a map operation with logical or or logical and as arguments. To compositionally represent the original none operation, we introduce a new operation `logic_not`, which takes a boolean variable and outputs the negation. This enables the negation of the original operation `all`, and also enables the contrast set creation from *all* \rightarrow *either none or only some*.

We refactor and create `choose`, `query`, `verify`, `filter`, `keep_if_values_count` and `compare` operations following the same pattern. We merge the redundant operation `relation_between_nouns` with the `choose` operation and replace two sanity check operations as `unique` and `assert_unique` with an automatic grammar checker. For the rest of operations, we keep them in CLF as the original version.

A.2 Default Value Grammar Checker

The default value grammar checker takes the sanity check responsibility from the original `unique` operation. However, unlike the original `unique`, which raises an error when the queried object is not unique, the grammar checker automatically fixes the error by taking the first object as the queried object. For example, for a query “*Is the boy wearing a hat?*”, if the `find` operation returns multiple “*boy*” nodes, the grammar checker automatically choose the first “*boy*” node and records an error. Otherwise, if the `find` operation returns zero “*boy*” nodes, the grammar checker will raise an “object not found” error, same as the missing object error in Table 5.

To make sure the program is executable, the default value grammar checker also assigns a default value for the arguments for each operation. The default value is *0* for the integer type and *False* for the boolean type. For example, if a `compare` operation has only one argument, the value will directly compare to *0*.

B Experiment Details

B.1 Dataset Statistics

The image data of COVR consists of GQA and imSitu (Yatskar et al., 2016). There are about 275k images in total. VQA has human annotated queries, and the language annotation of GQA and COVR is generated by template. VQA has 440k training questions, 214k validation questions and 448k testing questions. GQA has 943k training questions, 132k validation questions and 95k testing questions. For COVR, each example contains 1 to 5 query images, and there are 248k training questions, 7k validation questions and 7k testing questions.

For datasets with scene graph annotations, GQA and COVR, before generating scene graphs for the images in each dataset’s validation set, both the object detection and the relation prediction modules in the scene graph generator are tuned on the training set. For datasets without scene graph annotations, VQA, we directly apply the scene graph generator trained on GQA to generate scene graphs, and apply the constructed dictionary on GQA to map objects.

Model	# Params	Running time per experiment		
		COVR	GQA	VQA
T5	220M	15	15	-
GPT-2	117M	15	15	-
BART	110M	16	16	-
VinVL	-	-	0	-
ViLBERT	-	17	0	-
VisualBERT	-	17	0	-
LXMERT	-	-	4	8

Table 6: GPU hours are computed at 1 Quadro RTX 6000 GPU. *0* indicates the model is downloaded without further training.

B.2 Hyperparameters

The model parameters and GPU hours are listed in Table 6. We search hyperparameters manually, one per trial. For all language models, we choose a batch size of 256, train 30000 iterations at maximum, with an early-stopping with patience 3. We save the model each 500 iterations. For T5, we use an Adam optimizer with a learning rate $1e-4$. For GPT-2, The learning rate is $5e-5$ for GPT-2 and $2e-5$ for BART.

For all vision models on COVR, we choose a batch size of 12, train 8 epochs with no early-stopping and save the best model from the eval-

uation after each epoch. We use the AdamW optimizer with a learning rate of $1e-6$ and a weight decay of $1e-3$ for VisualBERT and ViLBERT. VisualBERT and ViLBERT take a batch size of 32 on GQA with the same AdamW optimizer and learning rate. LXMERT takes a batch size of 32 on GQA and VQA with the Adam optimizer and a learning rate of $1e-5$.

B.3 Few-shot Training

Corresponding to Figure 3, the whole few-shot training table is Table 8.

Original operation	Compositional operation	Additional arguments	Additional comments
some	map	or	
all	map	and	
none	map	or	add logic_not in front of it
choose_name	choose	name	
choose_attr	choose	attr	
choose_relation	choose	rel	
query_name	query	name	
query_attr	query	attr	
verify_attr	verify	attr	
with_relation	filter	rel	
with_relation_object	filter	rel, other	
filter	filter	attr	
relation_between_nouns			merged with choose_relation
find	find		
count	count		
keys	keys		
unique_images	unique_images		
group_by_images	group_by_images		
scene	scene		
exists	exists		
logic_or	logic_or		
logic_and	logic_and		
	logic_not		new operation
keep_if_values_count_eq	keep_if_values_count	eq	
keep_if_values_count_geq	keep_if_values_count	geq	
keep_if_values_count_leq	keep_if_values_count	leq	
eq	compare	eq	
geq	compare	geq	
leq	compare	leq	
lt	compare	lt	
gt	compare	gt	
unique			replaced with grammar checker
assert_unique			replaced with grammar checker

Table 7: CLF Creation. Refine 33 original operations into 18 compositional operations with additional arguments.

Template	Transfer Setting	K-shot	VLE2E	NS
			VisualBERT	CLF w/ T5
COUNTGROUPBY	<i>at least → no less than</i>	0	54.0	21.5 [39.4]
		1	53.3	28.8 [47.5]
		5	56.3	39.2 [70.3]
VERIFYCOUNT	<i>at least → no less than</i>	0	68.3	48.4 [67.8]
		1	68.8	55.1 [77.9]
		5	78.7	69.8 [91.1]
	<i>at least → less than</i>	0	21.1	52.9 [65.6]
		1	20.1	58.0 [73.3]
		5	45.5	70.7 [91.5]
VERIFYCOUNTGROUPBY	At least → No less than	0	65.8	55.6 [76.2]
		1	66.2	59.5 [82.3]
		5	68.3	65.2 [92.0]
	<i>at least → less than</i>	0	28.5	51.1 [50.6]
		1	29.4	58.0 [70.5]
		5	51.9	66.5 [86.9]
QUANTIFIER	<i>no ↔ some</i>	0	74.9	75.6 [97.8]
		1	73.5	76.1 [99.5]
		5	75.4	76.1 [99.5]
	<i>no → at least one</i>	0	83.9	53.9 [61.1]
		1	84.5	61.2 [80.2]
		5	84.6	71.3 [95.1]
	<i>some → none of the</i>	0	67.8	75.5 [97.8]
		1	67.3	75.8 [98.6]
		5	66.3	75.9 [99.5]
	<i>all → either none or only some</i>	0	36.2	57.5 [69.8]
		1	52.5	69.2 [89.0]
		5	56.0	73.9 [97.9]

Table 8: Few-shot Contrast Set on COVR: Comparison of VisualBERT and the NS method on language contrast sets of four templates on COVR. NS performance quickly improve from 1-shot training, VLE2E instead, needs more examples to learn, except for 1 unnatural perturbation, *all→either none or only some*.