Distractors-Immune Representation Learning with Cross-modal Contrastive Regularization for Change Captioning

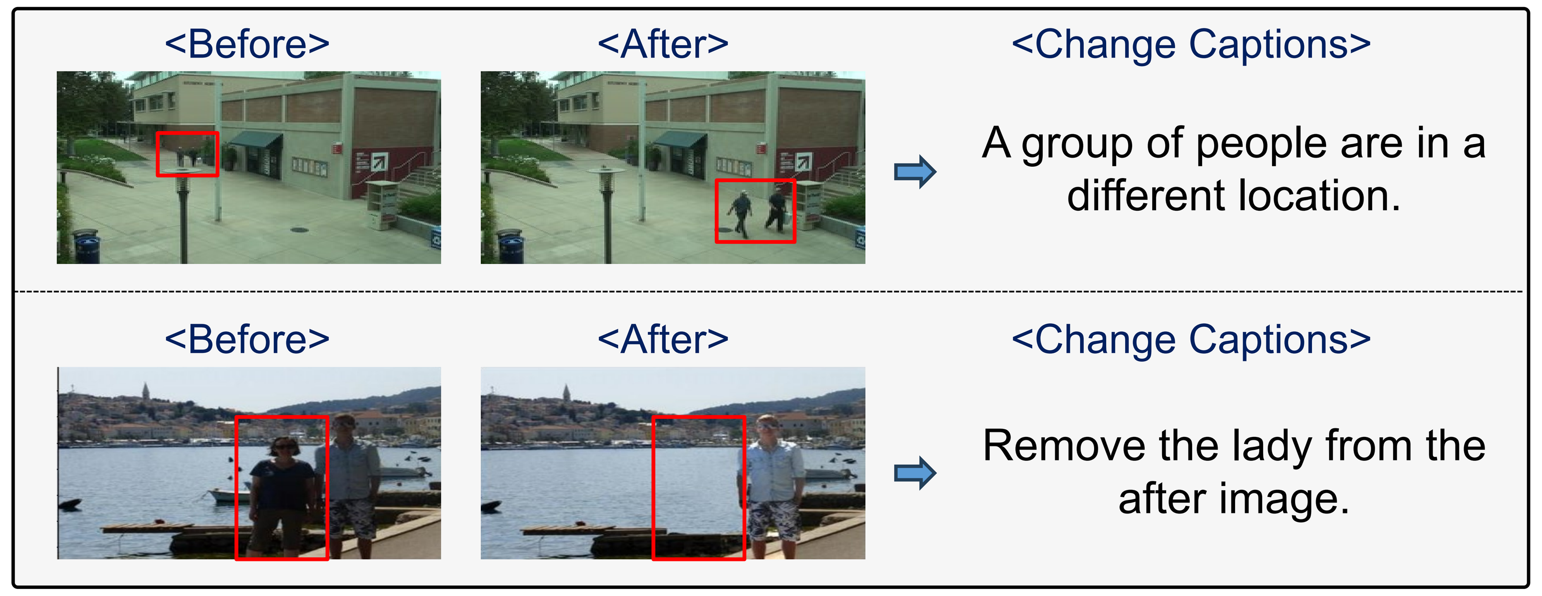
Yunbin Tu[1], Liang Li[2], Li Su[1], Chenggang Yan[3] and Qingming Huang[1]

[1]University of Chinese Academy of Sciences, [2]Institute of Computing Technology, CAS

[3]Hangzhou Dianzi University

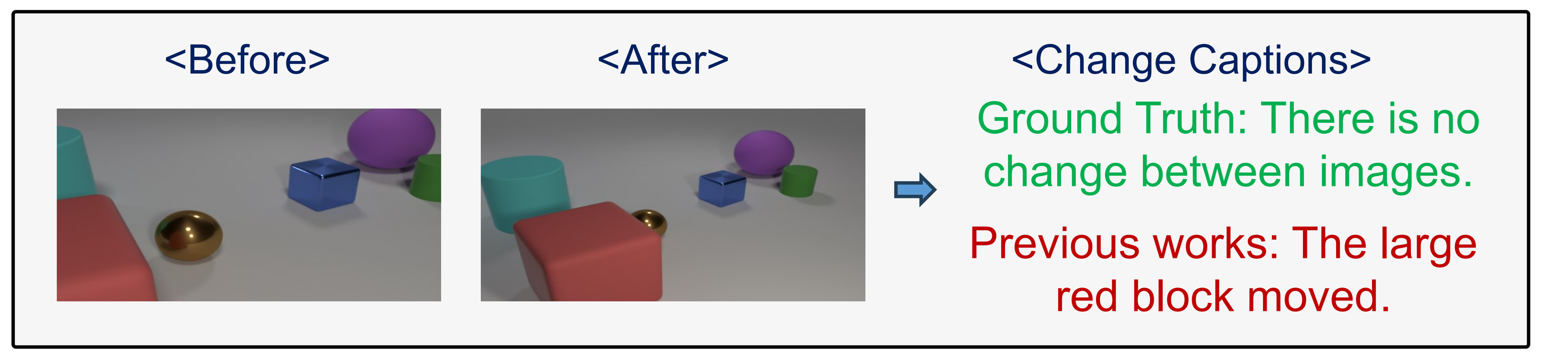
Problem Definition and Contribution

Goal: Change captioning is to describe the semantic change, while being immune to distractors (viewpoint / illumination changes) within an image pair in natural language.



Motivations:

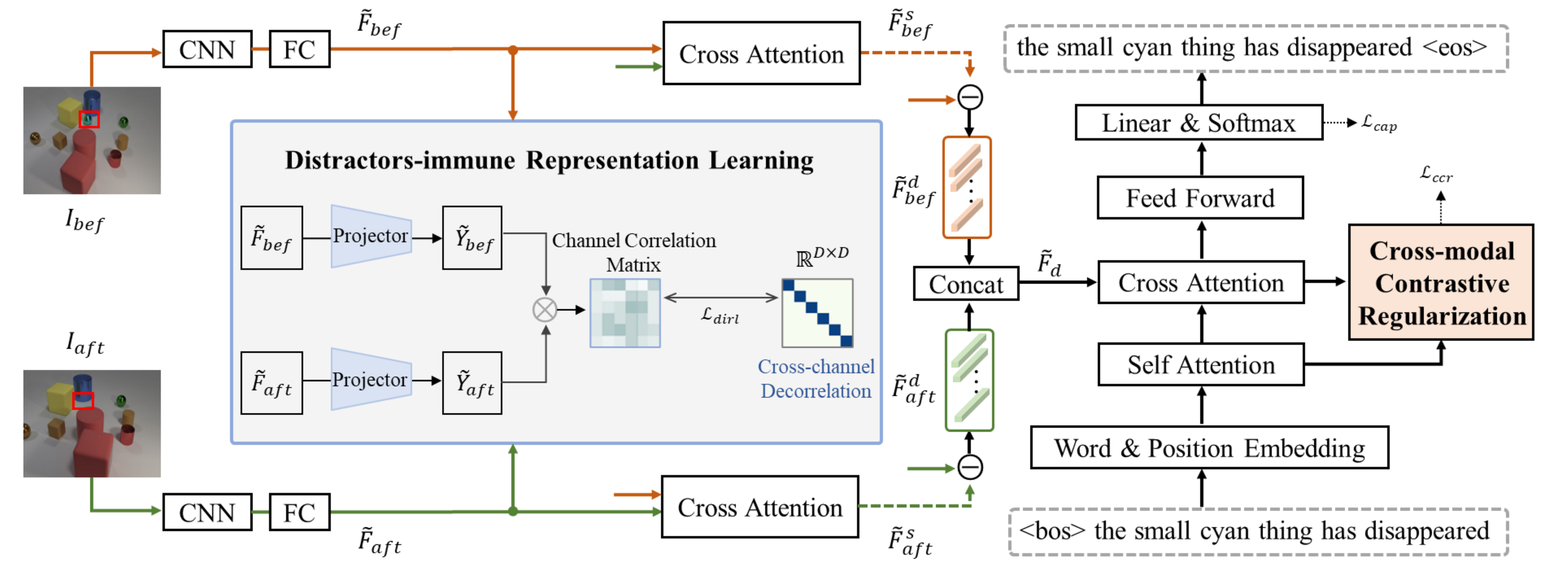
- Most unchanged objects appear **pseudo changes** and partially **overlap others**: features might be **perturbational** and **discrimination-degraded** under distractors.
- Previous works **directly subtract or match** between two unstable image features, yielding incorrect sentences.



Contributions:

- Distractors-Immune Representation Learning (DIRL) **captures two distractors-immune image features**, so the model can learn the robust difference features.
- Cross-modal Contrastive Regularization (CCR) **regularizes cross-modal alignment**, helping the decoder generate words based on the most related difference features.

Methodology



DIRL: Correlating corresponding channels and decorrelating different channels between two image features.

- Computing a channel correlation matrix:

$$\tilde{Y}_o = \mathrm{MLP}(\tilde{F}_o), o \in (bef, aft), \quad C_{ij} = \frac{\sum_b \tilde{y}_{b,i}^{bef} \tilde{y}_{b,j}^{aft}}{\sqrt{\sum_b (\tilde{y}_{b,i}^{bef})^2}\sqrt{\sum_b (\tilde{y}_{b,j}^{aft})^2}} \in \mathbb{R}^{D\times D}$$

- Enforcing matrix to be identity matrix by $\mathcal{L}_2$ -norm minimization:

$$\mathcal{L}_{dirl} = \sum_i (1 - C_{ii})^2 + \alpha \sum_i \sum_{j\neq i} C_{ij}^2$$

- Matching two updated features to gain unchanged features:

$$\tilde{F}_{bef}^s = \mathrm{CA}(\tilde{F}_{bef}, \tilde{F}_{aft}, \tilde{F}_{aft}), \qquad \tilde{F}_{aft}^s = \mathrm{CA}(\tilde{F}_{aft}, \tilde{F}_{bef}, \tilde{F}_{bef})$$

- Removing both from two images to learn difference features:

$$\tilde{F}_{bef}^d = \tilde{F}_{bef} - \tilde{F}_{bef}^s, \quad \tilde{F}_{aft}^d = \tilde{F}_{aft} - \tilde{F}_{aft}^s, \quad \tilde{F}_d = \mathrm{ReLU}([\tilde{F}_{bef}^d; \tilde{F}_{aft}^d]W_c + b_c)$$

CCR: Maximizing the contrastive alignment between the features of attended difference and generated words.

- Computing the global representation for word embeddings and attended difference features from the transformer decoder:

$$\tilde{E}[W] = \mathrm{Avg}\,(\mathrm{SA}(\hat{E}[W], \hat{E}[W], \hat{E}[W])), \qquad \tilde{V} = \mathrm{Avg}\,(\mathrm{CA}(\tilde{E}[W], \tilde{F}_d, \tilde{F}_d))$$

- Enforcing the contrastive alignment between $\tilde{E}[W]$ and $\tilde{V}$:

$$\mathcal{L}_{ccr} = \mathrm{InforNCE}(\mathrm{sim}(\tilde{E}[W], \tilde{V}))$$

  - (sim: dot-product operation)

Experimental Results

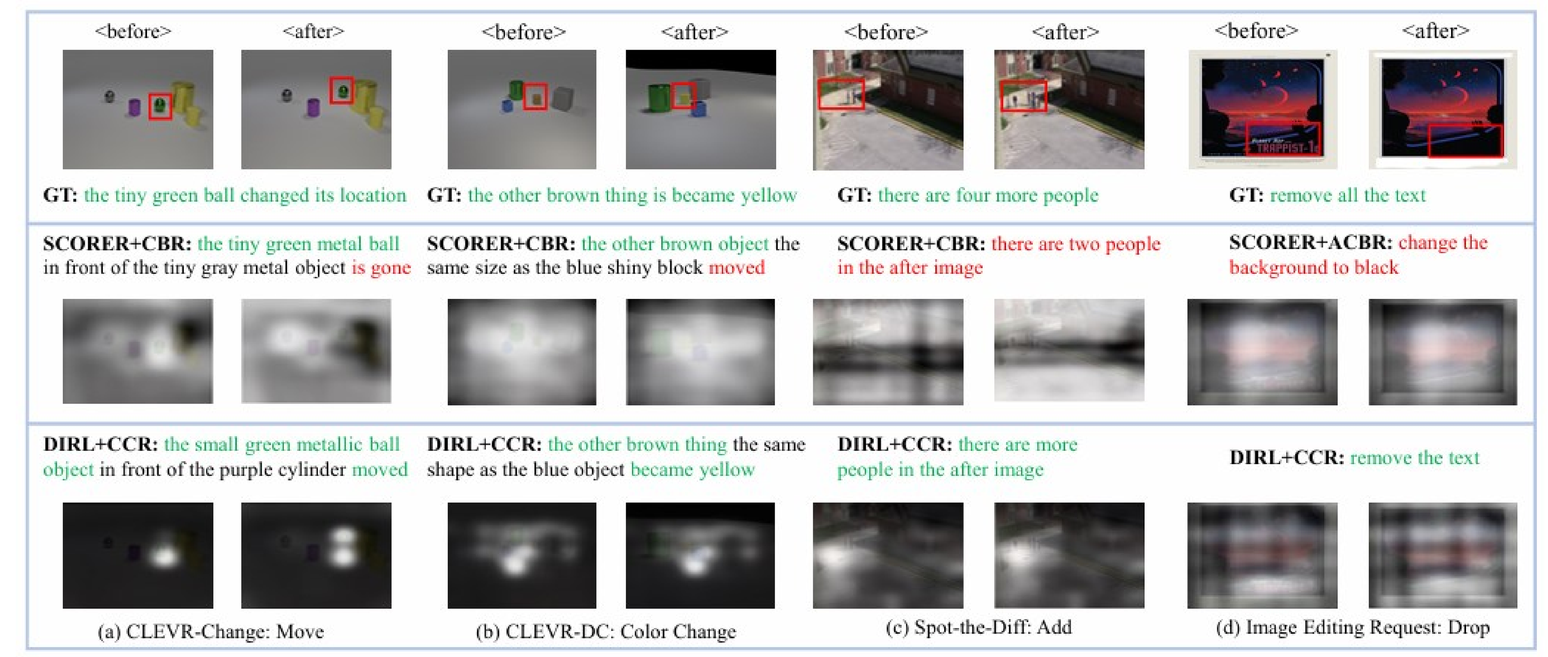
Comparison with existing methods on CLEVR-Change:

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DUDA [24] (ICCV 2019)	42.9	29.7	-	94.6	19.9
DUDA+TIRG [9] (CVPR 2021)	49.9	34.3	65.4	101.3	27.9
MCCFormers-D [25] (CVPR 2021)	53.3	37.1	70.8	119.1	30.4
R^{3}Net+SSP [37] (EMNLP 2021)	52.7	36.2	69.8	116.6	30.3
IFDC [11] (TMM 2022)	47.2	29.3	63.7	105.4	-
I3N [47] (TMM 2023)	53.1	37.0	70.8	117.0	32.1
NCT [33] (TMM 2023)	53.1	36.5	70.7	118.4	30.9
VARD-Trans [31] (TIP 2023)	53.6	36.7	71.0	119.1	30.5
SCORER+CBR [35] (ICCV 2023)	54.4	37.6	71.7	122.4	31.6
SMART [34] (TPAMI 2024)	54.3	37.4	71.8	**123.6**	**32.0**
DIRL+CCR (Ours)	**54.6**	**38.1**	**71.9**	**123.6**	31.8

Ablation study on CLEVR-DC:

Ablation	DIRL	CCR	BLEU-4	ROUGE-L	CIDEr	SPICE
Transformer	×	×	48.9	65.6	79.6	15.7
Transformer	✓	×	50.5	65.8	81.8	16.2
Transformer	×	✓	49.3	65.5	82.7	16.4
Transformer	✓	✓	**51.4**	**66.3**	**84.1**	**16.8**

Visualization for change localization and caption:



- SCORER+CBR [ICCV 2023]

- More experimental results are shown in our paper.
- Code is available at: https://github.com/tuyunbin/DIRL.