On the Limitations of Vision-Language Models in Understanding Image Transforms

Ahmad Mustafa Anis[1],
[1]Cohere for AI Community,
Corresponding to: ahmadanis5050@gmail.com

Abstract

Vision Language Models like CLIP and SigLIP have shown great potential in downstream tasks such as Image/Video Generation, Visual Question Answering, Multimodal Chatbots, and Video Understanding, however downstream tasks often seem to lack basic understanding of common image transforms. In this paper, we investigate the image level understanding of Vision Language Models CLIP by OpenAI and SigLIP by Google. We conclude that these models lack in understanding of multiple image-level augmentations. We expanded our study to see how lack of understanding affects downstream models especially the use-case of image editing and how current state-of-the-art Image2Image models perform on simple transforms.

1. Introduction

Vision Language Models like CLIP [19] and SigLIP [30] have emerged as powerful frameworks that incorporate visual and text encoders aligned via large-scale pre-training on image-text pairs. These models have demonstrated impressive performance across various downstream tasks, including Text-to-Image Generation [21, 20], Video Action Recognition [27], and applications in the Biomedical domain [32]. CLIP-like pre-training has been extended to other modalities as well such as [4] introduces CLAP for Audio and Language. While these models exhibit remarkable capabilities and adaptability to diverse tasks, recent studies have revealed significant limitations in their fundamental understanding and reasoning abilities. In this paper, we investigate a fundamental aspect of visual reasoning in these models by addressing the question: *"Can Vision Language Embedding Models understand simple Image Transformations?"*

Understanding image augmentations is crucial for robust visual reasoning, as real-world images often appear with variations in brightness, contrast, rotation, and other transformations. While humans can recognize and describe such modifications, it remains unclear whether Vision Language Models truly comprehend these basic image manipulations. This question is also important to understand because while Vision Language models are now increasingly being used for Image Editing purposes, do these models truly understand images at a global level and can perform simple transforms that are commonly used in day-to-day image editing? Through a systematic evaluation of CLIP and SigLIP responses to various controlled augmentations, we show the limitations of the model's ability to reason about simple image transformations. We further discuss how this lack of understanding affects downstream tasks that use such models.

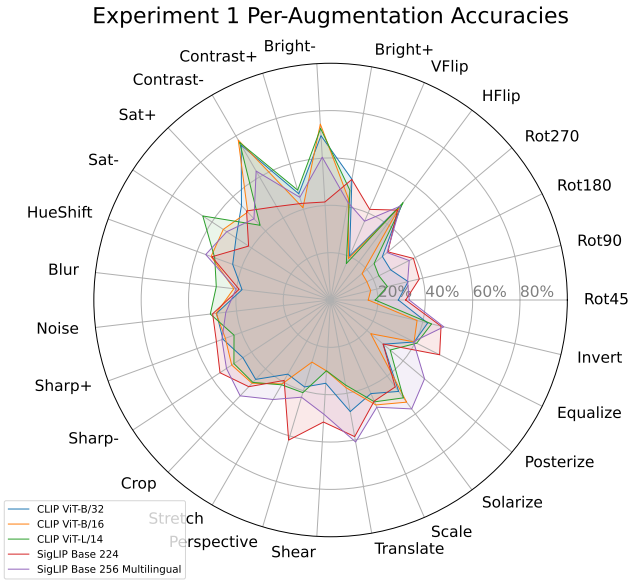

Figure 1. Comparison of image augmentation understanding between humans and Vision Language Models (CLIP/SigLIP). While humans can recognize and describe image transformations like rotation, brightness adjustment, and contrast changes, Vision Language Models show significant limitations in comprehending these basic image manipulations.

2. Related Works

Spatial Reasoning: Multiple works have been done to evaluate spatial reasoning in CLIP-related models. The paper "Visual-Spatial Reasoning" [11] shows that Vision Language models like CLIP are not good in spatial reasoning. ReCLIP [24] re-purposes CLIP to extend it to tasks related to Spatial Reasoning by introducing a Spatial Relation Resolver. Lewis et al. [9] show that CLIP models perform poorly on compositional visual reasoning tasks and cannot encode compositional concepts or bind variables in a structure-sensitive way (e.g., differentiating "cube behind sphere" from "sphere behind cube"). OmniCLIP [12] shows that CLIP falls short in capturing and integrating spatial-temporal features which is essential for video recognition and proposes a framework to extend CLIP for spatial temporal features for video recognition.

Linguistic Reasoning: Studies have shown that CLIP also does not perform well on pure linguistic tasks. Sam et al. [22] show that CLIP's embedding space lacks the structure of their purely text-based alternatives (e.g., Text(*"King"*) − Text(*"Man"*) + Text(*"Woman"*) ≈ Text(*"Queen"*)). CyCLIP [5] demonstrates that image and text representations learned by CLIP are not interchangeable and can lead to inconsistent downstream predictions.

Counting: Counting is an interesting challenge where the model must count the number of entities in an image. Paiss et al. [16] introduce a novel training framework and benchmark to improve the quantitative understanding of VLMs. Ma et al. [13] enhance CLIP's ability to count with a focus on estimating crowd sizes from images. Zhang et al. [33] studied the question "Can CLIP Count Stars?" and showed that CLIP is not reliable in counting stars and contains a quantity bias.

Robustness: Multiple studies have been done to evaluate the robustness of Vision Language Models like CLIP. Tu et al. [26] show that CLIP exhibits strong shape bias. Schlarmann et al. [23] propose an unsupervised adversarial fine-tuning technique to train a robust CLIP vision encoder that is safe against adversarial attacks. Laroudie et al. [8] demonstrate that CLIP is **overconfident** in incorrect predictions, making its predictions less reliable. They also show Domain Shift Vulnerability, where there is a significant accuracy drop when domains are shifted. They propose LP-CLIP, a novel Knowledge distillation framework to improve robustness in CLIP models.

3D Understanding: Recent works have explored CLIP's capabilities in understanding and generating 3D content. CLIP-Forge [**?**] introduces a zero-shot text-to-shape generation method that addresses the scarcity of paired text-shape data using CLIP's pre-trained image-text representations. Sbrolli et al. [**?**] propose unsupervised methods to enhance contrastive text-image-3D alignment by leveraging CLIP's knowledge of textual and 2D data for computing neural perceived similarity between 3D samples. CLIP2Scene [**?**] makes the first attempt to transfer CLIP knowledge to 3D scene understanding, achieving impressive results in annotation-free 3D semantic segmentation and fine-tuning scenarios. CISP [**?**] introduces a framework to enhance 3D shape synthesis from images by aligning 2D images with 3D shapes in a shared embedding space, showing that incorporating explicit 3D knowledge can improve generation coherence compared to standard CLIP-guided models.

3. Methodology

We divided our experiments in three parts to test how well CLIP and SigLIP perform with a focus on language, vision, and simple classification.

3.1. Data Collection

We used Flicker8k dataset [7] and developed a simple annotation technique to create our augmented dataset. For each image-caption pair, we apply a random augmentation from our defined set of transformations. The annotation follows a consistent pattern where we append the augmentation description to the original caption.

Original captions are augmented using F-strings in Python:

```
f"{original_caption},
f"{augmented_caption}
```

For example, given an original caption describing a scene and applying a sharpness reduction (using PyTorch's `RandomAdjustSharpness` with `sharpness_factor=0.5`), we get:

> "A child in a pink dress is climbing up a set of stairs in an entry way, this image has decreased sharpness"

This approach creates a parallel dataset where each augmented image is paired with an explicitly described transformation, enabling direct evaluation of vision-language models' capability to recognize image modifications.

3.1.1 Image Augmentation Methodology

We implemented a comprehensive set of 24 image transformations across six primary categories to evaluate vision-language models' robustness to image variations. Each transformation was paired with a natural language description to test the models' understanding of visual modifications. All augmentations were implemented using PyTorch's `torchvision.transforms` library[14].

3.1.2 Geometric Transformations

We implemented rotational and flip-based transformations to test spatial understanding:

- **Rotations:** Four distinct rotation angles (45, 90, 180, 270) using `RandomRotation` to assess orientation sensitivity
- **Flips:** Horizontal and vertical flips (`RandomHorizontalFlip`, `RandomVerticalFlip`) to evaluate mirror symmetry comprehension

3.1.3 Color Space Modifications

To evaluate color perception robustness, we implemented bidirectional adjustments using `ColorJitter`:

- **Brightness:** $\pm$50% modifications (factor: 1.5 for increase, 0.5 for decrease)
- **Contrast:** Similar bidirectional adjustments
- **Saturation:** Controlled adjustments to color intensity
- **Hue:** Warm color shifts implemented via `ColorJitter(hue=0.1)`

3.1.4 Clarity and Focus Transformations

To test visual acuity understanding:

- **Blur:** Gaussian blur with kernel size (5,5) and variable sigma (0.1-2.0)
- **Sharpness:** Bidirectional sharpness modifications ($\pm$50% from baseline) using `RandomAdjustSharpness`

3.1.5 Geometric Distortions

Complex spatial modifications to test structural understanding:

- **Perspective:** Controlled perspective shifts (`RandomPerspective`, `distortion_scale=0.3`)
- **Affine Transformations:**
  - Shear: 30horizontally
  - Translation: 20% in both dimensions
  - Scale: 20% enlargement

3.1.6 Resolution and Size Modifications

To evaluate scale invariance:

- **Center Crop:** 224px crop from 256px resized images using `Compose([Resize(256), CenterCrop(224)])`
- **Aspect Ratio:** Horizontal stretching to 160$\times$256 pixels

3.1.7 Image Processing Effects

Advanced transformations to test robustness to common image processing operations:

- **Noise:** Gaussian noise injection ($\sigma = 0.1$)
- **Intensity:**
  - Solarization (`RandomSolarize`, threshold=128)
  - Posterization (`RandomPosterize`, 2-bit)
  - Histogram equalization (`RandomEqualize`)
- **Color Inversion:** Complete color space inversion (`RandomInvert`)

Each transformation was implemented with precise control parameters to ensure reproducibility. Each augmented image was paired with its original caption plus a natural language description of the applied transformation, forming evaluation pairs for our vision-language models. For example, a rotation transformation would be described as "this image is rotated 90 degrees clockwise," maintaining a natural language format consistent with typical image descriptions.

3.1.8 Data Distribution

Our augmentation strategy involved applying diverse image transformations to the Flickr8k dataset. As shown in Figure 2, we implemented a comprehensive set of augmentations spanning different aspects of image manipulation. The distribution reveals balanced coverage across transformation types, with geometric transformations like rotation and flips being among the most frequently applied modifications.

The categorical analysis (Figure 3) demonstrates the diversity of our augmentation strategy. Color transformations constitute approximately 43.6% of all augmentations, followed by processing transformations (approximately 41.1%), distortion effects (approximately 7.8%), and clarity adjustments (approximately 7.5%). This distribution ensures comprehensive evaluation of vision-language models'

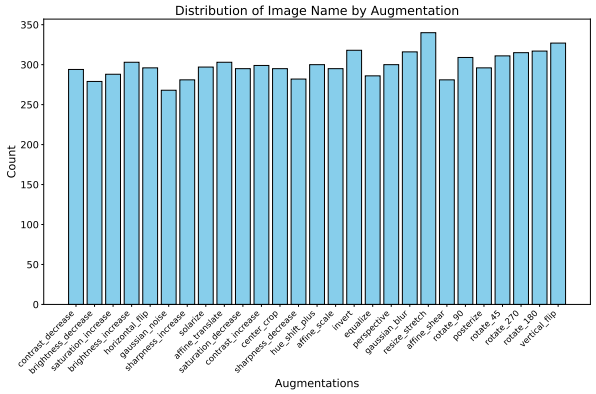

Figure 2. Distribution of individual augmentations applied to the Flickr8k dataset. The augmentations span across multiple transformation types including geometric (rotations, flips), color adjustments (brightness, contrast, saturation), clarity modifications (blur, sharpness), and various image processing effects.

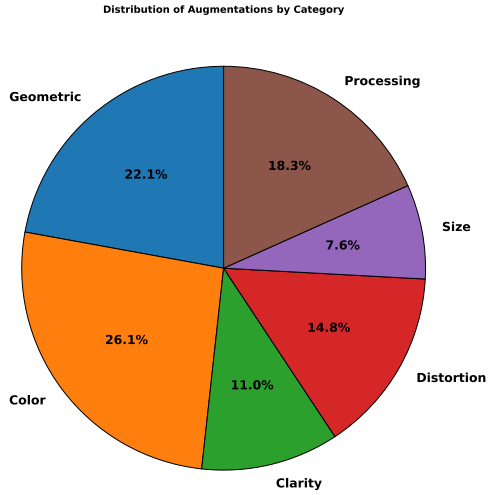

Figure 3. Distribution of augmentations applied to the dataset. The augmentations are grouped into six primary categories: Geometric (rotations and flips), Color (brightness, contrast, saturation, and hue adjustments), Clarity (blur and sharpness), Distortion (perspective and affine transformations), Size (cropping and stretching), and Processing (noise, solarization, posterization, and other effects).

robustness and interpretative capability across various image modification types.

Each category serves a specific purpose in our evaluation:

- **Geometric Transformations:** Test the ability to handle spatial modifications such as translation, rotation, flipping, and affine transformations.
- **Color Transformations:** Evaluate perception of color intensity, brightness, saturation, hue, and contrast variations.
- **Clarity Modifications:** Assess recognition capabilities under different levels of image sharpness and blurring.
- **Distortion Effects:** Test robustness to noise and other distortion effects like perspective transformations and affine distortions.
- **Size Modifications:** Evaluate performance under changes in image dimensions, such as cropping and stretching.
- **Processing Transformations:** Assess robustness to high-level image processing effects like solarization, posterization, inversion, and equalization.

The balanced distribution across these categories ensures comprehensive evaluation of model capabilities across different types of image modifications. Each augmentation is paired with a natural language description, enabling direct assessment of the model's ability to recognize and articulate specific types of image transformations.

3.2. Experiment 1

We first evaluate whether vision-language models can correctly associate augmented descriptions with their corresponding modified images. This experiment tests the model's ability to understand the relationship between textual descriptions of image modifications and the actual visual changes.

3.2.1 Methodology

For each image-caption pair (I, C), we:

1. Generate an augmented image I_{aug} using a random transformation T
2. Create an augmented caption C_{aug} by appending the transformation description to the original caption
3. Compare similarity scores:
   - $s_1 = sim(I_{aug}, C_{aug})$: Similarity between augmented image and augmented caption
   - $s_2 = sim(I_{orig}, C_{aug})$: Similarity between original image and augmented caption
4. Consider prediction correct if $s_1 > s_2$

Mathematically, the accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[sim(I_{aug}^{(i)}, C_{aug}^{(i)}) > sim(I_{orig}^{(i)}, C_{aug}^{(i)})] \tag{1}$$

where N is the total number of samples and $\mathbb{1}[\cdot]$ is the indicator function.

3.2.2 Results

Table 1 shows the detailed performance metrics across different model variants:

Model	Accuracy (%)
CLIP ViT-B/32	42.80
CLIP ViT-B/16	40.87
CLIP ViT-L/14	43.10
SigLIP Base 224	45.78
SigLIP Base 256 Multilingual	47.21

Table 1. Experiment 1 Overall Accuracy Comparison Across Models

3.2.3 Analysis

The results reveal several key insights:

- **Model Architecture Impact:** Larger models (e.g., ViT-L/14) generally show improved performance, suggesting that increased model capacity helps in understanding transformation descriptions. Similarly, CLIP models seem to perform better compared to SigLIP models on some individual types of transformations, as shown in Figure 4 but SigLIP outperforms CLIP when comparing mean accuracy.

- **Transformation Types:** Models show varying performance across different types of augmentations:

  - CLIP and SigLIP perform better in Color and Distortion based augmentations as compared to rest of augmentations however SiGLIP seems to perform better in size and processing based augmentations as shown in Figure 5

3.3. Experiment 2

Our second experiment evaluates the Vision Language Models' ability to match transformed images with the augmented textual description. This experiment assesses whether models can recognize when an augmented image better matches a description that contains transformation details or a description that has no transformation details.

3.3.1 Methodology

For each sample i in the dataset, we perform the following steps:

First, we select an original image $I^{(i)}$ and apply an augmentation transformation $T^{(i)}$ to obtain the augmented image:

$$I_{\text{aug}}^{(i)} = T^{(i)}\left(I^{(i)}\right) \tag{2}$$

Next, we prepare the corresponding captions. We obtain the original caption $C_{\text{orig}}^{(i)}$ associated with $I^{(i)}$ and define the textual description of the transformation $T^{(i)}$ as $\text{desc}\left(T^{(i)}\right)$. The augmented caption is then created by appending the augmentation description to the original caption:

$$C_{\text{aug}}^{(i)} = C_{\text{orig}}^{(i)} + \text{", "} + \text{desc}\left(T^{(i)}\right) \tag{3}$$

We compute the similarity between the augmented image and both the original and augmented captions. The similarity with the original caption is:

$$s_1^{(i)} = \text{sim}\left(I_{\text{aug}}^{(i)}, C_{\text{orig}}^{(i)}\right) \tag{4}$$

and the similarity with the augmented caption is:

$$s_2^{(i)} = \text{sim}\left(I_{\text{aug}}^{(i)}, C_{\text{aug}}^{(i)}\right) \tag{5}$$

where $\text{sim}(I, C)$ denotes the similarity function (e.g., cosine similarity) between the embeddings of image I and caption C.

The model is considered to have correctly associated the augmented image with the augmented caption if:

$$s_2^{(i)} > s_1^{(i)} \tag{6}$$

The overall accuracy over the dataset is computed as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\, s_2^{(i)} > s_1^{(i)} \,\right] \tag{7}$$

where N is the total number of samples, and $\mathbb{1}[\,\cdot\,]$ is the indicator function defined as:

$$\mathbb{1}\left[\,\text{condition}\,\right] = \begin{cases} 1, & \text{if condition is true} \\ 0, & \text{if condition is false} \end{cases} \tag{8}$$

3.3.2 Analysis

The results of Experiment 2 reveal some interesting analysis as shown in Table 2. In experiment 2, all CLIP models perform really well in terms of accuracy showing given an augmented image, Vision Language Models have a better tendency to recognize the augmented prompt in contrast to the actual prompt. However, **figure 6** shows that there is a very small difference in the similarity score indicating that even though CLIP models perform very well, they can not differentiate between the normal prompt and augmented prompt really well.

3.3.3 Per-Augmentation Analysis

Figure 7 shows the results of CLIP and SigLIP for experiment 2 per augmentation category, these results reflect our

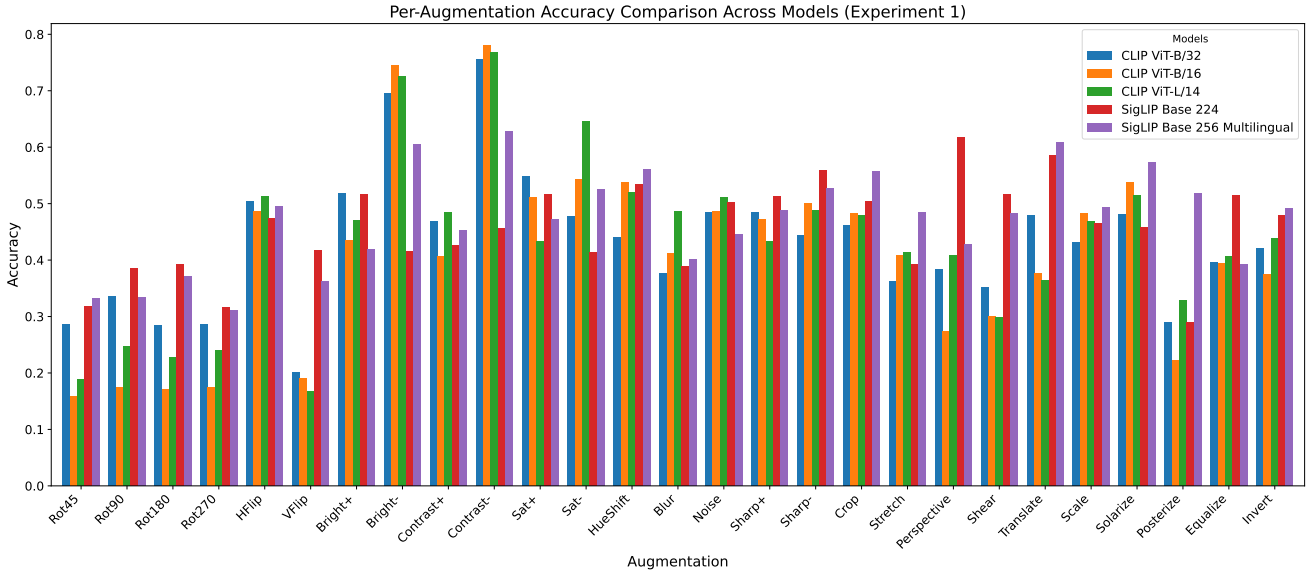

Figure 4. Accuracy comparison of model performance on augmented prompt recognition. Higher values indicate better understanding of the relationship between textual descriptions of transformations and their visual manifestations.

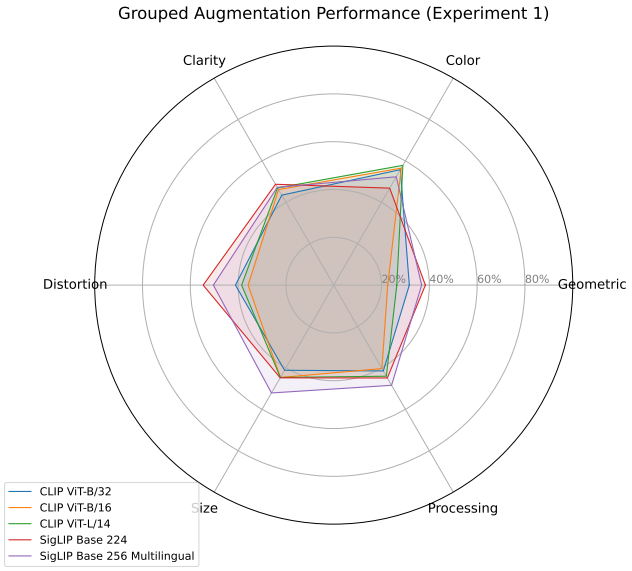

Figure 5. Comparison of model performance on augmentations grouped according to their properties.

Table 2. Experiment 2 Mean Accuracy Comparison

Model	Mean Accuracy
CLIP ViT-B/16	99.57%
CLIP ViT-B/32	98.67%
CLIP ViT-L/14	98.15%
SigLIP Base 224	64.40%
SigLIP Base 256 Multilingual	47.41%

initial analysis that CLIP model is performing well in differentiating between original prompt and augmented prompt when we calculate the similarity. Figure 8 shows the performance of CLIP and SigLIP when grouped by the categories mentioned earlier and further strengthen the result that CLIP models are performing better.

3.4. Experiment 3: Augmentation Classification

In our third experiment, we evaluate models' ability to identify specific image transformations from a predefined set of augmentations. Unlike Experiments 1 and 2, which focus on pairwise comparisons, this experiment tests direct classification capability across all possible augmentation types.

3.4.1 Methodology

For each augmented image I_{aug}, we perform the following steps:

First, we present the model with the augmented image I_{aug} and compare it against all possible augmentation descriptions $\mathcal{A}$, consisting of 27 types as described in Section 3.1. For each augmentation description $a \in \mathcal{A}$, we calculate the similarity score between the image and the textual description:

$$\text{score}_a = \text{sim}\left(I_{\text{aug}}, \text{“}a\text{”}\right) \tag{9}$$

We rank all augmentation descriptions based on their similarity scores in descending order. The rank of the true augmentation description t is determined by:

$$\text{rank}_t = \left|\{a \in \mathcal{A} : \text{score}_a > \text{score}_t\}\right| + 1 \tag{10}$$

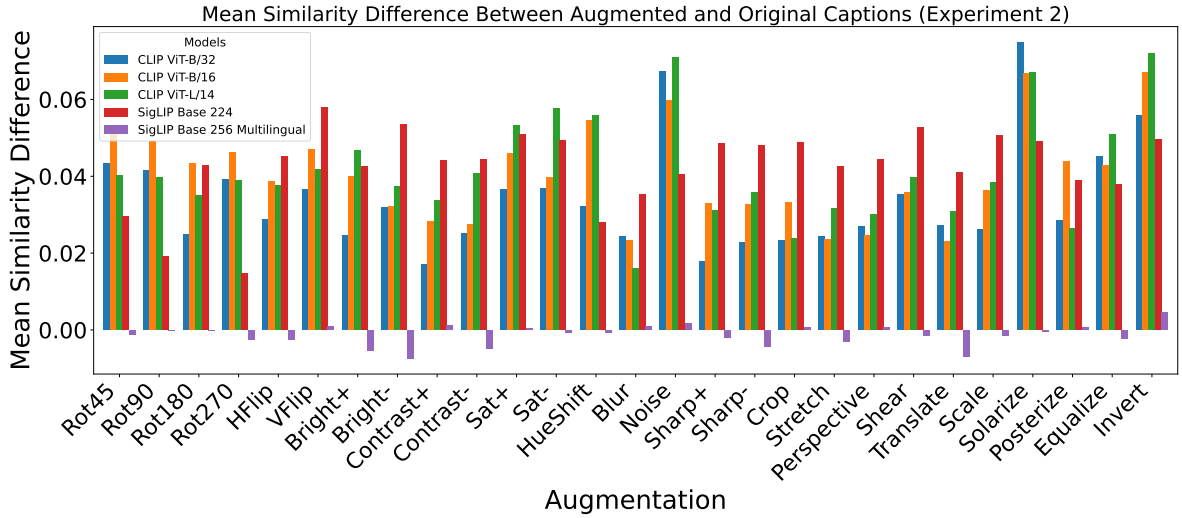

Figure 6. Mean difference between similarity of augmented image with actual prompt and augmented image with augmented prompt

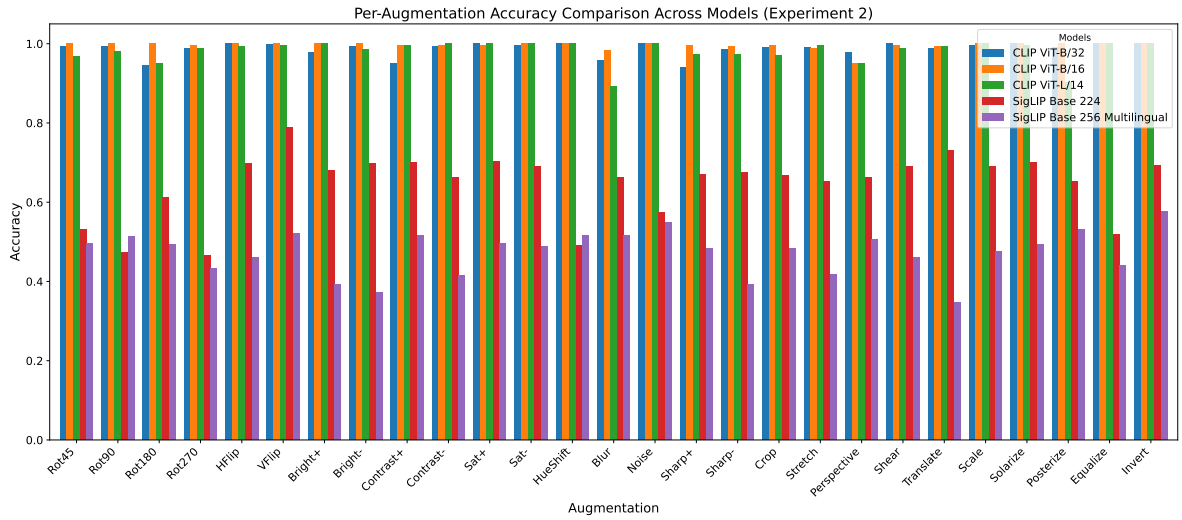

Figure 7. Per Augmentation Accuracy Experiment 2

We evaluate the model's performance using the following metrics:

- **Top-1 Accuracy**: The proportion of times the correct augmentation t is ranked first ($\text{rank}_t = 1$).
- **Top-5 Accuracy**: The proportion of times t is among the top five predictions ($\text{rank}_t \leq 5$).
- **Mean Rank**: The average rank position of the correct augmentation t across all samples.

This approach assesses the model's ability to accurately identify the augmentation applied to an image by matching it with the correct textual description.

3.4.2 Results

This experiment shows Vision Language Understanding of Augmentations where can a model associate itself with the correct Augmentation. Figure 10 shows the Top-1% accuracy performance of Vision Language Models on just identifying the correct augmentation class where for most of the augmentation, the accuracy is 0% and model was not able to identify a single correct example. Table 3 compares the Top-1% and Top-5% accuracy and shows that Vision Language Model can not classify the augmentation correctly. 9 shows that while Vision Language Models are bad at classifying the correct augmentation, they perform extremely poor at processing augmentations such as translation, rotation, scaling, and cropping, these are the transformations where humans excel in general.

Table 3. Comparison of Top-1 and Top-5 Accuracies for Each Model

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)
ViT-B/32	3.61	18.40
ViT-B/16	3.50	17.12
ViT-L/14	3.57	15.28
SigLIP Base 224	2.81	16.40
SigLIP Base 256 Multilingual	3.19	18.06

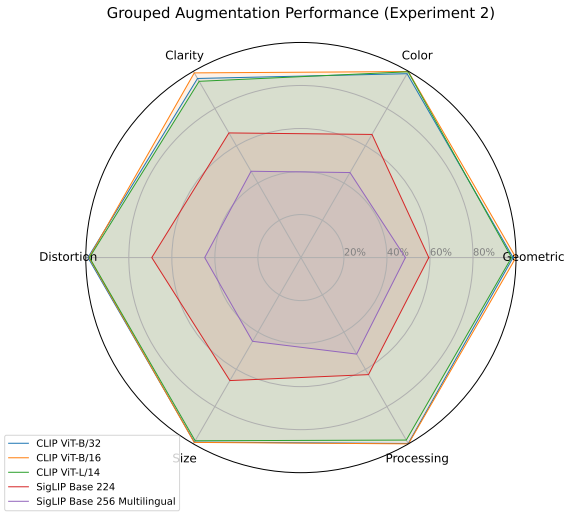

Figure 8. Per Augmentation Accuracy Experiment 2

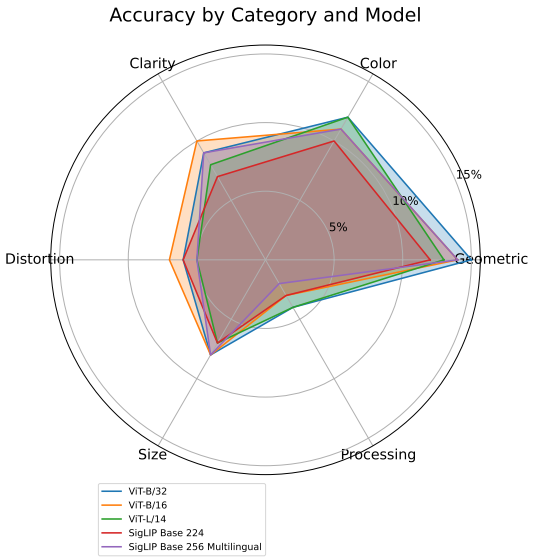

Figure 9. Mean difference between similarity of augmented image with actual prompt and augmented image with augmented prompt

4. Impact on Downstream task

With the rise of AI in Image/Video Editing[25], this study reveals an important lack of understanding of the image level in vision language models. These models are backbone of the models that are used in downstream tasks such as Image Generation[21][27], Controlled Image Generation models [31][10][29][18], Image-to-Image Editing[17][3][15] and multiple other downstream tasks. Different types of image transformation are a basic tool in traditional image editing tools such as Photoshop[1]. Table 4 shows some example of the common AI Image editing models, Instruct Pix2Pix[2], Dall.E 3[6], and IP Adapter[28] with the prompt **"Rotate the input image 90 degrees"**. The results show that none of these commonly used models was able to understand this basic instruction and failed to generate a an image with basic transformation applied.

This paper also motivates us to think about newer training pradigms for Vision Language Models where the model can have a global context, understands image at a deeper level that will help us to unlock newer capabalities in downstream tasks.

5. Acknowledgement

Valuable comments and insightful feedback from Anas Zafar are gratefully acknowledged.

References

[1] Adobe. Transform and rotate image selections and layers. https://helpx.adobe.com/photoshop/using/transforming-objects.html, 2024. Adobe Photoshop User Guide.

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[3] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models, 2023.

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022.

[5] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022.

[6] Gabriel Goh, James Betker, Li Jing, and Aditya Ramesh. Dall·e 3, 2024.

[7] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and

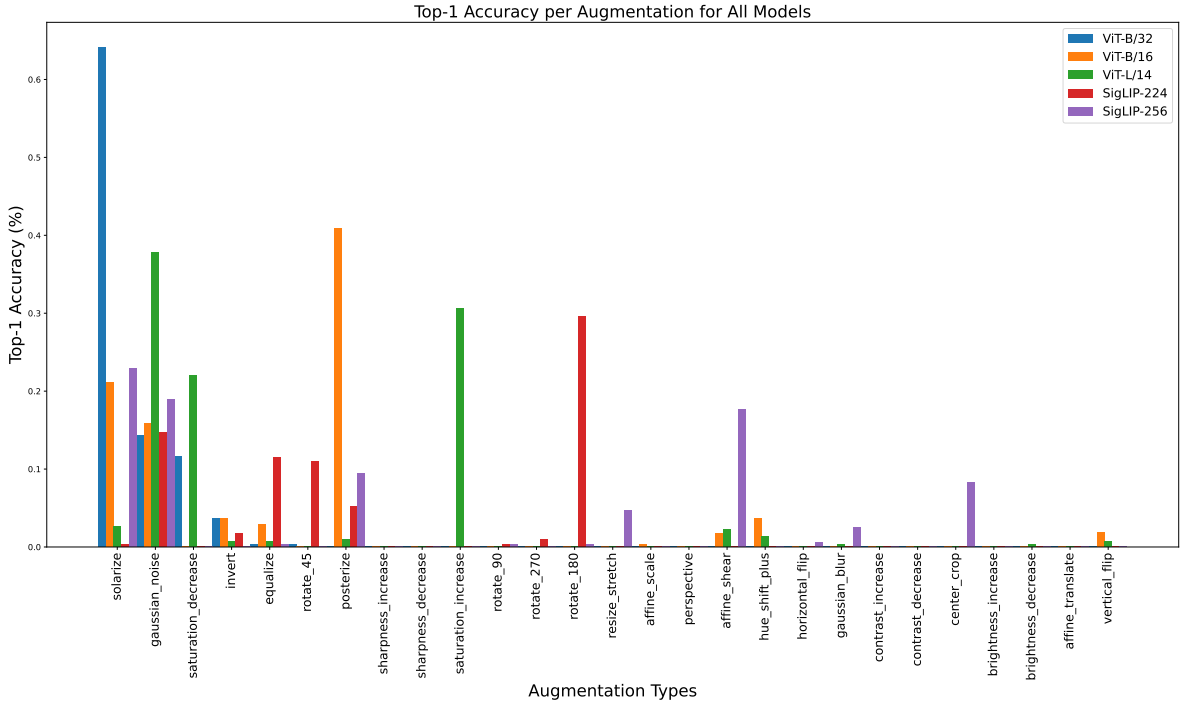

Figure 10. Top-1 Accuracy per Augmentation type for all models

Model	Input Image	Output Image
DALL·E		
Instruct Pix2Pix		
IP Adapter		

Table 4. Qualitative analysis table comparing input images and output transformations (rotation 90 degrees) for different models.

evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[8] Clement Laroudie, Andrei Bursuc, Mai Lan Ha, and Gianni Franchi. Improving clip robustness with knowledge distillation and self-training, 2023.

[9] Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack

Merullo, Stephen H. Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models, 2024.

[10] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback, 2024.

[11] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023.

[12] Mushui Liu, Bozheng Li, and Yunlong Yu. Omniclip: Adapting clip for video recognition with spatial-temporal omniscale feature learning, 2024.

[13] Yiming Ma, Victor Sanchez, and Tanaya Guha. Clip-ebc: Clip can count accurately through enhanced blockwise classification, 2024.

[14] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

[15] Naoki Matsunaga, Masato Ishii, Akio Hayakawa, Kenji Suzuki, and Takuya Narihira. Fine-grained image editing by pixel-wise guidance using diffusion models, 2023.

[16] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten, 2023.

[17] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023.

[18] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild, 2023.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[22] Dylan Sam, Devin Willmott, Joao D. Semedo, and J. Zico Kolter. Finetuning clip to reason about pairwise differences, 2024.

[23] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models, 2024.

[24] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension, 2022.

[25] Yuying Tang, Ningning Zhang, Mariana Ciancia, and Zhigang Wang. Exploring the impact of ai-generated image tools on professional and non-professional users in the art and design fields, 2024.

[26] Weijie Tu, Weijian Deng, and Tom Gedeon. Toward a holistic evaluation of robustness in clip models, 2024.

[27] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition, 2021.

[28] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

[29] Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems, 2024.

[30] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

[31] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[32] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024.

[33] Zeliang Zhang, Zhuo Liu, Mingqian Feng, and Chenliang Xu. Can clip count stars? an empirical study on quantity bias in clip, 2024.