Preserve or Modify? Context-Aware Evaluation for Balancing Preservation and Modification in Text-Guided Image Editing

KAIST

MLILAB Machine Learning & Intelligence

Yoonjeon Kim[1], Soohyun Ryu*, Yeonsung Jeong, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang[2], Eunho Yang[1,3]*

[1]KAIST [2]Korea Naval Academy [3]AITRICS (* Equal contribution)

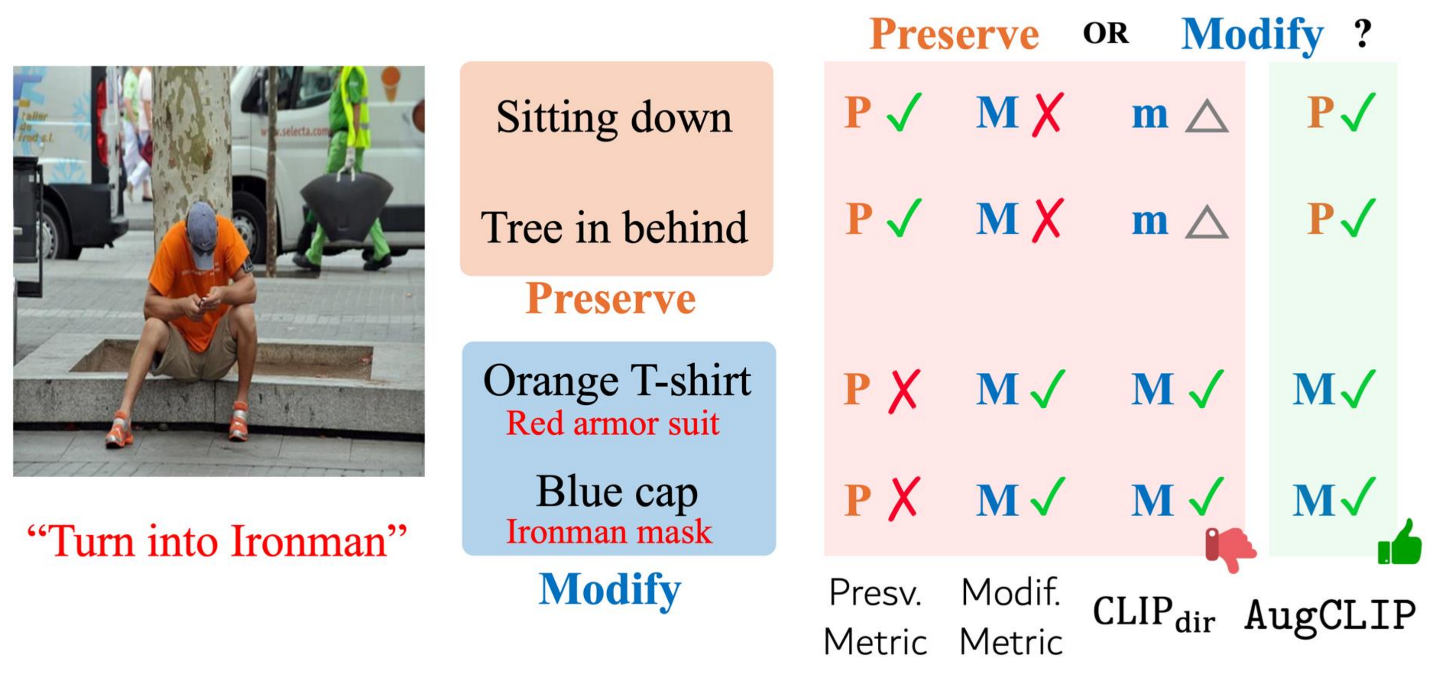
CVPR Nashville JUNE 11-15, 2025

Project Page Link · Paper Link · Code Link

Motivation & Problem Statement

Evaluation Metric for Text-guided Image Editing

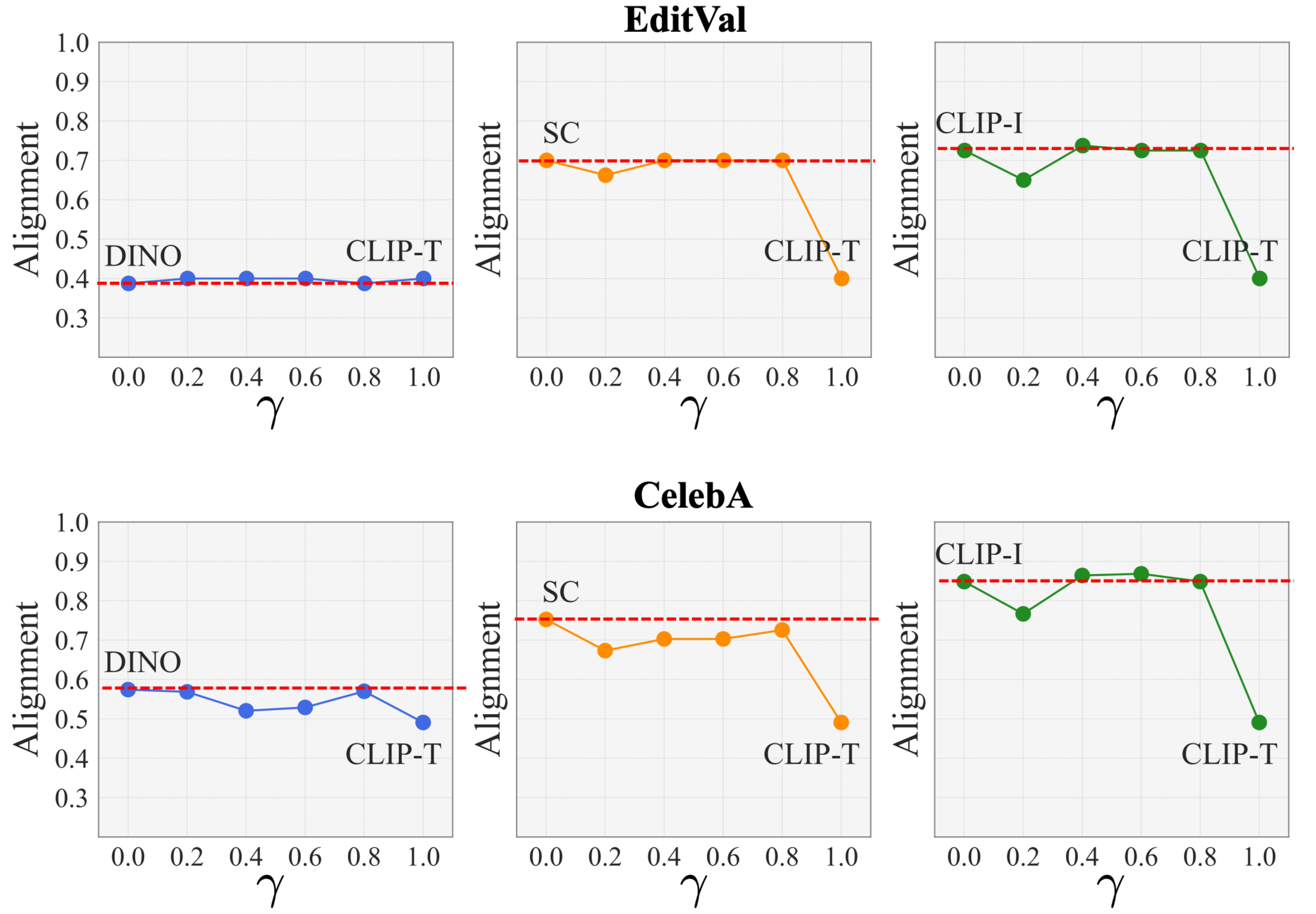
- Evaluation on text-guided editing has been relying on either preservation or modification-centric metrics.
- No single reliable metric capable of evaluating both aspects exists.



(a) Given Context (b) Guideline (c) Evaluation Standard of Metrics

Combination of Existing Metrics

- Interpolating preservation and modification centric metrics do not improve alignment with human preferences (orange line).

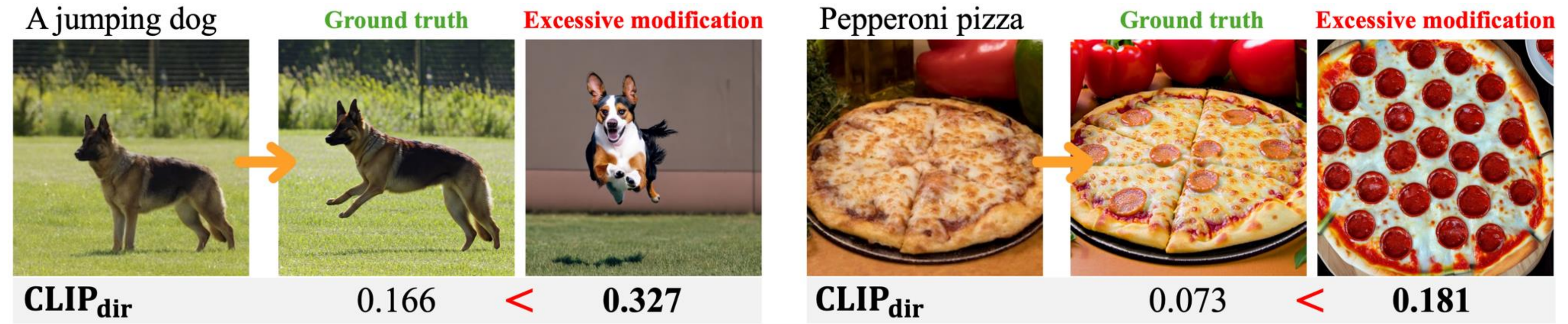


Directional CLIP Similarity Prefer Excessive Modification

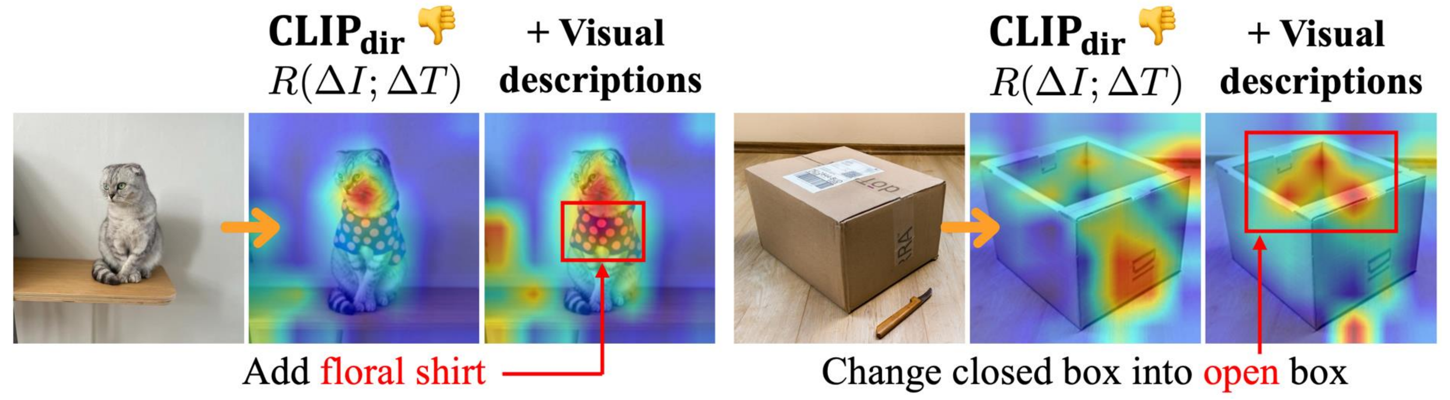
- Directional CLIP Similarity,

$$\mathrm{cs}\Big(E(I_{\mathrm{edit}}) - E(I_{\mathrm{src}}), E(T_{\mathrm{trg}}) - E(T_{\mathrm{src}})\Big)$$

blindly prefer excessively modified images.

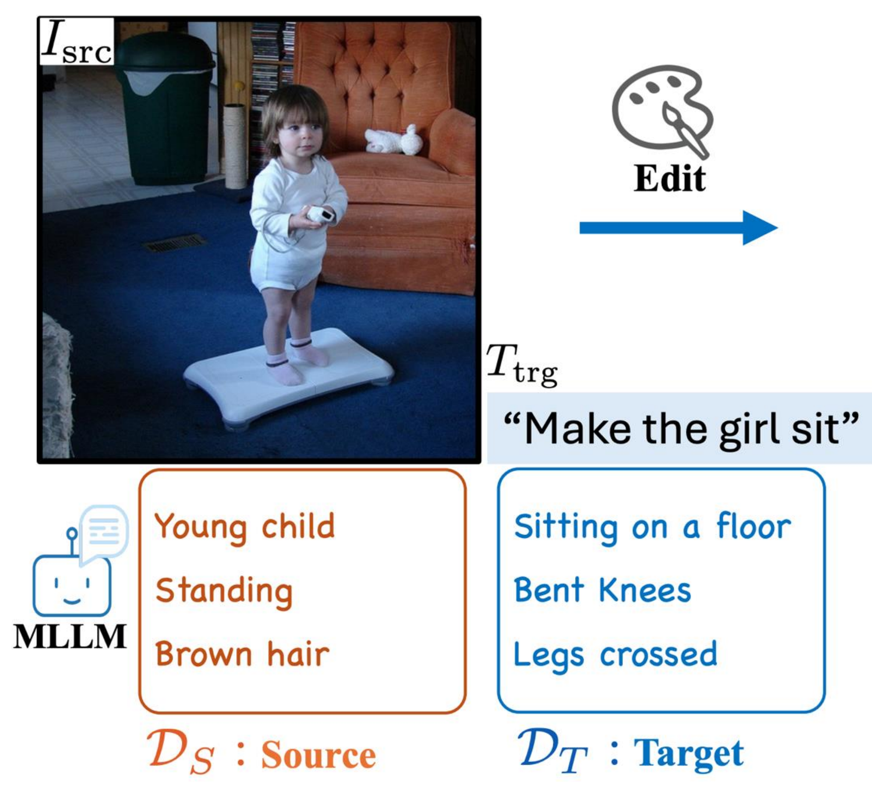


- Directional CLIP Similarity fail to pinpoint edited regions on the image, focusing on edit-irrelevant regions.

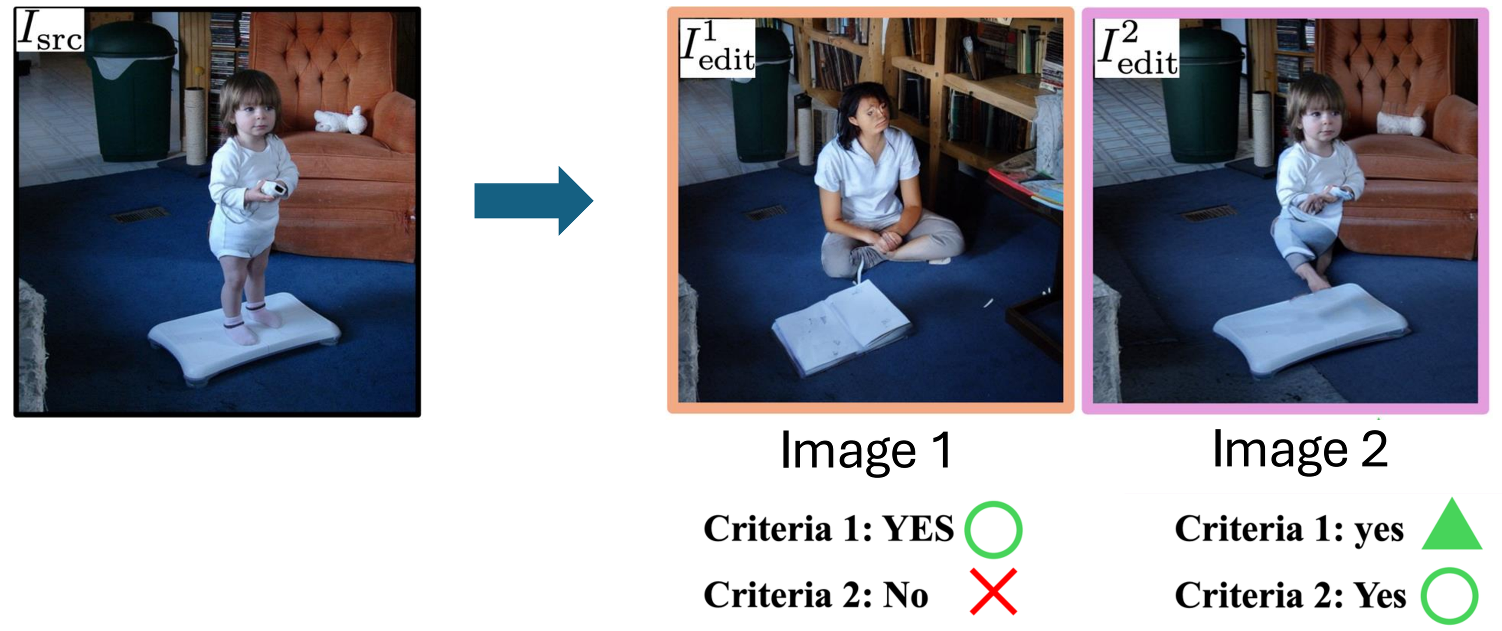


Proposed Method – AugCLIP

AugCLIP – Overview



- Attribute extraction from given source image and target text.
- Source attributes: CLIP representation of texts that describe the source image.
- Target attributes: CLIP representation of the texts that describe the target text.



- AugCLIP is a cosine similarity measured between the edited image and the ideal edited image, represented by adding a modification vector to the CLIP representation of source image: $\mathrm{CLIP}(I_{\mathrm{src}}) + v$

$$\texttt{AugCLIP} := \mathrm{cs}\left(\mathrm{CLIP}\left(I_{\mathrm{edit}}\right), \mathrm{CLIP}(I_{\mathrm{src}}) + v\right)$$

- How to define the "ideal edited image" in CLIP?

Criteria 1: Modified into target?

Criteria 2: Preserve source?

Minimum modification satisfies both criteria 1 & 2

- Step 1: Find a classifier in CLIP space that determines if an edited image is classified as target text or source image.

$$g(x) \;=\; \mathbf{w}^T x + b$$

- Step 2: Fit the classifier to CLIP representations of source and target attributes.
- Step 3: Derive the CLIP representation of ideal image that satisfies

$$\min_{\mathbf{v}} \|\mathbf{v}\| \quad \text{subject to} \quad \mathbf{w}^T \left(\boxed{E\left(I_{\mathrm{src}}\right) + \mathbf{v}}\right) + b > 0.$$

- Derivation of ideal modification vector v:

$$\mathbf{v} = c_{\min}\mathbf{w} = \frac{-(\mathbf{w}^\top I_{\mathrm{src}} + b)}{\|\mathbf{w}\|^2}\mathbf{w}$$

Experiments

Superior Alignment with Human Preferences

- The ranking of different editing models perfectly aligns with human evaluation result, where models are demonstrated in the order of rankings by human evaluation.

Dataset	Models	Rank	$\mathrm{CLIP_{dir}}$ ↑	LPIPS ↓	AugCLIP ↑	Human ↑
DreamBooth	ELITE	1	0.1132 (2)	71.38 (2)	0.7642 (1)	0.8478
	BlipDiffusion	2	0.0836 (3)	70.88 (1)	0.7579 (2)	0.6525
	CustomDiffusion	3	0.1348 (1)	73.84 (3)	0.6156 (3)	0.0263
EditVal	P2P	1	0.1771 (3)	15.04 (1)	0.8521 (1)	0.6133
	InstructPix2Pix	2	0.1774 (2)	25.75 (3)	0.8242 (2)	0.4855
	DiffEdit	3	0.2272 (1)	20.41 (2)	0.8155 (3)	0.3214
CelebA	StyleCLIP	1	0.0376 (2)	27.04 (1)	0.8484 (1)	0.6831
	Multi2One	2	0.0414 (1)	27.95 (2)	0.8152 (2)	0.5469
	Asyrp	3	-0.0001 (3)	36.98 (3)	0.7750 (3)	0.3197

- AugCLIP shows the highest pearson correlation with human preferences.

Metrics	Presv.	Modif.	CelebA	EditVal	DreamBooth
L2	✓	✗	0.653	0.348	0.464
LPIPS	✓	✗	0.465	0.360	0.286
DINO	✓	✗	0.574	0.348	0.286
SC	✓	✗	0.752	<u>0.764</u>	<u>0.571</u>
CLIP-I	✓	✗	<u>0.848</u>	0.730	**0.857**
CLIP-T	✗	✓	0.491	0.399	0.321
$\mathrm{CLIP_{dir}}$	▲	✓	0.673	0.697	0.357
AugCLIP	✓	✓	**0.883**	**0.831**	**0.857**

Evaluation Robustness to Various Editing Scenarios

	Pos. Add	Obj. repl.	Alter Parts	Background
$\mathrm{CLIP_{dir}}$	0.667	0.688	0.730	0.5
AugCLIP	**1.0**	**0.75**	**0.838**	**1.0**
	Texture	**Color**	**Action**	**Style**
$\mathrm{CLIP_{dir}}$	**0.806**	**1.0**	**1.0**	0.529
AugCLIP	0.742	**1.0**	**1.0**	**0.647**

- Alignment with human preferences excel directional CLIP similarity across various editing scenarios.