Target-Free Text-guided Image Manipulation

Wan-Cyuan Fan, Cheng-Fu Yang, Chiao-An Yang, Yu-Chiang Frank Wang

STRUCT Group Seminar
Presenter: Yexiang Cheng
2023.01.29

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

BACKGROUND: Text-guided image manipulation

Object-centric image editing

- Modify visual attributes of particular objects in the image, or change its style to match the given description.
- Related works
  - ControlGAN
  - ManiGAN
  - TediGAN

BACKGROUND: Text-guided image manipulation

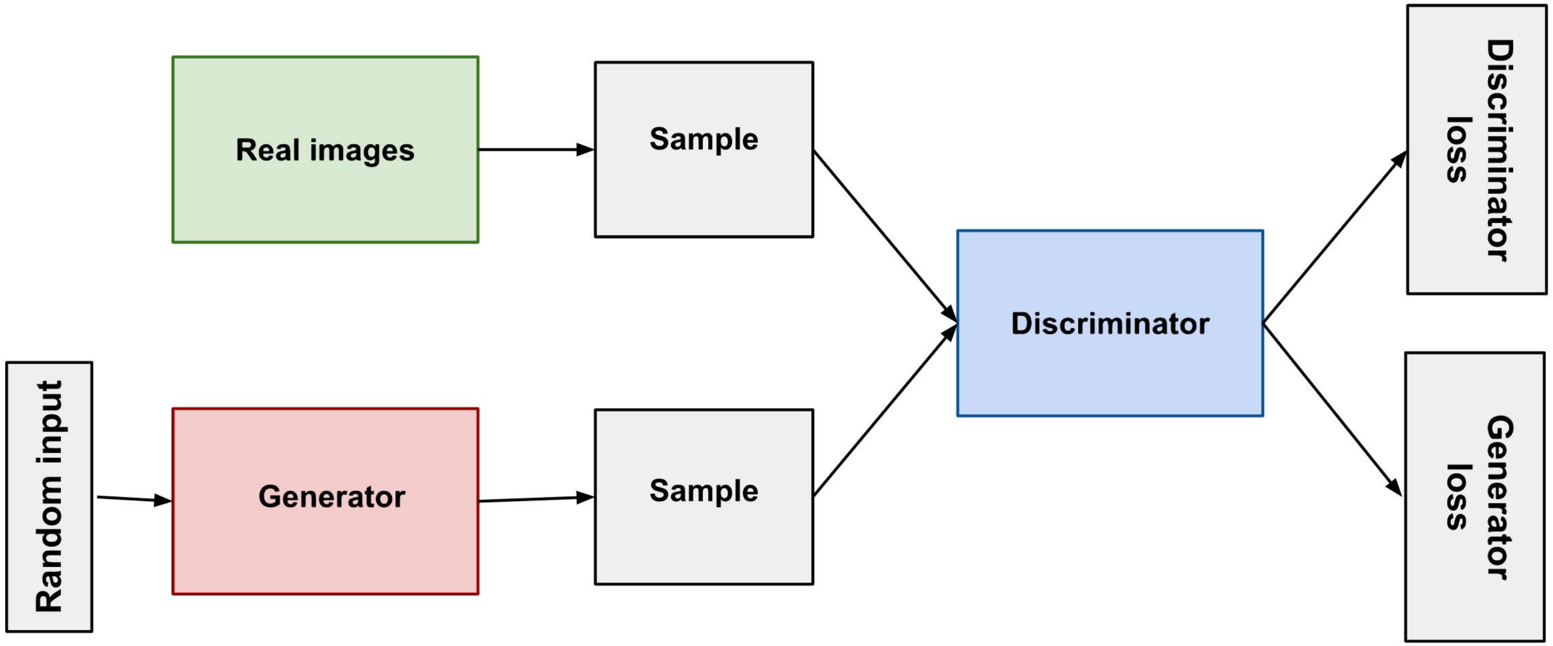
Scene-level image manipulation

- Reorganize the composition of input image based on the given instruction.
- Related works
  - GeNeVa
  - TIM-GAN
  - ASE

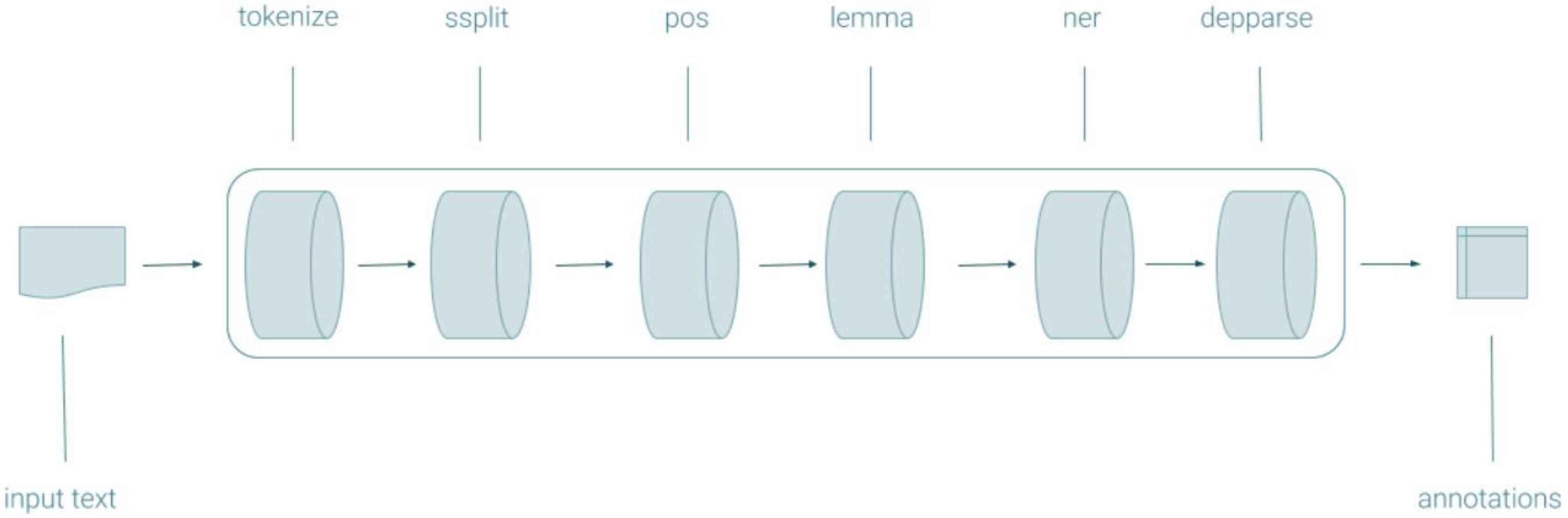
BACKGROUND: Text-guided image manipulation

Methods	Input data				Manipulation type		
	Instruction	Description	GT image	Auxiliary info	Change visual attribute	Remove object	Add object
ManiGAN (Li et al. 2020a)	-	✓	No	-	✓	-	-
TediGAN (Xia et al. 2021a)	-	✓	No	-	✓	-	-
ASE (Shetty, Fritz, and Schiele 2018)	-	-	No	Image-level labels	-	✓	-
GeNeVa (El-Nouby et al. 2019)	✓	-	Yes	-	-	-	✓
TIM-GAN (Zhang et al. 2021)	✓	-	Yes	-	✓	✓	✓
Ours	✓	-	No	Image-level labels	✓	✓	✓

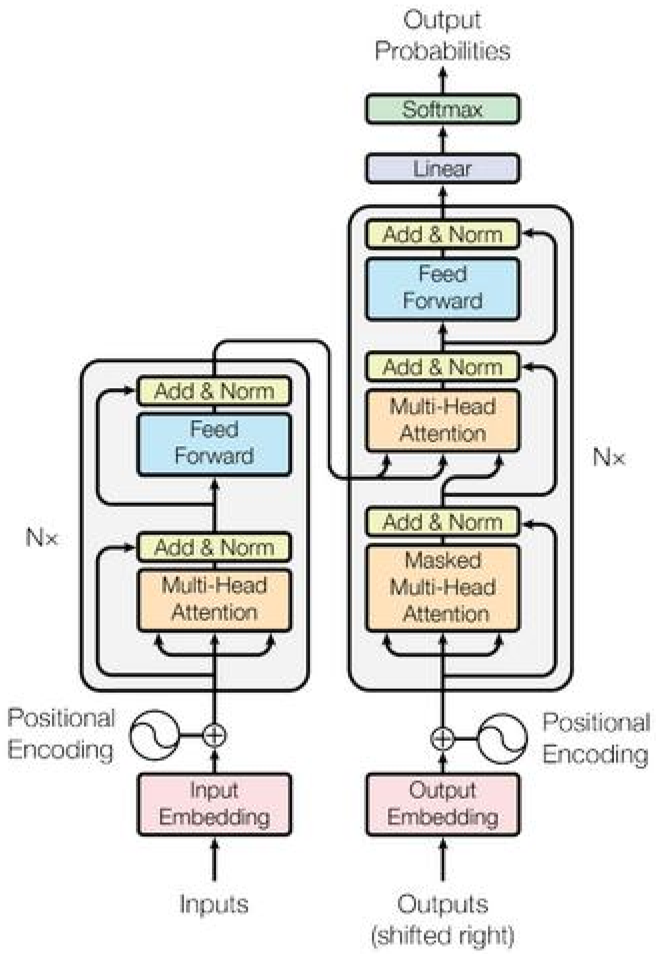
BACKGROUND: GAN



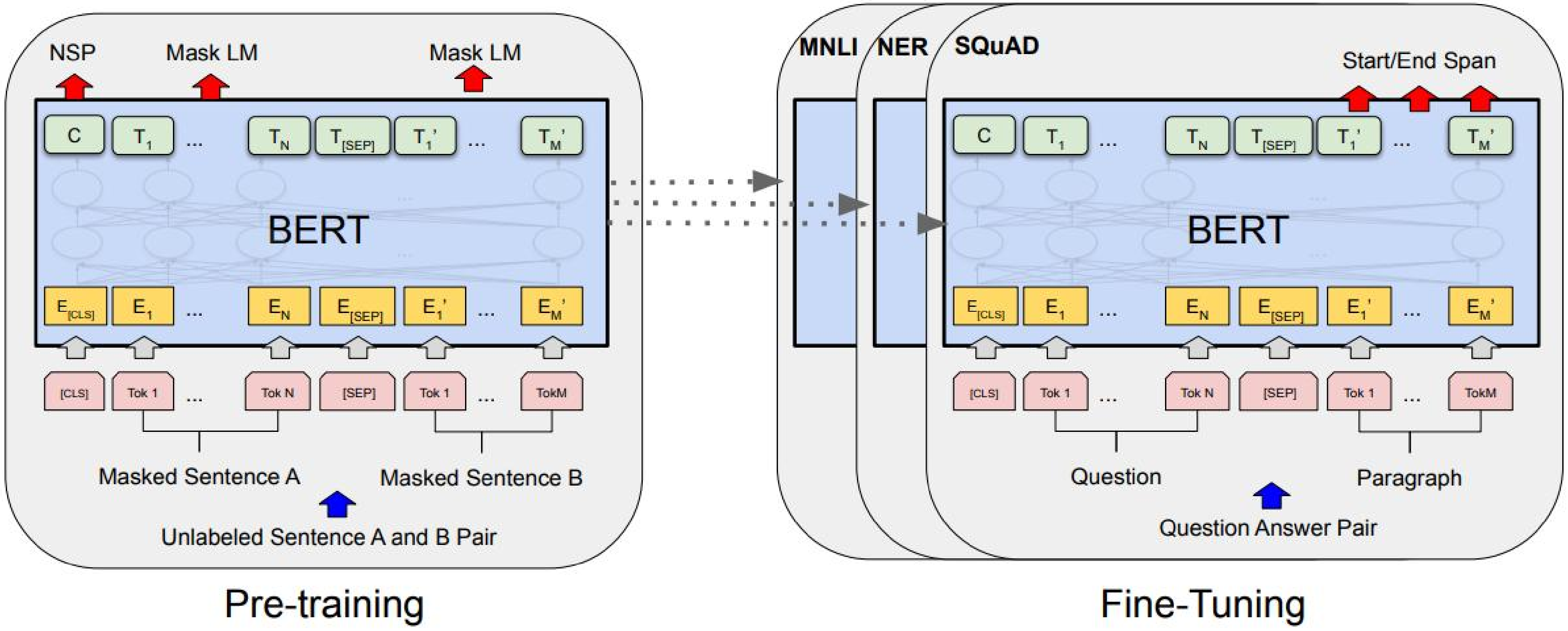
BACKGROUND: CoreNLP



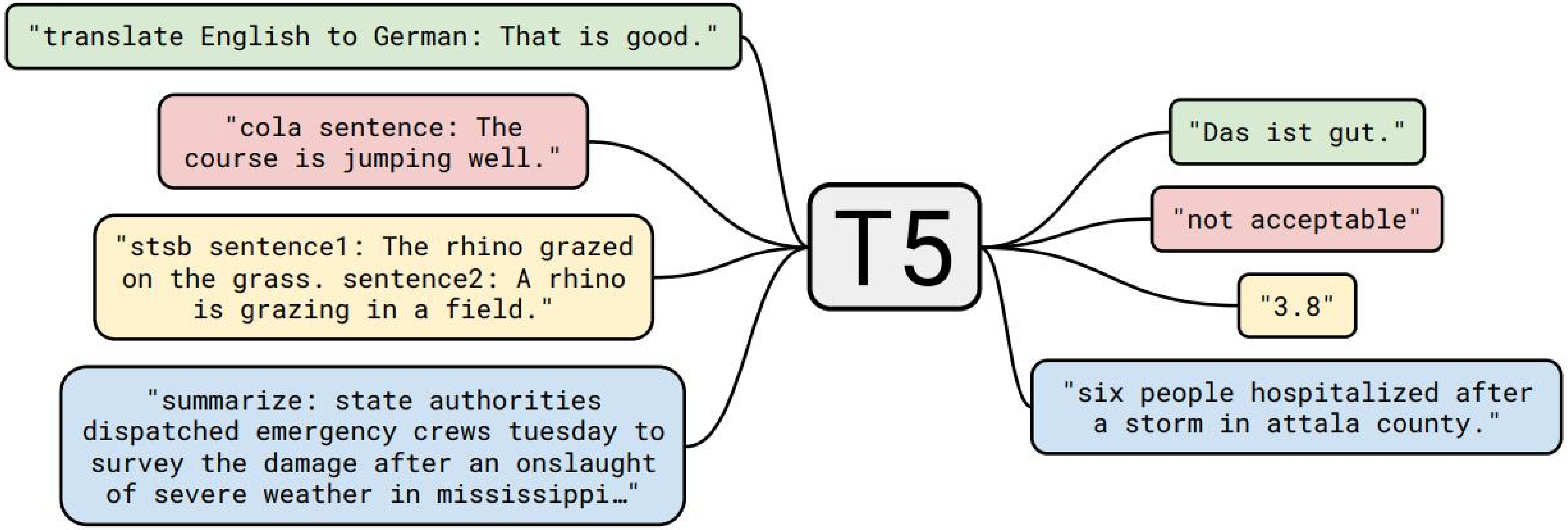
BACKGROUND: Transformer



BACKGROUND: BERT



BACKGROUND: T5

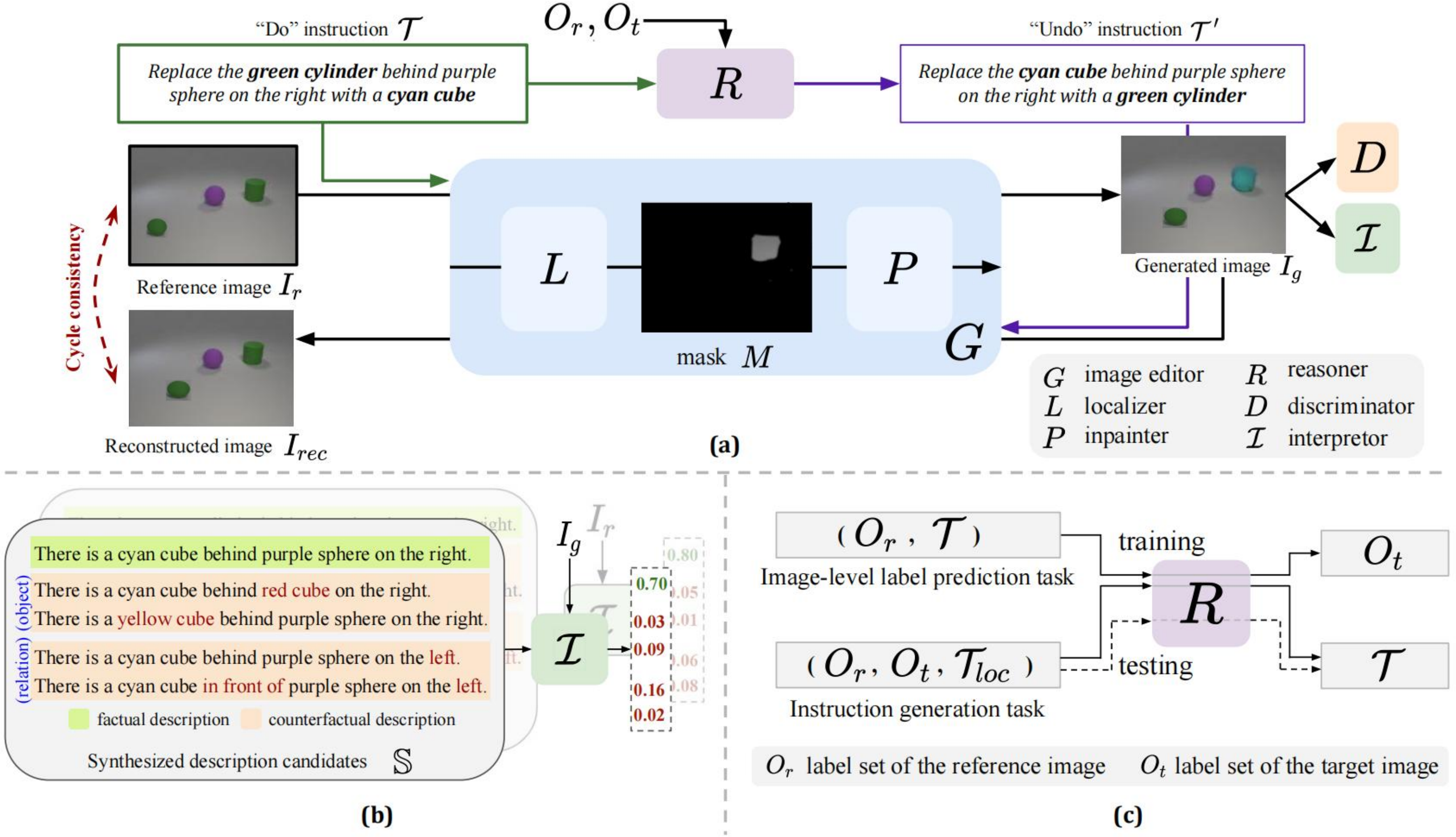


OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

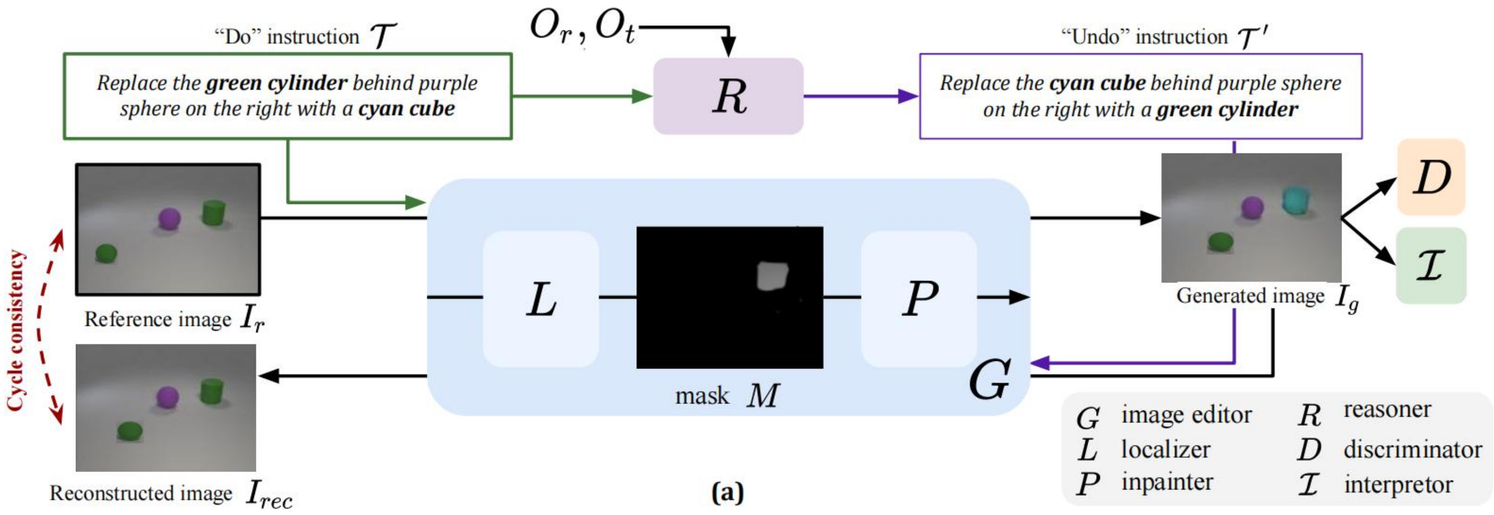
METHOD

Overview



Method

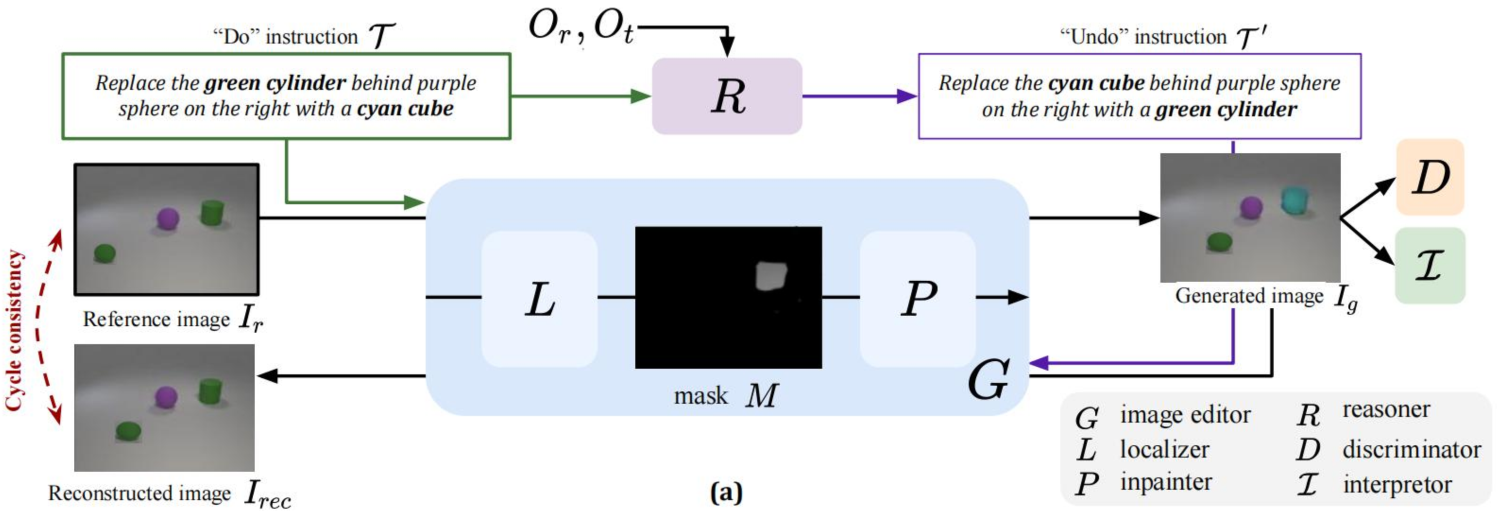
Localizer



- Identify the target object/location in I_r based on the adverb $\mathcal{T}_{loc}$ extracted from instruction $\mathcal{T}$ via CoreNLP.
- Achieved by performing cross modal attention between $f_{loc}^{\mathcal{T}}$(embedding of the location of interest encoded by a pre-trained BERT) and the feature map of I_r, followed by a mask decoder to produce M.

Method

Localizer

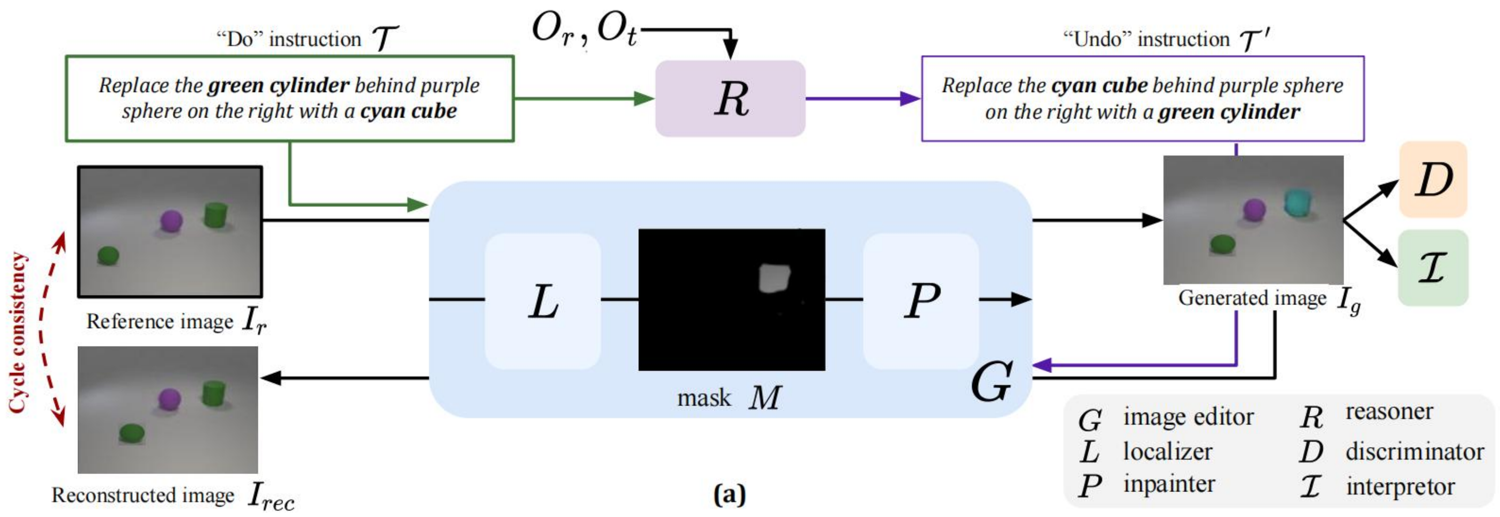


(a)

- Objective
  - $\mathcal{L}_{in}^{L} = \mathcal{L}_{CE}\big(MLP\big(E(M \cdot I_r)\big), y_{in}^{r}\big)$
  - $\mathcal{L}_{out}^{L} = \mathcal{L}_{BCE}\left(MLP\left(E((1 - M) \cdot I_r)\right), y_{out}\right)$

Method

Image In-painter



- Given I_r, M, $f_{how}^{\mathcal{T}}$(extracted from $\mathcal{T}$ by pre-trained BERT), produce I_g.

Method

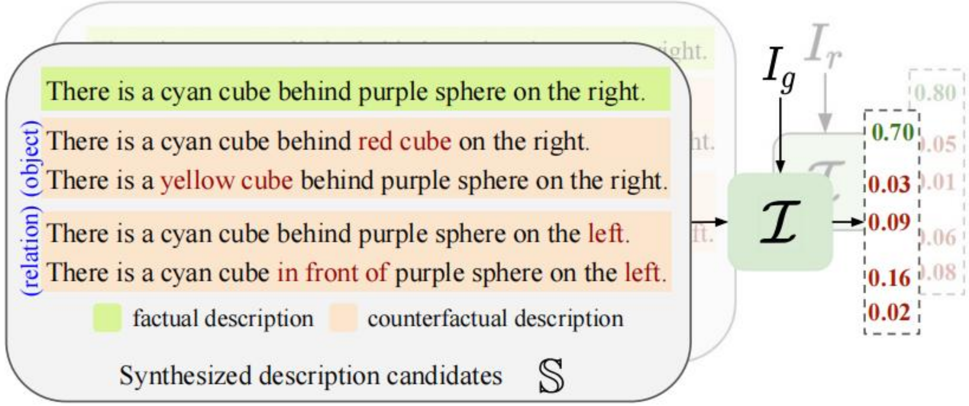
Image In-painter

- Objective
  - Adversarial loss
  - $\mathcal{L}_{rec}^{P} = \mathcal{L}_{MSE}\big((1 - M) \cdot I_r, (1 - M) \cdot I_g\big)$
  - $\mathcal{L}_{out}^{P} = \mathcal{L}_{BCE}\big(\mathcal{C}\big((1 - M) \cdot I_g\big), y_{out}\big)$
  - $\mathcal{L}_{in}^{P} = \mathcal{L}_{CE}\big(\mathcal{C}\big(M \cdot I_g\big), y_{in}\big)$

Method

Cross-Modal Interpreter

- Authenticates the output image via factual/counterfactual descriptions.
- Learning from Factual/counterfactual Descriptions:
  - Description template: There is a [OBJ][LOC]
  - OBJ: the symmetry difference between reference image label set O_r and target image label set O_t.
  - LOC: Adverb of the place of $\mathcal{T}$, extracted by CoreNLP.
- Authenticating Semantic Correctness of I_g.

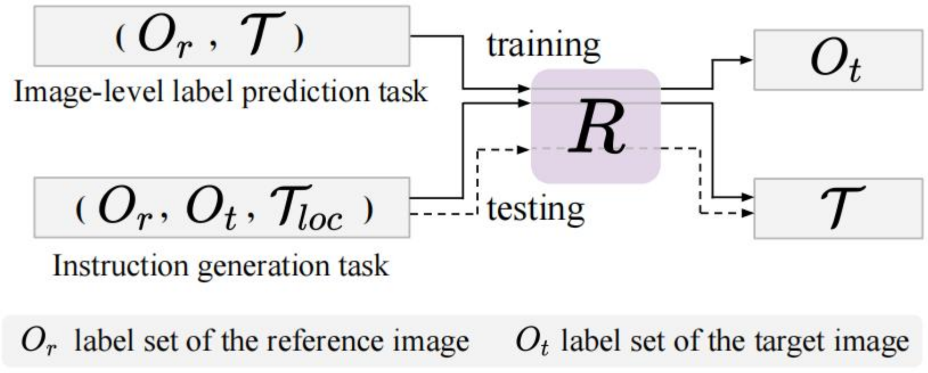


(b)

Method

Reasoner

- Produce the undo instruction for cross-modal cycle consistency.
- Purpose: minimizing the difference between I_r and I_{rec}.
- Two learning tasks:
  - O_r, $\mathcal{T}$ to O_t
  - O_r, O_t, $\mathcal{T}_{loc}$ to undo instruction $\mathcal{T}'$
- Objective:
  - $\mathcal{L}_R = \mathcal{L}_{s2s}\left(R(\mathcal{T}_r^O \oplus \mathcal{T}_t^O \oplus \mathcal{T}_{loc}), \mathcal{T}\right) + \mathcal{L}_{s2s}\left(R(\mathcal{T}_r^O \oplus \mathcal{T}), \mathcal{T}_t^O\right)$



OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

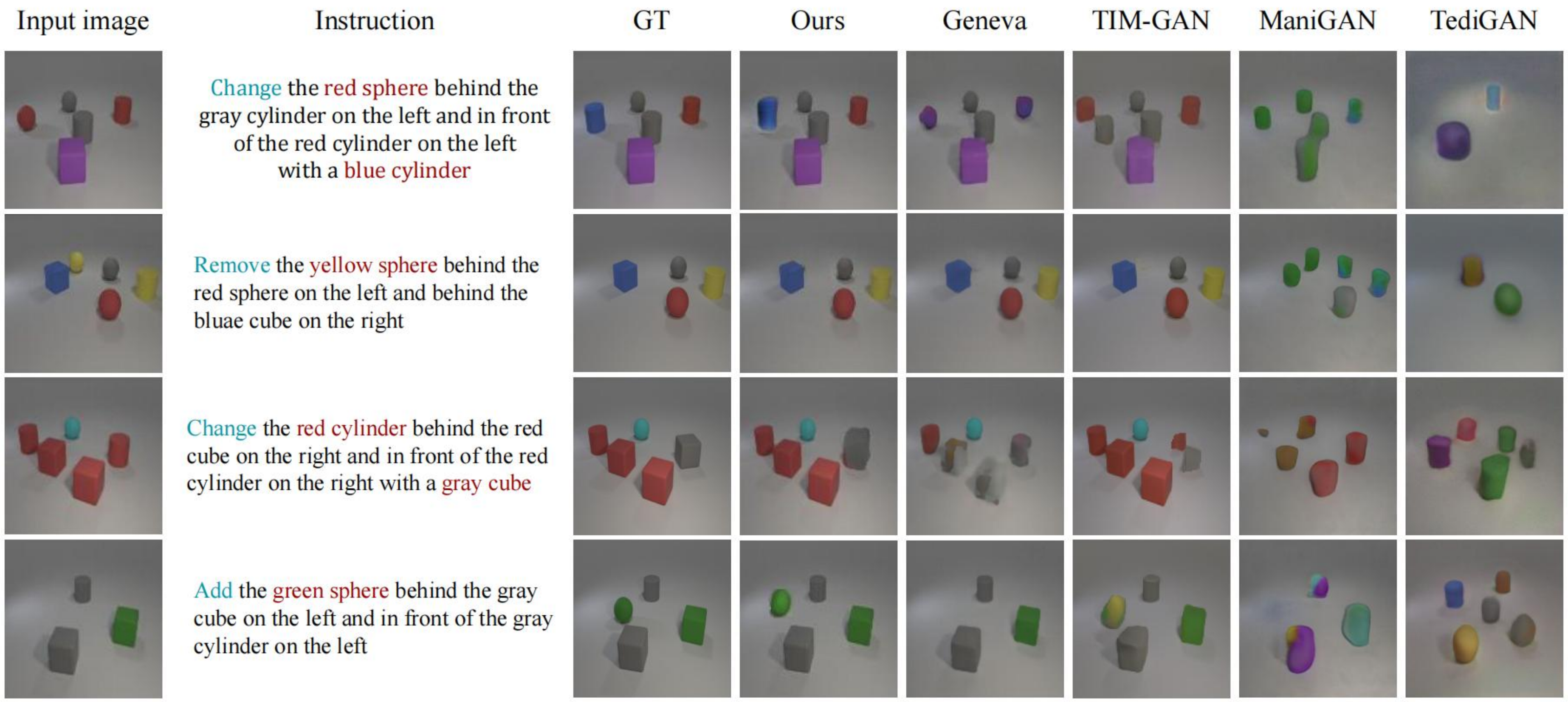
EXPERIMENTS

Datasets

- CLEVR: Created for multimodal learning tasks such as visual question answering, cross-modal retrieval, and iterative story generation. Synthesized version of CLEVR were considered(24 object categories, 28.1K/4.6K paired images with instructions)
- COCO: 118K real-world scene images. The subset with 20 object categories(overlapped with Pascal-VOC) were used.

EXPERIMENTS

Qualitative Evaluation on CLEVR



EXPERIMENTS

Qualitative Evaluation on CLEVR

Operation	Type 1: remove + add							Type 2: attribute change / shape						
Matrics	FID ↓	IS ↑	image acc (%)	In-mask acc (%)	Interp. acc (%)	R@1	R@5	FID ↓	IS ↑	image acc (%)	In-mask acc (%)	Interp. acc (%)	R@1	R@5
Upper bound	-	-	99.25	88.66	67.16	72.12	99.77	-	-	98.71	90.91	67.19	96.27	99.85
GeNeVa†	54.80	2.336	92.93	40.08	34.27	33.32	79.23	52.91	2.017	88.65	7.18	11.18	64.17	76.75
TIM-GAN†	**43.38**	2.192	93.40	25.50	38.17	33.72	80.81	54.66	2.122	90.05	4.67	10.79	58.73	76.37
ManiGAN	168.5	2.390	75.68	20.12	0.88	0.01	0.09	170.1	2.234	73.78	2.3	0.42	0.08	0.17
TediGAN	172.2	**2.760**	69.60	26.07	4.02	0.01	0.49	168.1	**2.672**	69.47	2.46	0.76	0.04	0.64
Ours	45.88	2.214	**93.59**	**43.01**	**40.85**	**47.95**	**94.04**	**38.26**	2.210	**93.18**	**39.18**	**33.74**	**87.46**	**94.01**

- Image acc: whether the objects in the generated image match the labels of the target image.
- In-mask acc: whether the generated object in the masked part can be recognized by a pretrained classification model.
- In-terp. Acc: whether the generated image semantically matches its factual description via a cross-modal interpreter.
- RS: the manipulation correctness of the manipulation by applying the existing text-guided image retrieval method of TIRG.

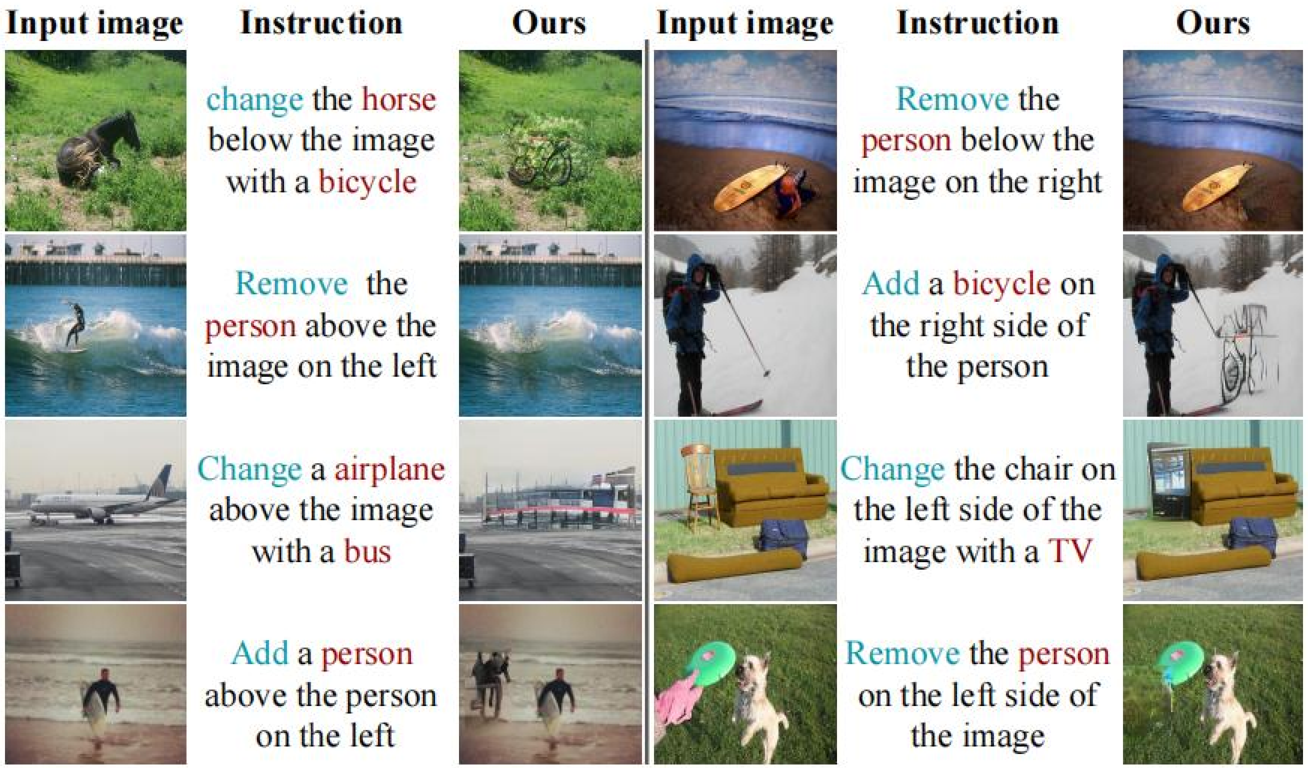
EXPERIMENTS

Qualitative Evaluation on COCO

	FID ↓	IS ↑	image acc (%)	Inside-mask acc (%)	Inpterp. acc (%)
Upper bound	-	-	91.47	92.49	68.71
Ours	166.18	4.64	86.04	17.17	13.54
ASE†	132.04	6.37	86.99	41.66	33.34
Ours†	**104.77**	**7.21**	**89.73**	**50.03**	**46.20**

EXPERIMENTS

Qualitative Evaluation on COCO



EXPERIMENTS

Ablation Studies

	FID ↓	IS ↑	image acc (%)	Inside-mask acc (%)	Interp. acc (%)
Upper bound	-	-	98.96	89.56	67.16
Ours w/o L	228.7	1.11	72.52	24.38	0.667
Ours w/o R (cycle)	68.08	2.07	83.47	41.40	28.27
Ours w/o $\mathcal{I}$	44.08	2.11	**93.73**	41.01	35.14
Ours w/o R, $\mathcal{I}$	77.56	2.08	80.22	39.70	26.55
Ours	**39.41**	**2.22**	93.41	**41.92**	**37.46**

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

CONCLUSION

- A Cyclic Manipulation GAN (cManiGAN) for target-free text-guided image manipulation.
- Using localizer and in-painter to decide “where” and “how” to edit given image.
- Using cross-modal interpreter to enforces the authenticity and correctness of the output image.
- Using reasoner to provide additional pixel-level guidance.

Thanks for listening!