Uni-NLX: Unifying Textual Explanations for Vision and Vision-Language Tasks (Supplementary Material)

Fawaz Sammani and Nikos Deligiannis
ETRO Department, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
imec, Kapeldreef 75, B-3001 Leuven, Belgium
fawaz.sammani@vub.be, ndeligia@etrovub.be

1. Prompts

In this section, we provide the prompts we use to formulate the VQA-ParaX and ImageNetX Natural Language Explanation (NLE) datasets.

VQA-ParaX : We prompt the Large Language Model (LLM) with <I, S^i>. This consists of the paragraph sample S^i and the instruction I. The instruction I is constructed with the following considerations:

- An overview of the task that the LLM has to perform
- A guideline to generate a short answer, typical of the standard Visual Question Answering (VQA) scenario
- A guideline to avoid trivial cases
- An example of the task and of the trivial case, following the few-shot learning paradigm.
- An output format to facilitate the post-processing stage.

I is formulated as follows: `You are an assistant which helps formulate a VQA dataset with Textual Explanations to train deep learning models. Read the following text and formulate 3 samples, as unique as possible, each consisting of a question (Q), answer (A) and more information about the answer to help in better understanding it (E). The answers should be short, maximum of 3 words. Here is an example for Q, A and E, respectively: Q: What sport is being played?, A: baseball, E: they are playing on a baseball diamond with a ball and a bat. Also, E should be non-trivial. For example, if Q is: Where is the green tennis ball? and A is: above her head, then E should NOT BE: there is a green tennis ball above the woman's head. This is considered as trivial. Please generate the output in a single line strictly following this format for the 3 samples, where <r> indicates your response: [Q:<r>, A:<r>, E:<r>, Q:<r>, A:<r>, E:<r>, Q:<r>, A:<r>, E:<r>]. Here is the text:`

It is worth noting that in preliminary stages of this work, the sub-instruction: `and more information about the answer to help in better understanding it (E)` was formulated as `and an explanation (E) to explain the answer (A)`. However, we observed that in the majority of cases, the outputs from the LLM were primarily focused on its own reasoning process, consistently generating trivial statements: "the text describes that..." or "it is mentioned in the text that...". Consequently, we decided to avoid this particular sub-instruction from further consideration.

ImageNetX : We prompt the LLM with <I, c>, where I represents the instruction and c represents the class category. The instruction I consists of the following:

- An overview of the task that the LLM has to perform
- An additional guideline to tune the generated output to be general and short, avoiding an extensive enumeration of individual elements pertaining to class c

The instruction I is formulated as: `You are an assistant which helps humans describe objects. what are physical features and characteristics describing a` c`? Please answer in a short, brief and concise way, with a maximum of 50 words.`

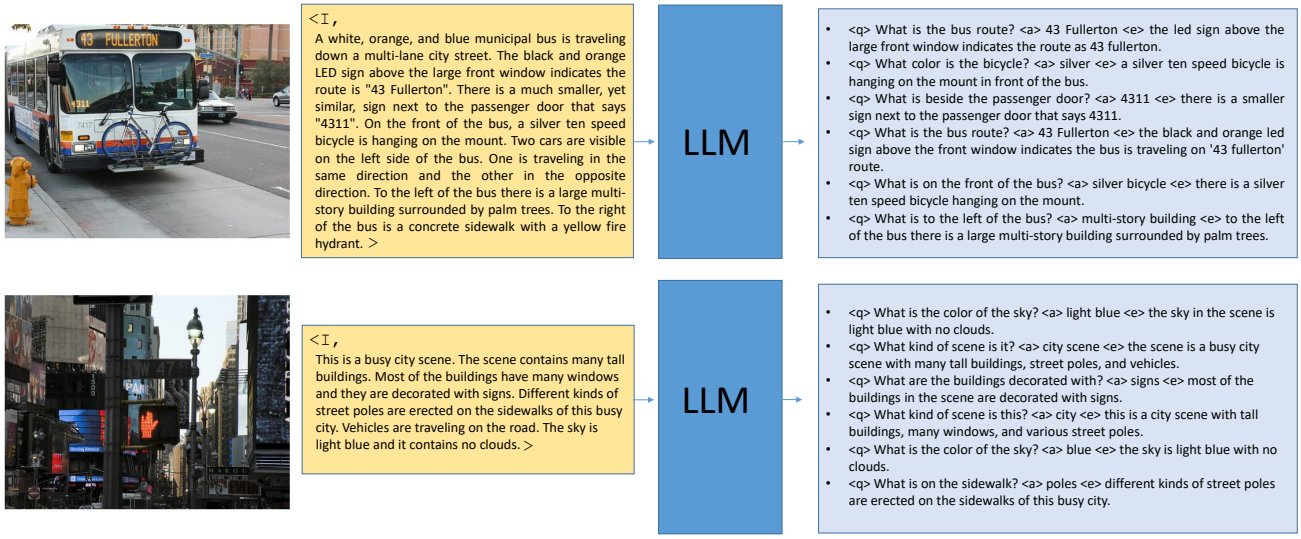

Figure 1. The process of generating VQA-ParaX leveraging a Large Language Model (LLM). The instruction `I` and the text fragment describing the image jointly serve as the input prompt (yellow box) for the LLM, which reformulates the text fragment into 6 samples, each consisting of a question $< q >$, answer $< a >$ and explanation $< e >$. It is important to note that the image is not provided to the LLM.

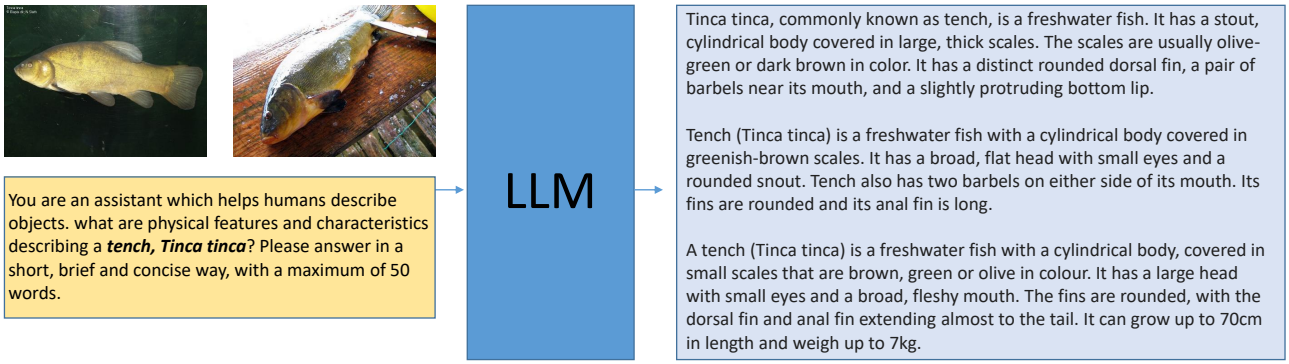

Figure 2. The process of generating ImageNetX leveraging a LLM. The instruction containing the ImageNet category (*tench* in this example) serves as the input prompt (yellow box) for the LLM, which outputs distinctive features describing that category. Although 3 generated samples are shown, it is important to clarify that we generate a single sample for each run. It is important to also note that no image is provided to the LLM.

2. Qualitative Data Samples

Figure 1 shows 2 examples depicting the process of generating VQA-ParaX. The instruction `I` and the text fragment describing the image jointly serve as the input prompt for the LLM. The output of the LLM is a re-formulation of the text fragment into 6 samples, each consisting of a question $< q >$, answer $< a >$ and explanation $< e >$. Similarly, Figure 2 shows an example illustrating the process of generating ImageNetX. The instruction containing the ImageNet category serves as the input prompt for the LLM. The output is a textual description describing attributes and distinctive features of that category.

3. Data Analysis and Quality Assessment

In this section, we perform analysis on the newly introduced datasets VQA-ParaX and ImageNetX. Subsequently, we evaluate their quality through ablation experiments. Table 1 presents the average and maximum word lengths of the explanations for both VQA-ParaX and ImageNetX. As observed, the average word length of VQA-ParaX explanations is similar to the average word length of explanations from other NLE datasets. ImageNetX has a larger average word length describing distinctive features requires more words. In Table 2, we present question repetition statistics on VQA-ParaX. As the prompt requests 6 samples, the LLM might reiterate the constructed questions when the

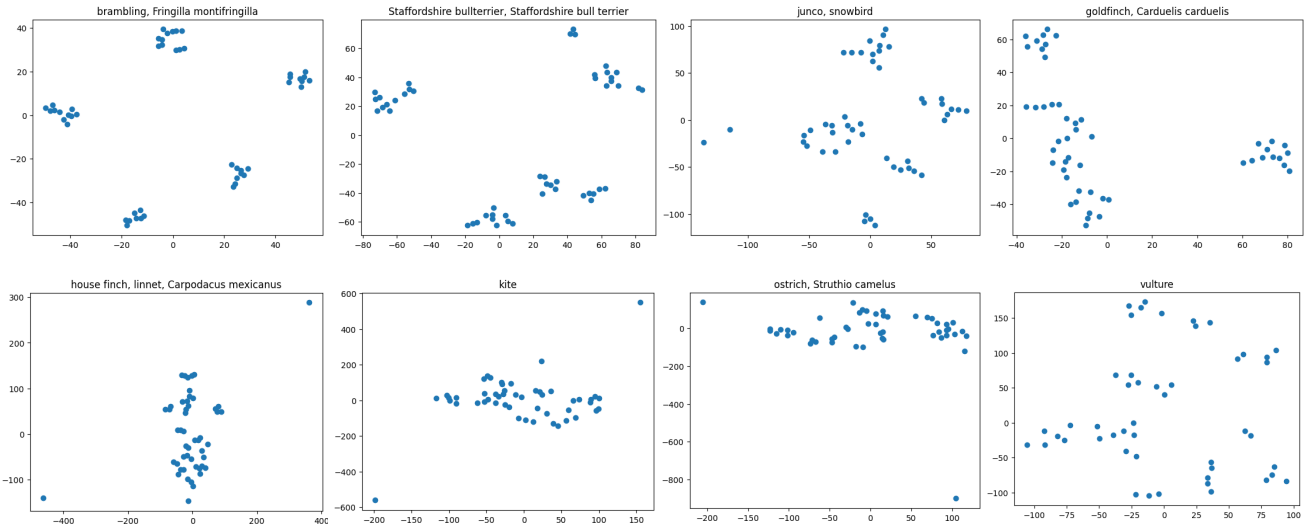

Figure 3. A 2D visualization using t-SNE [3] of the different textual descriptions for ImageNet categories generated using the LLM. Each plot depicts a distinct ImageNet category, with each data point representing 1 of 50 sample descriptions produced for each category.

provided textual description of the image is overly brief and lacks information (*e.g.,* a textual description as: *a man in a white t-shirt and blue jeans and a cell phone*). However, even though the question and answer might be replicated across a sample, the explanation is typically formulated differently.

Next, we examine the uniqueness of the 50 different samples we generate for each ImageNet category. We first encode each sample through MPNet [2] finetuned on 1B sentence pairs using the self-supervised contrastive learning objective, utilizing the Sentence-Transformers [1] library[1] to obtain a 768-d vector representing the sample description. We apply t-SNE [3] to reduce the vector into a 2-d space for visualization. Upon plotting the 50 samples, we observe distinct clusters emerging among them, highlighting the diversity present across the samples, as illustrated in the upper row. Conversely, the initial three instances depicted in the lower row reveal that these samples tend to exhibit greater similarity, resulting in the formation of a singular cluster.

Lastly, we perform ablation studies on the newly introduced datasets VQA-ParaX and ImageNetX in Table 3. We start by analyzing the exclusion of VQA-ParaX from the training NLE corpus. In 60% of the cases (across all datasets and metrics), the inclusion of VQA-ParaX improves the evaluation metrics across the board. This suggests that VQA-ParaX contribute positively to the performance of the model. Next, we evaluate the exclusion of ImageNetX. In 94% of the cases, excluding ImageNetX improves the performance of evaluation metrics across all VQA NLE tasks (A-OKVQA, VQA-X and VQA-ParaX), and including it improves the performance of visual recognition and visual reasoning tasks in 84% of the cases. This suggests that ImageNetX has a negative impact on VQA tasks, but positive impact on the other tasks. This can be rationalized by considering the complexity of ImageNetX, which requires the model to additionally learn how to describe coarse-grained distinctive textual features of an object. This task is more challenging compared to VQA tasks. As a result, incorporating a broader range of such complex information might lead to a trade-off, resulting in a decrease in performance for VQA tasks. Finally, we investigate the performance of the model when excluding both VQA-ParaX and ImageNetX. We find that in 58% of the cases, excluding both these datasets leads to an improvement in performance.

[1]https://github.com/UKPLab/sentence-transformers

Table 1. Average Word Length of VQA-ParaX and ImageNetX

	VQA-ParaX	ImageNetX
Average Word Length	13	49
Maximum Word Length	90	110

Table 2. Question Repetition statistics for VQA-ParaX

Repetitions	**Value**
Maximum repetitions across all samples	3
Percentage of samples with 3 repetitions	1.96%
Percentage of samples with 2 repetitions	12.96%
Percentage of samples with 1 repetition	36.84%
Percentage of samples with no repetitions	48.23%

4. Additional Quantitative Evaluation

In Table 4, we present quantitative evaluation results on unfiltered scores for Uni-NLX, achieved through fine-

Table 3. Ablation Studies on the newly introduced datasets (Unfiltered Scores, w/o pretraining). B-N, M R, C, S are short for: BLEU-N, METEOR, ROUGE-L, CIDER and SPICE.

Dataset	Setting	B1	B4	M	R	C	S
ACT-X	All Data (Uni-NLX)	0.654	0.265	**0.220**	0.485	0.677	**0.167**
	w/o VQA-ParaX	**0.658**	0.265	0.219	0.484	0.680	0.166
	w/o ImageNetX	0.656	0.263	0.219	0.483	0.675	**0.167**
	w/o VQA-ParaX, ImageNetX	0.655	**0.271**	0.219	**0.486**	**0.685**	0.165
A-OKVQA	All Data (Uni-NLX)	**0.582**	0.185	0.171	0.440	0.581	0.160
	w/o VQA-ParaX	0.561	**0.209**	0.168	**0.458**	**0.652**	0.152
	w/o ImageNetX	0.576	0.194	**0.173**	0.445	0.608	**0.161**
	w/o VQA-ParaX, ImageNetX	0.558	0.198	0.166	0.455	0.624	0.155
VQA-X	All Data (Uni-NLX)	0.579	0.217	0.194	0.459	0.811	0.178
	w/o VQA-ParaX	0.578	0.224	0.196	0.463	0.833	0.176
	w/o ImageNetX	**0.588**	**0.232**	**0.202**	**0.472**	**0.865**	**0.182**
	w/o VQA-ParaX, ImageNetX	0.578	0.221	0.196	0.462	0.818	0.179
VQA-ParaX	All Data (Uni-NLX)	0.351	0.148	0.182	0.408	1.399	**0.316**
	w/o VQA-ParaX	0.165	0.058	0.120	0.319	0.769	0.230
	w/o ImageNetX	**0.360**	**0.151**	**0.183**	**0.409**	**1.416**	**0.316**
	w/o VQA-ParaX, ImageNetX	0.164	0.058	0.118	0.317	0.758	0.226
e-SNLI-VE	All Data (Uni-NLX)	**0.353**	**0.118**	**0.178**	**0.322**	1.065	0.313
	w/o VQA-ParaX	0.343	0.112	0.174	0.318	1.044	0.308
	w/o ImageNetX	0.327	0.104	0.169	0.311	1.023	0.307
	w/o VQA-ParaX, ImageNetX	0.348	0.115	0.177	0.321	**1.066**	**0.319**

tuning the pretrained captioning model of NLX-GPT. Our findings demonstrate that Uni-NLX achieves results comparable to NLX-GPT on VQA-X, ACT-X, and VQA-ParaX. Furthermore, it surpasses NLX-GPT performance on e-SNLI-VE and ImageNetX. Additionally, in the A-OKVQA task, our model outperforms NLX-GPT across three metrics, while also achieving comparable results on other metrics. In Table 5, we provide results of Uni-NLX for the filtered setting without finetuning the pretrained captioning model. Our findings reveal that Uni-NLX surpasses NLX-GPT on ACT-X, e-SNLI-VE, and VCR, while achieving comparable results on the other tasks. By conducting a comparative analysis of both settings, it becomes evident that Uni-NLX exhibits superior performance in reasoning tasks such as e-SNLI-VE and VCR, as well as in visual recognition tasks such as ImageNetX and ACT-X.

5. Additional Qualitative Examples

We provide additional qualitative examples for each of the seven NLE tasks in Figure 4. As evidenced in our observations, our model generates an answer to the provided question about a given image, complemented by an explanation. For ImageNetX, it becomes apparent that Uni-NLX offers distinctive, fine-grained explanations for the predicted answer (*e.g.,* white head and tail, dark brown body, yellow beak, large wingspan, the weight), all conveyed in a manner easily understandable to humans. In Figure 5, we visualize the attention maps for the generated answers from the last layer of the model. We analyze ImageNetX answers in the top row, VQA-ParaX answers in the bottom-left and ACT-X answers in the bottom-right. As demonstrated for ImageNetX, the presented heatmaps exhibit distinctive features within the image, in contrast to conventional explainability techniques that usually yield heatmaps encompassing the entire main object in the image.

6. Limitations and Collapse of VQA tasks

Acknowledging the limitations of our proposed model is of crucial importance. In particular, we address the issue of *shortcut learning* that arises in explanations for some VQA tasks including A-OKVQA and VQA-ParaX. Specifically, the model generates an explanation which is composed of the answer and the question itself. For instance, consider the question "what is on the table?" and the predicted answer "cake", the generated explanation would be "there is cake on the table". Similarly, when presented with the question "what is on top of the plate?" and the predicted answer "pizza", the generated explanation would be "there is pizza on top of a plate". By using this shortcut approach, the model fails to reason correctly about the generated answer. This phenomenon becomes further evident by examining Figure 6, wherein we conduct an analysis of the heatmaps for various questions of the same images shown in Figure 5. As illustrated, the heatmaps exhibit no distinctions from those depicted in Figure 5, thereby indicating a lack of reasoning capabilities in generating explanations. Conse-

Figure 4. Qualitative Examples of Uni-NLX on the 7 NLE tasks. We show the *question*, **answer** and explanation under each image.

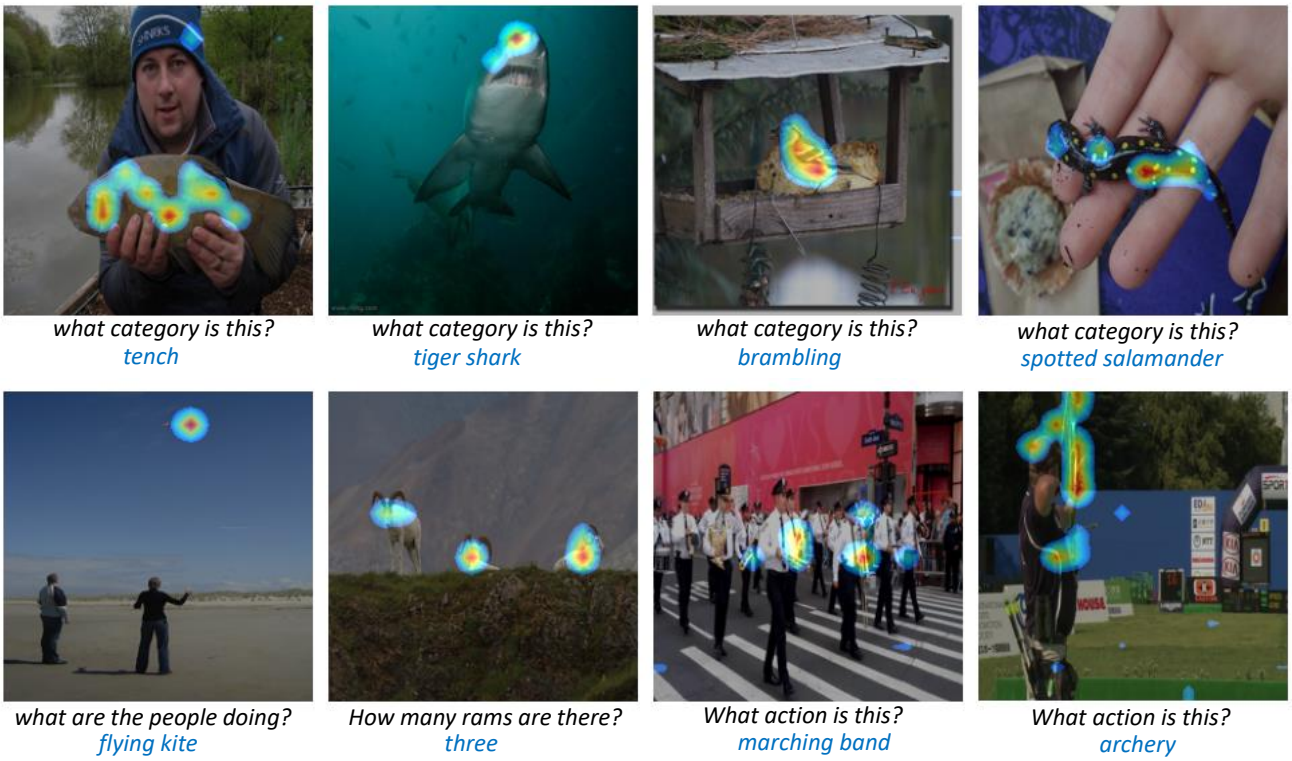

Figure 5. The attention maps for the generated answers of ImageNetX (top row), VQA-ParaX (bottom-left) and ACT-X (bottom-right).

quently, this finding corroborates the presence of the shortcut learning problem in the model. This problem is also observed in the individual uni-task models for A-OKVQA and VQA-ParaX. We intend to explore this matter in future research.

Table 4. Unfiltered Scores for Uni-NLX compared to NLX-GPT on the 7 downstream tasks. Both models are w/ pretraining. B-N, M R, C, S are short for: BLEU-N, METEOR, ROUGE-L, CIDER and SPICE.

	VQA-X							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**61.2**	**46.1**	**34.3**	**25.6**	**21.5**	**48.7**	**97.2**	**20.2**
Uni-NLX	60.2	44.7	32.8	24.1	20.8	47.2	89.9	19.5
	ACT-X							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**67.0**	**50.5**	**37.5**	**28.1**	**22.6**	**49.7**	**74.9**	**17.6**
Uni-NLX	66.6	**50.5**	37.3	27.7	22.4	49.3	72.5	17.2
	e-SNLI-VE							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**34.3**	**22.7**	15.6	10.9	**17.5**	31.7	106.6	**31.5**
Uni-NLX	33.9	**22.7**	**15.8**	**11.3**	**17.5**	**32.1**	**107.5**	**31.5**
	VQA-ParaX							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**37.9**	**28.0**	**21.5**	**16.6**	**19.5**	**42.5**	**156.6**	**34.0**
Uni-NLX	36.8	27.2	20.8	16.1	19.1	42.0	152.6	33.5
	A-OKVQA							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	57.1	**41.1**	**30.4**	**21.7**	17.4	**46.8**	**69.0**	16.0
Uni-NLX	**58.6**	40.2	28.2	18.9	**17.5**	44.8	61.1	**16.9**
	ImageNetX							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	61.7	45.2	34.2	26.4	20.7	37.6	76.4	20.2
Uni-NLX	**63.2**	**47.0**	**36.0**	**28.2**	**21.4**	**38.9**	**82.8**	**21.1**
	VCR							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	-	-	-	-	-	-	-	-
Uni-NLX	19.1	10.1	5.8	3.6	9.1	20.0	24.9	12.5

Figure 6. An example of the shortcut learning problem

Table 5. Filtered Scores for Uni-NLX compared to NLX-GPT on the 7 downstream tasks. Both models are w/o pretraining.

	VQA-X							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**63.3**	**48.5**	**36.9**	**28.1**	**22.6**	**50.9**	**108.5**	**21.2**
Uni-NLX	60.3	45.0	33.0	24.2	20.7	48.2	91.9	19.5
	ACT-X							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	69.5	53.5	40.7	31.3	24.8	52.3	99.6	20.9
Uni-NLX	**70.6**	**55.6**	**42.7**	**32.9**	**25.3**	**52.9**	**104.6**	**22.4**
	e-SNLI-VE							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	35.7	24.0	16.8	11.9	18.1	33.4	**114.7**	**32.1**
Uni-NLX	**36.3**	**24.8**	**17.6**	**12.8**	**18.4**	**33.8**	114.6	32.0
	VQA-ParaX							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**40.7**	**30.0**	**23.2**	**18.2**	**21.1**	**45.6**	**187.5**	**40.1**
Uni-NLX	38.5	28.4	21.7	17.0	20.6	44.9	179.8	39.6
	A-OKVQA							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**61.1**	**45.7**	**34.8**	**26.1**	**19.9**	**51.0**	**89.1**	**19.6**
Uni-NLX	62.8	43.6	30.7	21.3	19.1	47.9	73.5	19.4
	ImageNetX							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	**72.3**	**56.7**	**45.0**	**36.2**	**26.0**	44.6	**119.1**	**28.1**
Uni-NLX	71.4	55.9	44.3	35.6	25.8	**44.7**	117.7	27.9
	VCR							
	B1	B2	B3	B4	M	R	C	S
NLX-GPT	24.7	15.0	9.6	6.6	12.2	26.4	46.9	18.8
Uni-NLX	**25.3**	**19.6**	**16.1**	**14.0**	**14.9**	**30.5**	**66.1**	**20.5**

References

[1] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 3

[2] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *ArXiv*, abs/2004.09297, 2020. 3

[3] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 3