NeIn: Telling What You Don't Want

CVPR 2025 Workshop Syntagen

Nhat-Tan Bui[1], Dinh-Hieu Hoang[2], Quoc-Huy Trinh[3,4], Minh-Triet Tran[2], Truong Nguyen[5], Susan Gauch[1]

[1]University of Arkansas, USA [2]University of Science, VNU-HCM, Vietnam

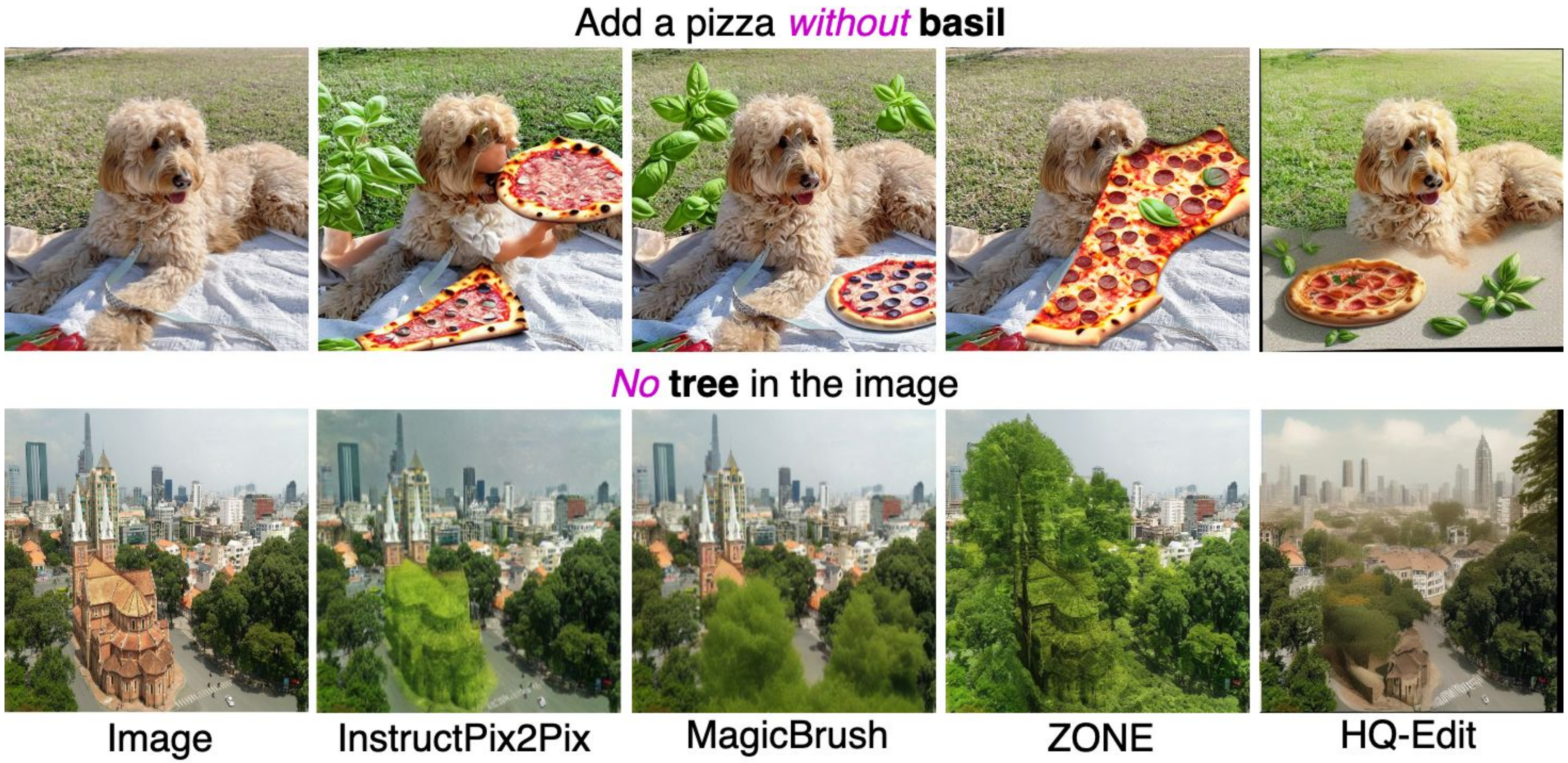
[3]Aalto University, Finland [4]SpexAI GmbH, Germany [5]University of California, San Diego, USA

Text-guided Image Editing

Given **an image** and **a textual instruction**, the model is able to make appropriate edit based on the instruction.

Image Credit: [InstructPix2Pix: Learning to Follow Image Editing Instructions]

Negation Understanding



The failures of recent text-guided image editing methods in understanding the **negative queries**.

Motivation

Datasets	Tasks	Train		Validation		Total	
		#Negative	#All	#Negative	#All	#Negative	#All
CC12M [3]	Pre-training	–	–	–	–	314,181 (2.53%)	12,423,374
LAION-400M [22]	Pre-training	–	–	–	–	2,404,784 (0.58%)	413,862,224
MS-COCO'14 [11]	Image Captioning	1,761 (0.43%)	414,113	886 (0.44%)	202,654	–	616,767
SBU Captions [17]	Image Captioning	–	–	–	–	26,222 (2.62%)	1,000,000
CC3M [23]	Image Captioning	54,219 (1.63%)	3,318,333	–	–	–	3,369,218
CIRR [13]	Composed Image Retrieval	868 (3.08%)	28,225	130 (3.11%)	4,181	–	36,554
InstructPix2Pix[1]	Image Editing	77 (0.02%)	313,010	–	–	–	313,010
MagicBrush [28]	Image Editing	54 (0.61%)	8,807	6 (1.17%)	528	–	10,388

Statistic of captions in current image-caption pair datasets. The number of **negative sentences** in those datasets is very small.

→ We need a dataset specifically for **negation understanding** in **vision-language tasks**.

Contributions

1. We investigate the ability of VLMs to **interpret negation cues** in text-guided image editing, leading to the creation of the first large-scale vision-language negation dataset for this task, termed **NeIn**.

2. We introduce a **pipeline** to generate NeIn, an extensive dataset comprising **366,957** quintuplets. This dataset focuses on the understanding of negation, a fundamental linguistic concept, for image editing VLMs.

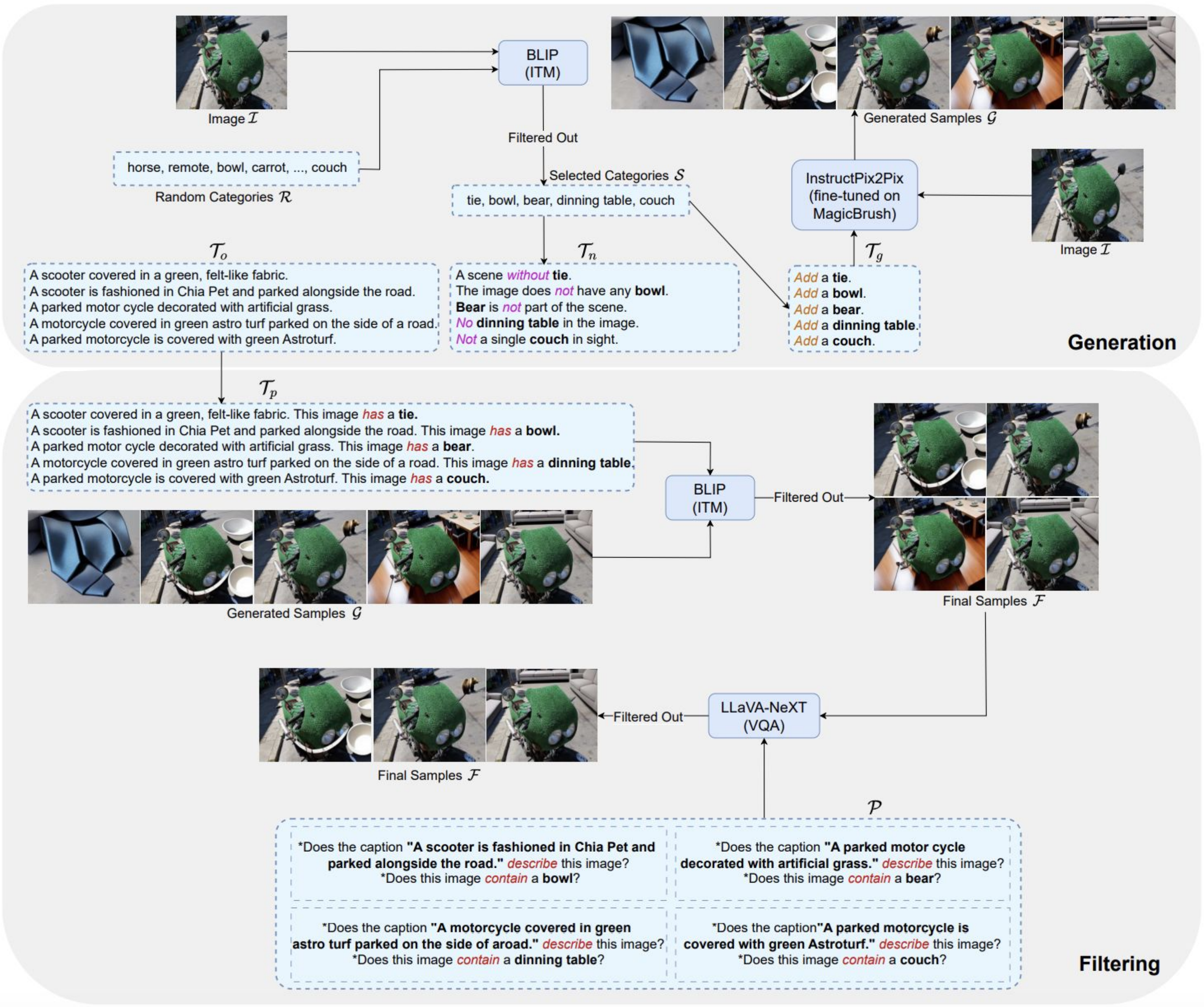
3. We propose an **evaluation method** for negation understanding that can be used by future researchers. Using our evaluation method, we observe that VLMs in image editing task have **difficulty comprehending** negative instructions. This insight opens a new research direction for improving negation understanding for VLMs.

Negative Instruction (NeIn)

We present NeIn, the **first** large-scale vision-language negation dataset for image editing.

It comprises of **366,957** samples with **342,775** queries for training and **24,182** queries for benchmarking.

The creation of NeIn involves two primary stages: **generation** and **filtering**.



Negative Instruction (NeIn)

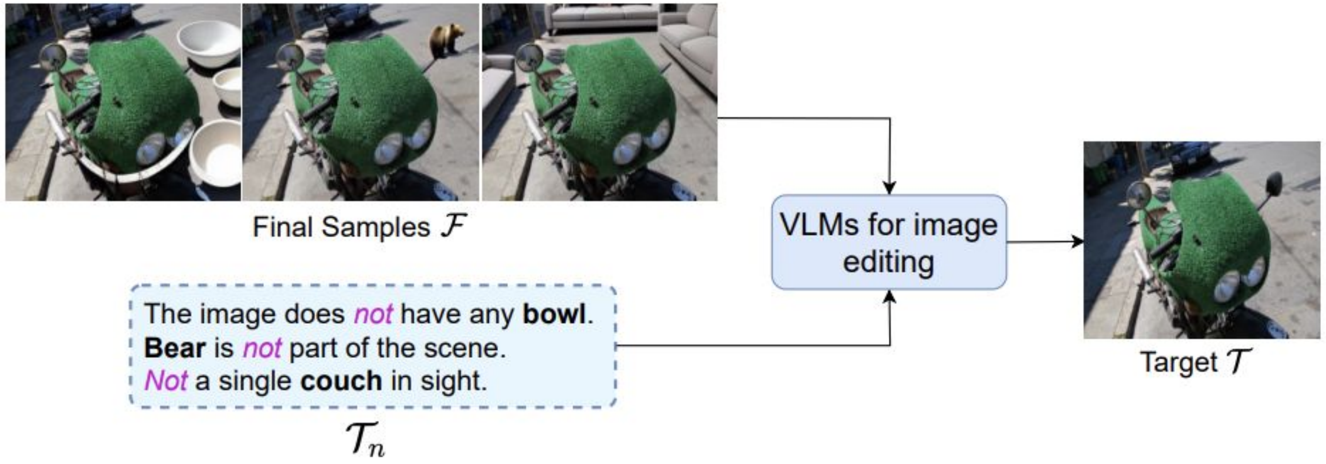


Illustration for fine-tuning and benchmarking process.

Proposed Evaluation Metrics

We consider whether image editing methods:

1. Can **eliminate** the object categories *specified* in the negative sentence.
2. Can **preserve** the object categories *not mentioned* in the negative sentence.

The first is determined by the **Removal Evaluation**, while the second is assessed using the **Retention Evaluation**.

Since the purpose of both metrics is to **identify** objects, we consider the **visual question answering (VQA)** and the **open-vocabulary object detection (OVD)**.

Removal Evaluation

Algorithm 3 Removal Evaluation by VQA

Input:
$\mathcal{T}$: considered model's outputs
$\mathcal{S}$: objects to be removed
Output:
s: removal score

$s := 0$
for each tuple $(\mathcal{T}^{(i)}, \mathcal{S}^{(i)})$ in $(\mathcal{T}, \mathcal{S})$ **do**
    # pre-defined prompt
    $p \leftarrow$ "Does this image contain a/an $\mathcal{S}^{(i)}$?"
    if VQA($\mathcal{T}^{(i)}, p$) = "No" **then** ▷ Object is removed
        $s \leftarrow s + 1$
    end if
end for
$s \leftarrow s/|\mathcal{T}|$
return s

Algorithm 5 Removal Evaluation by OVD

Input:
$\mathcal{T}$: considered model's outputs
$\mathcal{S}$: objects to be removed
Output:
s: removal score

$s := 0$
for each tuple $(\mathcal{T}^{(i)}, \mathcal{S}^{(i)})$ in $(\mathcal{T}, \mathcal{S})$ **do**
    $p \leftarrow$ OVD($\mathcal{T}^{(i)}, \mathcal{S}^{(i)}$) ▷ Prediction list
    if length of $p = 0$ **then** ▷ Object is removed
        $s \leftarrow s + 1$
    end if
end for
$s \leftarrow s/|\mathcal{T}|$
return s

Retention Evaluation

Algorithm 4 Retention Evaluation by VQA

Input:
$\mathcal{F}$: samples of NeIn
$\mathcal{T}_o$: original caption from MS-COCO
$\mathcal{T}$: considered model's outputs
Output:
s: retention score

$s := 0$
for each tuple $(\mathcal{F}^{(i)}, \mathcal{T}_o^{(i)}, \mathcal{T}^{(i)})$ in $(\mathcal{F}, \mathcal{T}_o, \mathcal{T})$ **do**
  list^1 :=[], list^2 :=[]
  $\mathcal{O} \leftarrow$ extractor($\mathcal{T}_o^{(i)}$) ▷ Original objects in $\mathcal{I}$
  # check $\mathcal{O}$ in $\mathcal{F}$
  for each *object* in $\mathcal{O}$ **do**
    $p \leftarrow$ "Does this image contain a/an *object* ?"
    $b \leftarrow$ VQA($\mathcal{F}^{(i)}, p$) ▷ Boolean result
    if $b =$ "Yes" **then** ▷ Object is still in $\mathcal{F}^{(i)}$
      append *object* to list^1
    end if
  end for
  # check $\mathcal{O}$ in both $\mathcal{F}$ and $\mathcal{T}$
  for each *object* in list^1 **do**
    $p \leftarrow$ "Does this image contain a/an *object* ?"
    $b \leftarrow$ VQA($\mathcal{T}^{(i)}, p$)) ▷ Boolean result
    if $b =$ "Yes" **then** ▷ Object is in $\mathcal{F}^{(i)}$ & $\mathcal{T}^{(i)}$
      append *object* to list^2
    end if
  end for
  score $\leftarrow$ length of list^2 / length of list^1
  $s \leftarrow s + score$
end for
$s \leftarrow s \, / \, |\mathcal{T}|$
return s

Algorithm 6 Retention Evaluation by OVD

Input:
$\mathcal{F}$: samples of NeIn
$\mathcal{T}_o$: original caption from MS-COCO
$\mathcal{T}$: considered model's outputs
Output:
s: retention score

$s := 0$
for each tuple $(\mathcal{F}^{(i)}, \mathcal{T}_o^{(i)}, \mathcal{T}^{(i)})$ in $(\mathcal{F}, \mathcal{T}_o, \mathcal{T})$ **do**
  list^1 :=[], list^2 :=[]
  $\mathcal{O} \leftarrow$ extractor($\mathcal{T}_o^{(i)}$) ▷ Original objects in $\mathcal{I}$
  $p^1 \leftarrow$ OVD($\mathcal{F}^{(i)}, \mathcal{O}$) ▷ Objects are still in $\mathcal{F}^{(i)}$
  for each *object* in p^1 **do**
    # each object in p^1 may overlap with multiple confidence scores; store each object only once
    append unique *object* to list^1
  end for
  $p^2 \leftarrow$ OVD($\mathcal{T}^{(i)}$, list^1) ▷ Objects in $\mathcal{F}^{(i)}$ & $\mathcal{T}^{(i)}$
  for each *object* in p^2 **do**
    append unique *object* to list^2
  end for
  score $\leftarrow$ length of list^2 / length of list^1
  $s \leftarrow s + score$
end for
$s \leftarrow s \, / \, |\mathcal{T}|$
return s

Quantitative Results

Methods	Image Quality						Negation Understanding				
							LLaVA-NeXT		OWLv2		
	L1 ↓	L2 ↓	CLIP-I ↑	DINO ↑	FID ↓	LPIPS ↓	Removal ↑	Retention ↑	Removal ↑	AUC-Removal ↑	Retention ↑
InstructPix2Pix [1]	11.24	3.59	81.68	73.53	10.60	0.43	3.83	81.96	6.70	50.11	81.63
InstructPix2Pix	**8.32**	**2.32**	**93.11**	**91.67**	**4.08**	**0.33**	**93.62**	**98.26**	**92.66**	**97.89**	**95.83**
MagicBrush [28]	8.95	2.69	88.29	84.91	7.80	0.36	5.06	93.86	8.13	52.48	91.39
MagicBrush	**8.38**	**2.35**	**93.04**	**91.53**	**4.15**	**0.33**	**92.18**	**98.21**	**91.24**	**97.34**	**98.07**
ZONE [10]	11.95	3.67	74.12	63.18	14.95	0.46	2.93	72.38	6.47	46.04	69.07
HQ-Edit [8]	23.48	9.61	62.84	46.60	27.61	0.67	32.23	54.75	40.42	70.29	57.43
InstructDiffusion [4]	8.54	2.54	90.57	88.62	6.89	0.34	31.46	97.55	30.00	67.99	97.58

Quantitative results of five image editing SOTA methods on the **evaluation set** of NeIn. All the metrics are in (%). The InstructPix2Pix (2^{nd} row) and MagicBrush (4^{th} row) fine-tuned on NeIn's training set are highlighted.

Qualitative Results

Qualitative results of five methods on **NeIn's evaluation samples** (first two samples) and **random image-prompt pairs** (last two samples). The **fine-tuned** InstructPix2Pix (3rd column) and MagicBrush (5th column) on NeIn's training set are highlighted.

Conclusion

- We introduce **NeIn**, the **first** large-scale dataset for negation understanding for image editing task. Additionally, we present a comprehensive **evaluation protocol**, including *removal* and *retention* aspects, to assess the performance of current image editing models on negation understanding.

- **Limitations**: (1) we have only performed experiments using image editing models, and (2) the negative predefined prompts are relatively simple.

- **Future Directions**: (1) fine-tuning and benchmarking NeIn for *other tasks* in vision-language domain; (2) considering *complex negative sentences* involving words such as “except”, “neither-nor”, etc.

Paper

Project Page