This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

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

[1] KAIST [2] Republic of Korea Naval Academy [3] AITRICS

https://augclip.github.io/

Abstract

The development of vision-language and generative models has significantly advanced text-guided image editing, which seeks the preservation of core elements in the source image while implementing modifications based on the target text. However, existing metrics have a ***context-blindness*** *problem, indiscriminately applying the same evaluation criteria on completely different pairs of source image and target text, biasing towards either modification or preservation. Directional CLIP similarity, the only metric that considers both source image and target text, is also biased towards modification aspects and attends to irrelevant editing regions of the image. We propose* `AugCLIP`*, a* ***context-aware*** *metric that adaptively coordinates preservation and modification aspects, depending on the specific context of a given source image and target text. This is done by deriving the CLIP representation of an ideally edited image, that preserves the source image with necessary modifications to align with target text. More specifically, using a multi-modal large language model,* `AugCLIP` *augments the textual descriptions of the source and target, then calculates a modification vector through a hyperplane that separates source and target attributes in CLIP space. Extensive experiments on five benchmark datasets, encompassing a diverse range of editing scenarios, show that* `AugCLIP` *aligns remarkably well with human evaluation standards, outperforming existing metrics. The code is available at* https://github.com/augclip/augclip_eval.

1. Introduction

Building on advancements in vision-language models [7, 19, 24], recent generative models [2, 8, 12] have been widely utilized as creative tools for image editing via text instructions. Text-guided image editing models enable the modification of images in response to textual guidance, ensuring that changes are aligned with the provided instructions. The primary objective of these models is to apply the necessary *modifications* guided by the target text while *preserving* most of the source image.

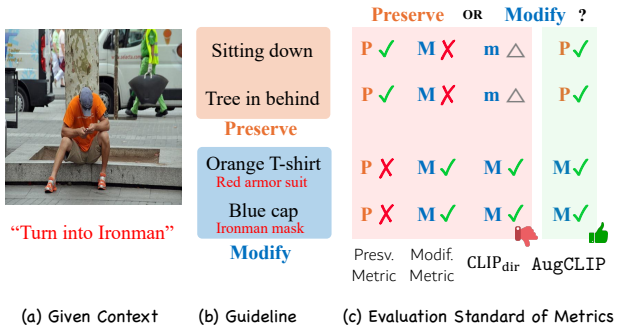


Figure 1. Context Blindness Problem of Existing Evaluation Metrics. Evaluation metrics should consider the specific context of a given source image and target text. However, existing metrics exhibit *context-blindness*, applying the same criteria of either 'preserve' (P) or 'modify' (M) across the entire image. Our proposed metric, `AugCLIP`, is a *context-aware* metric that flexibly applies different criteria to local regions of the image.

Despite the remarkable advancements in editing models, there has been a lack of rigorous evaluation methods, tailored specifically for text-guided image editing. Consequently, most studies [1, 2, 6, 9, 13, 15, 26] have heavily relied on human evaluation, which provides balanced consideration of preservation and modification aspects. However, as it is costly and impractical for real-world applications, researchers have adapted automatic evaluation metrics [3, 10, 11, 29] originally

*Equal contribution,† Corresponding Author.

designed for other vision tasks, such as image generation or captioning.

In text-guided editing scenarios, it is essential to assess which elements in the source to preserve and which in the target to modify into, based on the specific context of the given source image and target text (see Fig. 1(a)-(b) for example). Existing metrics, however, exhibit *context blindness*, meaning preservation-centric metrics completely ignore the target context while modification-centric metrics do not account for the source context. As shown in Fig. 1(c), preservation- and modification-centric metrics apply a fixed standard across the entire image, regardless of the true editing requirement of the given context. Even combining them fails to overcome the context blindness problem.

Among existing metrics, directional CLIP similarity [6] is unique because it provides a method to consider both the source image and target text in the evaluation. Even this metric suffers from a context-blindness problem, using a fixed standard without considering the specific relationship between the source image and target text. We show that directional CLIP similarity tends to prioritize modification over preservation, and often focuses on peripheral parts rather than regions that are edited as guided by target text, in Sec. 3. These observations underscore the need for a context-aware metric to dynamically adjust the evaluation standard, balancing both aspects in response to diverse editing contexts.

Based on our comprehensive analysis, we propose a novel context-aware metric, `AugCLIP`, which evaluates the quality of the edited image by comparing it with an estimated representation of a well-edited image. To balance preservation and modification, a well-edited image is represented by a minimum modification on the source image, while matching the target text. We use multi-modal large language models (MLLMs) to extract attributes that capture various visual aspects of the source image and target text. With these attributes, we derive a hyperplane that separates the source and target to derive the ideal modification as a projection to this hyperplane. `AugCLIP` evaluates how closely the edited image aligns with the estimated ideal image in CLIP space.

Our metric `AugCLIP` demonstrates remarkable improvement in alignment with human judgment on diverse editing scenarios such as object, attribute, and style alteration compared to all other existing metrics. Moreover, our metric is even applicable to personalized generation, the DreamBooth dataset, where the objective is to identify an object in the source image and edit it into a completely novel context. This shows the flexibility of `AugCLIP`, which seamlessly applies to a variety of editing directions. Notably, our metric excels in identifying minor differences between the source image and the edited image, showing superb ability in complex image editing scenarios such as MagicBrush.

The major contributions are summarized as follows.

- We point out the unreliability of current evaluation metrics in text-guided image editing, noting their inability to balance preservation and modification effectively due to inherent **context blindness**.
- We introduce `AugCLIP`, a **context-aware** metric for image editing by automatically **augmenting** descriptions via MLLM and estimating a balanced representation of preservation and modification, which takes into account the relative importance of each description.
- `AugCLIP` demonstrates a significantly high correlation with human evaluations across various editing scenarios, even in complex applications where existing metrics struggle.

2. Background on Existing Metrics

In the context of text-guided image editing, evaluation metrics should assess the quality of edited images in terms of both preservation of the source image and modification towards the target text.

Preservation-Centric Metrics FID [11], LPIPS [29], and L2 distance measure feature-level distance between images. These metrics are primarily used in image generation tasks to compare the distributions of real and generated images. When applied to editing scenarios, they assess the distance between the source and edited images, *completely ignoring the target text* guiding the editing process. Segmentation consistency (SC) [13] measures the structural change. Additionally, DINO similarity (DINO) and CLIP-I evaluate the semantic similarity between the source and edited images within the DINO [3] and CLIP [24] embedding spaces, also focusing solely on preservation.

Modification-Centric Metrics To assess how well an edited image aligns with the target text, the primary approach involves using a pre-trained multi-modal model [24]. CLIPScore (CLIP-T) [10] is the most widely adopted metric in this category, measuring the similarity between the edited image and the target text. This metric *completely ignores the source image* that should be considered for the preservation aspect.

Holistic Evaluation of Preservation and Modification Directional CLIP similarity (CLIP_{dir}) [6] is unique among existing metrics as it assesses both

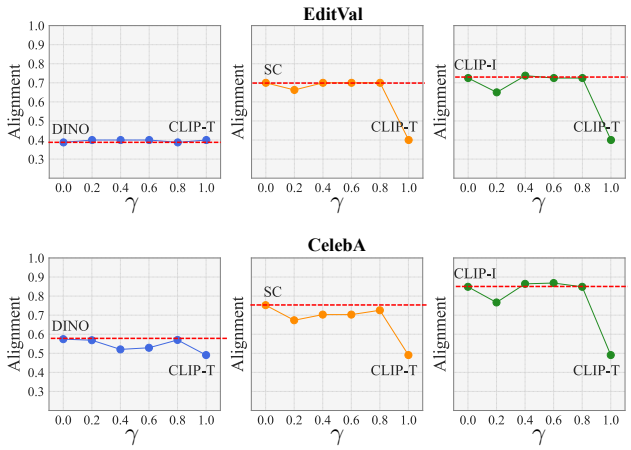


Figure 2. Combination of Preservation- and Modification-Centric Metrics Deteriorates in Performance. The plot shows the human alignment score s_{align} measured by a linear interpolation of {DINO, SC, CLIP-I} and CLIP-T. The results show that combining rather degrades the alignment with human judgment.

preservation and modification aspects by evaluating directional alignment between the original and edited images concerning the source and target text.

However, these metrics employ a fixed evaluation standard, disregarding the context of each editing scenario, which leads to *context blindness*. As discussed in Sec. 3, combining preservation- and modification-centric metrics or using directional CLIP similarity fails to balance preservation and modification adequately, resulting in misalignment with human judgment.

3. Problems of Existing Metrics in Balancing Preservation and Modification Aspects

Existing evaluation metrics, that focus on either one of the preservation or modification aspects, face a challenge in assessing both aspects. Does combining these two metrics successfully balance the preservation and modification aspects? We systematically analyze how these metrics and their combinations still fail in balancing the two colliding objectives, suffering from context-blindness (Sec. 3.1). Moreover, we focus on directional CLIP similarity, the unique metric that accounts for both source image and target text, to reveal that it presents two significant problems that question its reliability (Sec. 3.2).

3.1. Combination of Preservation- and Modification-Centric Metrics

We explore whether combining preservation- and modification-centric metrics can improve alignment with human judgments across two benchmark datasets: EditVal [1] and CelebA [20] (Results on the other three benchmark datasets are deferred to the Appendix). To do so, we combine CLIP-T, a modification-centric metric, and preservation-centric metrics (LPIPS, DINO similarity, CLIP-I) with varying interpolation values $\gamma \in [0, 1]$. Note that both metrics are scaled to the same range before interpolation, ensuring the interpolation value γ is properly reflected in the final combined score.

As illustrated in Fig. 2, the combination of two powerful metrics in preservation and modification often leads to performance degradation. This manifests that a simple combination of the colliding metrics cannot be a reliable evaluation metric. This is due to the context-blindness of these metrics, failing to holistically consider source image and target text. Preservation-centric metrics compare only the source and edited images, ignoring the target context. Similarly, modification-centric metrics consider only the target text, overlooking the source context. This highlights the need for a tailored metric that is contextually aware of both preservation and modification, ensuring a reliable alignment with human evaluations across diverse datasets and tasks.

3.2. Directional CLIP Similarity

Directional CLIP similarity (CLIP_{dir}) is designed to assess both preservation and modification in text-guided image editing by measuring the similarity between image and text directions as follows:

$$\text{CLIP}_{\text{dir}} = \texttt{cs}(\Delta I, \Delta T) \tag{1}$$

$$= \texttt{cs}\Big(E(I_{\text{edit}}) - E(I_{\text{src}}), E(T_{\text{trg}}) - E(T_{\text{src}})\Big), \tag{2}$$

where $\texttt{cs}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ denotes cosine similarity and $E(\cdot)$ is a CLIP encoder for either image or text. Here, I_{edit} is the edited image to be evaluated, which is edited from the source image I_{src} according to the target text T_{trg}. For evaluation of directional CLIP similarity, I_{src}, the corresponding text about the source image, is required. We analyze that CLIP_{dir} has two major problems that hinder the balanced evaluation of preservation and modification.

CLIP_{dir} overly emphasizes modification. Directional CLIP similarity overly emphasizes the alignment with the target text while neglecting the preservation

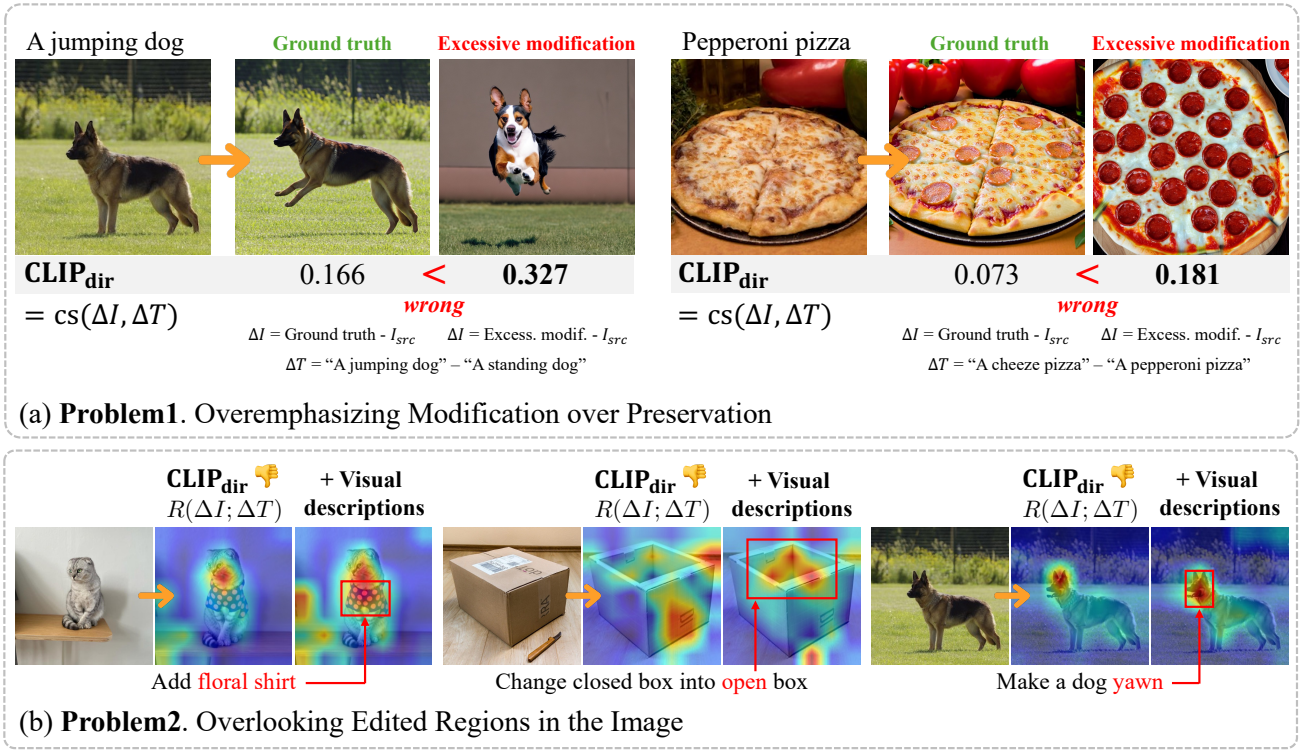


Figure 3. Problems of Directional CLIP Similarity. (a) CLIP_{dir} assigns higher scores to excessive modification, over well-edited ground truth images. (b) CLIP_{dir} evaluates edited images by attending to irrelevant regions of the image. Adding visual annotations helps CLIP_{dir} properly attend to edited regions.

of the source image. Despite attempting to incorporate preservation by subtracting T_{src} from T_{trg}, CLIP_{dir} frequently favors excessively modified images over ground truth ideally edited images as shown in samples in Fig. 3(a). Specifically, CLIP_{dir} metric fails on 63% of the 100 editing cases in TEdBench [12] and 35.1% of the 1053 cases in MagicBrush [28] (Details in Sec. 5). This problem happens since CLIP_{dir} is designed to assume that a well-edited image should primarily adhere to the target text. Thus, it blindly favors the visual features of the target text, leading to unreliable results.

CLIP_{dir} overlooks edited regions. CLIP_{dir} often focuses on peripheral or unmodified local areas of the edited image, overlooking the specific areas that are edited by the text. Using the relevancy map [4], $\boldsymbol{R}(\Delta I; \Delta T)$[1], to visualize attention on image regions corresponding to the text direction, we observe that CLIP_{dir} frequently fails to concentrate on the image regions that are directly relevant to the target text as shown in Fig. 3(b). The figure also shows that adding explicit visual details guides CLIP_{dir} to attend to edited regions properly.

[1]The original relevancy map $\boldsymbol{R}$ utilizes an image-text pair. To adapt it to CLIP_{dir}, we subtract two relevancy maps. Details are in the Appendix.

4. AugCLIP: A Context-Aware Evaluation Metric

We introduce a novel evaluation metric, `AugCLIP`, for scoring the editing quality of the image in terms of preservation and modification. Given a target text T_{trg} and a source image I_{src}, `AugCLIP` measures the cosine similarity between the edited image I_{edit} and the ideal editing. Since the ideal edited image is not given in most evaluation cases, we estimate the representation of an ideally edited image as a modification on the source image, $I_{\text{src}} + \mathbf{v}$, where $\mathbf{v}$ is the modification vector. Then, the editing quality of the edited image I_{edit} is evaluated as

$$\texttt{AugCLIP}\left(I_{\text{edit}}; I_{\text{src}}, T_{\text{trg}}\right) = \texttt{cs}\left(E(I_{\text{edit}}), E(I_{\text{src}}) + \mathbf{v}\right), \tag{3}$$

where $E(\cdot)$ is a CLIP encoder.

The evaluation score ranges between $[0, 1]$, where the maximum value 1 indicates that the edited image is ideally modified by an editing model. We outline the framework of `AugCLIP` in Fig. 4. We extract the visual

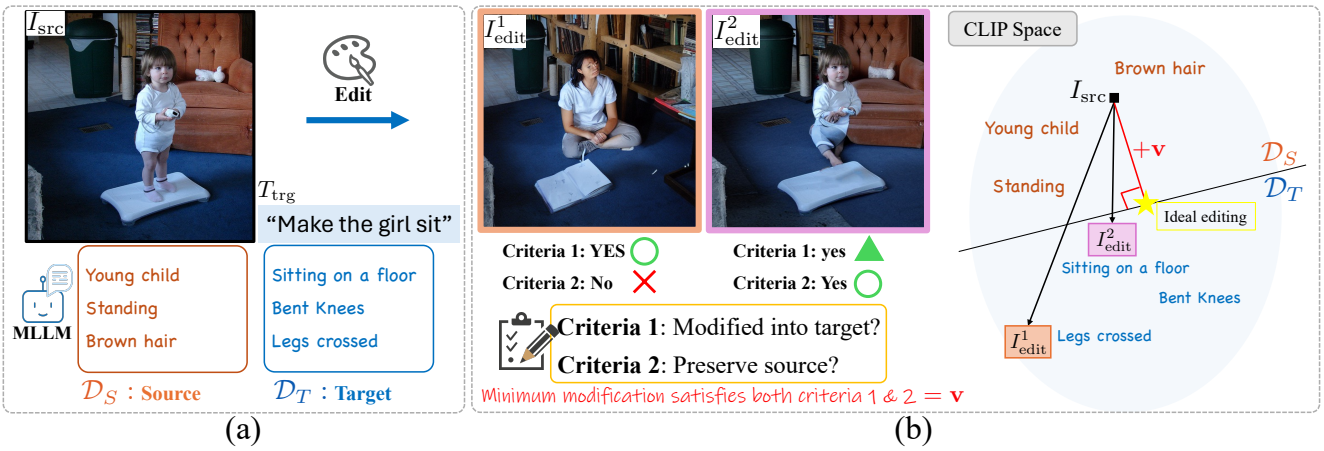


Figure 4. (a) Description Extraction Process. The source image describes a young child standing on the balance board. Target text guides the editing model to make the girl sit. The source and target attributes are extracted with MLLM. **(b) Evaluation Process of `AugCLIP`.** The two edited images demonstrate i) An older woman sitting down with legs crossed (I^1_{edit}) and ii) A young girl sitting on the floor (I^2_{edit}). `AugCLIP` derives the ideal image representation as a minimum modification $\mathbf{v}$ on the source image to be classified as *target*. The second image that is closer to $I_{\text{src}} + \mathbf{v}$ shows a higher score, while the first image that is excessively modified with lost source identity shows a lower score.

features of the source image and target text (Sec. 4.1), then derive the CLIP representation of an ideal modification that satisfies the two criteria: 'i) Does the edited image follow the target text?' and 'ii) Does the edited image preserve the source image?' (Sec. 4.2).

4.1. Extracting and Processing Attributes

Source & Target Attribute Extraction As described in Fig. 4(a), we extract visual descriptions about attributes of the source image and target text using a state-of-the-art multi-modal large language model (MLLM), GPT-4V [22]. From the source image, we extract visual features that can be the object of preservation. Likewise, from the target text, we extract visual features that can be the object of modification. The extraction process is instructed by five annotated in-context examples, which are provided in the Appendix.

Encoding into CLIP Space After attribute extraction from MLLM, each attribute is encoded into CLIP space. The source attributes are denoted as $\mathcal{D}_S$ and target attributes as $\mathcal{D}_T$.

Relative Weighting on Attributes In editing scenarios, some of the attributes are more important than others. For instance, drawing stripes is more essential than a short mane to represent a zebra. To prioritize this, we emphasize the similarity between key attributes within the same class. Since multi-modal large-language models sample based on the empirical probability, frequently extracted attributes indicate higher importance in representing the context. Formally, the importance of each visual attribute in $\mathcal{D}_S$ and $\mathcal{D}_T$ is represented as a cosine similarity between CLIP embeddings in the same set as

$$Importance(v_i) = \sum_{j \neq i} \frac{u_i \cdot v_j}{\|u_i\| \|v_j\|}, \tag{4}$$

for all u_j in the same attribute set, either $\mathcal{D}_S$ or $\mathcal{D}_T$.

Additionally, the importance should be down-weighted by how much the visual attribute hinders the preservation of the source image or modification of the target text. More specifically, a collision between source and target attributes is measured by averaged cosine distance to CLIP embeddings in the opposite set as

$$Collision(v_i) = -\sum_{l} \frac{v_i \cdot w_l}{\|v_i\| \|w_l\|}, \tag{5}$$

for all w_l in the opposite attribute set.

Adding these two measures, the weighting value for the attribute $v \in \mathcal{D}_S \cup \mathcal{D}_T$ is defined as

$$\alpha(v) = Importance(v) + Collision(v). \tag{6}$$

This weighting value accounts for the relative significance of each attribute in the evaluation process and is integrated into Eq. (7).

4.2. Deriving the Ideal Editing in CLIP Space

Based on the extracted source and target attributes, we describe how to derive a modification vector that satisfies i) modification into the target text and ii) preservation of the source image.

Criterion 1: Does the edited image follow the target text? To determine if the edited image faithfully follows the given target text, it is intuitive to employ a classifier that classifies the image into either *source* or *target*. Classification into *source* indicates the edited image is still closer to the source image, while classification into *target* means that the image is sufficiently modified. The classifier is derived by a hyperplane that separates the source $\mathcal{D}_S$ and target $\mathcal{D}_T$.

Formally, the classifier function $g(x) = \mathbf{w}^T x + b$ assigns a CLIP embedding x to the target class if $g(x) > 0$, or to the source class if $g(x) < 0$. Therefore, $E(I_{\text{edit}})$ should satisfy $\mathbf{w}^T E(I_{\text{edit}}) + b > 0$ if the edited image faithfully follows the target text. Here, the optimization form of the classifier function $g(x)$ is

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \alpha(v_i)\cdot\max(0, 1-y_i(\mathbf{w}\cdot v_i+b)), \quad (7)$$

where $y_i = +1$ for $v_i \in \mathcal{D}_T$ and $y_i = -1$ for $v_i \in \mathcal{D}_S$. Here, the precomputed weighting values $\alpha(\cdot)$ in Eq. (6) are integrated into the optimization process.

Criterion 2: Does the edited image preserve the source image? However, it is equally important to consider the second criterion, source image preservation. Therefore, the CLIP representation of modification should inflict minimum modification to the original embedding $E(I_{\text{src}})$, while being classified as a target class. Formally, we derive the modification $\mathbf{v}$ in CLIP space that satisfies

$$\min_{\mathbf{v}} \|\mathbf{v}\| \quad \text{subject to} \quad \mathbf{w}^T \left(E\left(I_{\text{src}}\right) + \mathbf{v}\right) + b > 0. \quad (8)$$

Derivation of $\mathbf{v}$ To minimize $\|\mathbf{v}\|$, $\mathbf{v}$ should be in the direction that most efficiently increases the classifier output $g(\cdot)$. This direction is given by the gradient of the classifier function concerning $\mathbf{x}$, which is $\mathbf{w} = \nabla g(\mathbf{x})$. Therefore, the optimal $\mathbf{v}$ is $c\mathbf{w}$ where $c > 0$ is a scalar to be determined. Substituting $\mathbf{v} = c\mathbf{w}$ to Eq. (8), we derive the following equation

$$c > \frac{-(\mathbf{w}^\top I_{\text{src}} + b)}{\|\mathbf{w}\|^2}. \quad (9)$$

In order to solve for c that minimizes $\|\mathbf{v}\|$, we derive the smallest c satisfying the condition as $c_{\min} = \frac{-(\mathbf{w}^\top I_{\text{src}}+b)}{\|\mathbf{w}\|^2}$. Then the modification vector $\mathbf{v}$ is derived as

$$\mathbf{v} = c_{\min}\mathbf{w} = \frac{-(\mathbf{w}^\top I_{\text{src}} + b)}{\|\mathbf{w}\|^2}\mathbf{w}. \quad (10)$$

Finally, `AugCLIP` is computed by substituting $\mathbf{v}$ to Eq. (3).

5. Experiments

Implementation Details. For our experiments, we employ a pre-trained CLIP-ViT 16/B model for CLIP-based metrics. Source and target attributes are generated using GPT-4V [22]. Further details on prompting the source and target attributes are in the Appendix.

Baseline Metrics. We compare `AugCLIP` with DINO similarity, LPIPS, L2, Segmentation Consistency (SC), Directional CLIP similarity (CLIP_{dir}), CLIP-I, and CLIP-T. The CLIP variants are summarized as follows:

$$\begin{aligned}
\text{CLIP}_{\text{dir}} &= \texttt{cs}\left(E(I_{\text{edit}}) - E(I_{\text{src}}), E(T_{\text{trg}}) - E(T_{\text{src}})\right)\\
\text{CLIP-I} &= \texttt{cs}\left(E(I_{\text{edit}}), E(I_{\text{src}})\right)\\
\text{CLIP-T} &= \texttt{cs}\left(E(I_{\text{edit}}), E(T_{\text{trg}})\right).
\end{aligned}$$

Note that FID measures the distance between two distributions of image features, this metric cannot be measured in a sample-wise manner. Therefore, FID cannot be tested in terms of alignment with human evaluation or preference test on ground truth samples.

Benchmark Datasets and Editing Models. We evaluate `AugCLIP` and existing metrics across several text-guided image editing benchmarks, including TEdBench [12], EditVal [1], MagicBrush [28], DreamBooth [26], and CelebA [20]. Each benchmark dataset represents varying editing scenarios as in Tab. 1. For each benchmark dataset, we employ multiple editing models [2, 5, 8, 9, 13, 14, 16–18, 23, 27] as specified in Tab. 1 for suitable type of benchmark dataset. Details on benchmark datasets are provided in the Appendix.

5.1. Correlation with Human Judgment

User Study To evaluate the effectiveness of `AugCLIP` compared to existing metrics, we conduct a Two-Alternative Forced Choice survey on human evaluators to collect preferences over diverse edited images. The alignment score $\boldsymbol{s}_{\text{2AFC}}$ measures if each metric prefers the same option as human evaluators. The survey was conducted using Amazon Mechanical Turk (AMT), over three benchmark datasets, encompassing a total of 105 participants as detailed in Tab. 2. The survey questionnaire is structured as described in the Appendix, providing clear criteria to consider source image preservation and target text alignment.

`AugCLIP` aligns better with human judgments. As shown in Tab. 3, `AugCLIP`, the context-aware metric that balances preservation and modification, is superior over all previous metrics across diverse benchmark datasets. A high level of alignment between

Table 1. Benchmark Datasets and Editing Models Used for Evaluating the Performance of Metrics.

	CelebA	EditVal	TEdBench	MagicBrush	DreamBooth
Dataset Types	Facial Attribute	General Object	Object Centric	Local Region Editing	Personalized Generation
Editing Models	[13, 14, 17, 23]	[2, 5, 8, 9, 12]	[2, 5, 8, 9, 12]	[2, 5, 8, 9, 12]	[16, 18, 27]
Dataset Size	50	648	100	1053	50

Table 2. User Study Statistics.

	CelebA	EditVal	DreamBooth
Survey questions	39	35	37
Participants	45	30	30

Table 3. **Correlation with Human Judgment.** Human alignment score s_{2AFC} (↑), scaled between 0 to 1.

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	0.764	0.571
CLIP-I	✓	✗	0.848	0.730	**0.857**
CLIP-T	✗	✓	0.491	0.399	0.321
CLIP$_{dir}$	▲	✓	0.673	0.697	0.357
`AugCLIP`	✓	✓	**0.883**	**0.831**	**0.857**

preservation-centric (L2, LPIPS, DINO, SC, CLIP-I) and human judgment is expected because these benchmarks do not significantly deviate from the source image.

5.2. Ground Truth Editing Selection Test

Creating Triplet Dataset We generate a triplet of images, each representing a well-edited, excessively preserved, and excessively modified image. Among the triplet of images, the well-edited image is provided in the benchmark datasets, TEdBench, and MagicBrush. Excessively modified images are generated with text-to-image models [25] instructed by the target text, completely ignoring the source image. Excessively preserved images are generated by applying a negligible transformation to the given source image, ignoring the target text. Then, we measure the accuracy where each metric correctly assigns the highest score to the well-edited image as **Acc**$_{\text{Both}}$. Yielding high accuracy in this test means that the evaluation metric can balance both preservation and modification aspects, without being biased to either side.

`AugCLIP` is superior in balancing preservation and modification. In Tab. 4, `AugCLIP` is on par with or even better than baseline metrics over all datasets. This emphasizes the balanced evaluation criteria of `AugCLIP`, which previous metrics fail to abide by. Segment Consistency (SC) excels in MagicBrush since the dataset requires editing a small local area of the image.

Table 4. **Ground Truth Selection Test.** Accuracy, **Acc**$_{\text{both}}$ (↑), of assigning higher scores to ground truth images over excessively preserved and modified images.

Metrics	Presv.	Modif.	TEdBench	MagicBrush
L2	✓	✗	0.310	0.002
LPIPS	✓	✗	0.090	0.000
DINO	✓	✗	0.280	0.008
SC	✓	✗	0.420	**0.936**
CLIP-I	✓	✗	0.281	0.810
CLIP-T	✗	✓	0.312	0.260
CLIP$_{dir}$	▲	✓	0.350	0.601
`AugCLIP`	✓	✓	**0.570**	0.889

6. Analysis

We use s_{2AFC} for CelebA, EditVal, and Dreambooth, and **Acc**$_{\text{Both}}$ for TEdBench and MagicBrush. We compare our metric, `AugCLIP` with CLIP$_{\text{dir}}$, which is the only metric that considers both preservation and modification.

`AugCLIP` generalizes to diverse editing scenarios. Text-guided image editing encompasses a wide range of tasks such as style editing, object replacement, local editing, partial alteration, and personalized generation. `AugCLIP` excels in the highly complex scenarios of local editing as conducted with the MagicBrush dataset. Moreover, `AugCLIP` highly aligns with human judgment in the personalized generation case of DreamBooth, which has been underexplored by previous metrics. We report how well `AugCLIP` aligns with human judgment with various editing cases provided in the EditVal dataset. Across all eight scenarios, except for the texture modification task, `AugCLIP` outperforms CLIP$_{\text{dir}}$ as shown in Tab. 5.

Augmenting CLIP$_{\text{dir}}$ cannot provide reliable evaluation result. We demonstrate the impact of augmenting CLIP$_{\text{dir}}$ with attributes, extracted in Sec. 4.1. Specifically, we replace the source and target text embeddings with the average of their respective attribute embeddings, $\mathcal{D}_S$ and $\mathcal{D}_T$. As shown in Tab. 6,

Table 5. Human Alignment Score s_{2AFC} on Various Text-guided Image Editing Scenarios.

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

this straightforward augmentation does not guarantee performance gain. AugCLIP consistently outperforms all description-augmented versions of CLIP$_{dir}$. This improvement is due to AugCLIP's approach, which derives the ideal representation by a minimum modification vector. This is fundamentally different from the naïve methodology of CLIP$_{dir}$, which relies on a simple subtraction of the source from the target.

Table 6. Effect of Augmenting Attributes.

	CelebA	EditVal	DreamBooth	TEdBench	MagicBrush
CLIP$_{dir}$	0.673	0.697	0.357	0.350	0.601
$+\mathcal{D}_S$	0.816	0.629	0.357	0.400	0.429
$+\mathcal{D}_T$	0.819	0.708	0.536	0.420	0.533
$+\mathcal{D}_S \cup \mathcal{D}_T$	0.816	0.607	0.536	0.440	0.402
AugCLIP	**0.883**	**0.831**	**0.857**	**0.570**	**0.889**

Weighting Strategy for Hyperplane We demonstrate the effectiveness of our weighting strategy, described in Eq. (6) in Tab. 7. Our weighting strategy enables AugCLIP to prioritize key features central to preservation and modification while estimating the representation of an ideally edited image, resulting in improvement in human alignment (s_{2AFC}) and balancing preservation and modification ($\mathbf{Acc}_{Both}$).

Table 7. Effect of Weighting Strategy.

	CelebA	EditVal	DreamBooth	TEdBench	MagicBrush
Unweighted	0.849	0.787	0.786	0.400	0.830
Weighted	**0.883**	**0.831**	**0.857**	**0.570**	**0.889**

Length and Number of Attributes In Tab. 8, we compare short (under 5 words) and long (over 5 words), as well as cases with a fixed number (10, 20, or 30) of attributes. We observe that using short attributes tends to outperform using long ones, likely because they focus on a single visual aspect, avoiding the potential entanglement of multiple aspects in a single attribute. Moreover, the number of attributes impacts the performance depending on the scene complexity of the benchmark dataset, as complex scenarios require a higher number of attributes to fully capture the scene's details. Therefore, allowing flexibility in the number of attributes, rather than imposing a strict limit, yields the best overall performance across all configurations.

Table 8. Effect of Length and Number of Attributes.

Length	Number	CelebA	EditVal	DreamBooth	TEdBench	MagicBrush
short	10	0.870	0.719	**0.857**	0.540	**0.889**
short	20	0.829	0.809	0.821	0.540	0.868
short	30	0.829	0.764	0.714	**0.570**	0.863
long	30	0.843	0.697	0.750	0.530	0.862
short	-	**0.883**	**0.831**	**0.857**	**0.570**	**0.889**

Description Randomness Using GPT-4V for description extraction introduces slight randomness across different seeds. Tests with five seeds show low variance (0.0197) in human alignment score (s_{2AFC}), indicating that the description generation process does not affect the scoring. We also demonstrate in the Appendix that extracted attributes with different seeds do not vary.

We provide analysis on the modification vector $\mathbf{v}$ and choice of classifier function, discussion on limitations, description prompting, and extensive qualitative results in the Appendix.

Limitations and Discussions

Our method, AugCLIP, adds additional computation time compared to CLIP$_{dir}$. It requires extracting attributes via MLLM, and then fitting a hyperplane to distinguish source and target attributes. This process takes around 12.3 seconds for attribute generation and 0.15 seconds for score computation. We will provide extracted attributes for all benchmark datasets to reduce this overhead, facilitating the evaluation process for future usage.

Conclusion

We present AugCLIP, a novel evaluation metric for text-guided image editing that balances source image preservation and target text modification. By leveraging a multi-modal language model to extract visual attributes and finding a separating hyperplane, AugCLIP estimates a representation of an ideal edited image that closely matches human preferences. Experiments on five benchmark datasets show that AugCLIP outperforms existing metrics, especially in challenging tasks, offering a more reliable tool for evaluating edits while preserving core attributes.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants (RS-2023-00209060 A Study on Optimization and Network Interpretation Method for Large-Scale Machine Learning), Institute for Information & communications Technology Planning & Evaluation(IITP) grants (No.2022-0-00713 Meta-learning applicable to real-world problems, RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)) funded by the Korea government(MSIT).

References

[1] Samyadeep Basu, Mehrdad Saberi, Shweta Bhardwaj, Atoosa Malemir Chegini, Daniela Massiceti, Maziar Sanjabi, Shell Xu Hu, and Soheil Feizi. Editval: Benchmarking diffusion based text-guided image editing methods. *arXiv preprint arXiv:2310.02426*, 2023. 1, 3, 6, 21

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1, 6, 7, 21

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[4] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 4, 29

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 6, 7, 21

[6] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1, 2

[7] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023. 1

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 6, 7, 21

[9] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 2023. 1, 6, 7, 21

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1, 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1, 2

[12] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 1, 4, 6, 7, 21

[13] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*, 2021. 1, 2, 6, 7, 21

[14] Yoonjeon Kim, Hyunsu Kim, Junho Kim, Yunjey Choi, and Eunho Yang. Learning input-agnostic manipulation directions in stylegan with text guidance. In *The Eleventh International Conference on Learning Representations*, 2022. 6, 7, 21

[15] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. Stylemc: Multi-channel based fast text-guided image generation and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 895–904, 2022. 1

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 6, 7, 21

[17] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 7, 21

[18] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7, 21

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 3, 6, 21

[21] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 17

[22] OpenAI. Gpt-4v(ision) system card, 2023. https://cdn.openai.com/papers/GPTV_System_Card.pdf/ [Accessed: 22-09-2024]. 5, 6, 27

[23] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 6, 7, 21

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 21

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 7

[26] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1, 6, 21

[27] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 6, 7, 21

[28] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 6, 21

[29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 2, 21

[30] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023. 27