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

Ahmad Mustafa Anis[1], Hasnain Ali[2], Saquib Sarfraz[3]

[1]Cohere for AI Community, [2]Arbisoft, [3]Karlsruhe Institute of Technology

Corresponding to: ahmadanis5050@gmail.com

Abstract

Vision Language Models (VLMs) have demonstrated significant potential in various downstream tasks, including Image/Video Generation, Visual Question Answering, Multimodal Chatbots, and Video Understanding. However, these models often struggle with basic image transformations. This paper investigates the image-level understanding of VLMs, specifically CLIP by OpenAI and SigLIP by Google. Our findings reveal that these models lack comprehension of multiple image-level augmentations. To facilitate this study, we created an augmented version of the Flickr8k dataset, pairing each image with a detailed description of the applied transformation. We further explore how this deficiency impacts downstream tasks, particularly in image editing, and evaluate the performance of state-of-the-art Image2Image models on simple transformations.

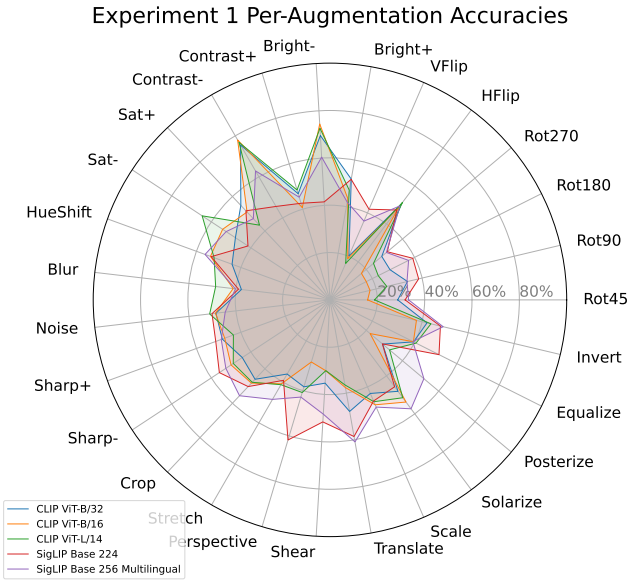

Figure 1. Comparison of image augmentation understanding between humans and Vision Language Models (CLIP/SigLIP). While humans can recognize and describe image transformations like rotation, brightness adjustment, and contrast changes, Vision Language Models show significant limitations in comprehending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

1. Introduction

Vision Language Models like CLIP [21] and SigLIP [36] 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 [22, 23], Video Action Recognition [33], and applications in the Biomedical domain [38]. CLIP-like pre-training has been extended to other modalities as well, such as CLAP for Audio and Language [5].

However, despite their broad success, a fundamental question remains unanswered: *"Can Vision Language Embedding Models understand simple Image Transformations?"* This question is particularly crucial as these models are increasingly deployed for image editing tasks, where understanding basic transformations is essential for meaningful manipulation. As shown in Figure 1, our systematic evaluation reveals a significant gap between human and machine understanding of common image modifications.

Understanding image augmentations is fundamental for robust visual reasoning, as real-world images frequently appear with variations in brightness, contrast, rotation, and other transformations. While these models are designed to exhibit robustness and invariant behavior to standard image transforms, we argue that this invariance might come at the cost of explicit understanding. Although invariance was valuable for earlier models trained in data-constrained environments [8] on datasets like ImageNet [24], modern foundation models trained on vast amounts of data should ideally possess both invariance when needed and explicit understanding of transformations when required.

Through comprehensive evaluation of CLIP and SigLIP responses to various controlled augmentations, we demonstrate significant limitations in these models' ability to reason about simple image transformations. Our findings have important implications for downstream tasks that rely on these models, particularly in applications requiring explicit understanding of image modifications. This work not only highlights a critical gap in current Vision Language Models but also emphasizes the need for developing approaches that can better capture fundamental aspects of visual reasoning.

2. Related Works

Spatial Reasoning: Multiple works have been done to evaluate spatial reasoning in CLIP-related models. The paper "Visual-Spatial Reasoning" [13] shows that Vision Language models like CLIP are not good in spatial reasoning. ReCLIP [30] re-purposes CLIP to extend it to tasks related to Spatial Reasoning by introducing a Spatial Relation Resolver. Lewis et al. [11] 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 [14] 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. [25] 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 [6] 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. [18] introduce a novel training framework and benchmark to improve the quantitative understanding of VLMs. Ma et al. [15] enhance CLIP's ability to count with a focus on estimating crowd sizes from images. Zhang et al. [39] 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. [32] show that CLIP exhibits strong shape bias. Schlarmann et al. [29] propose an unsupervised adversarial fine-tuning technique to train a robust CLIP vision encoder that is safe against adversarial attacks. Laroudie et al. [10] 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 [26] 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. [27] 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 [3] 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 [28] 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. Dataset & Augmentation Methodology

To thoroughly evaluate the image-level understanding of VLMs, we created an augmented version of the Flickr8k [9] dataset. This dataset was chosen for its diverse range of images and corresponding captions, providing a robust foundation for our experiments. We developed a systematic approach to apply a variety of image transformations, ensuring each augmented image was paired with a detailed natural language description of the applied modification. This section outlines our data collection process, the specific augmentation techniques employed, and the distribution of these augmentations across the dataset.

3.1. Data Collection

We used Flicker8k dataset [9] and developed a simple annotation technique to create our augmented dataset. For each image-caption pair, we apply a random augmentation and append the transformation description to the original caption:

> "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.

3.2. Image Augmentation Methodology

We implemented 24 image transformations across six categories using PyTorch's `torchvision.transforms` library[16]:

3.2.1. Geometric Transformations

- **Rotations:** Four angles (45, 90, 180, 270)
- **Flips:** Horizontal and vertical

3.2.2. Color Space Modifications

Bidirectional adjustments for:

- **Brightness:** $\pm$50% modifications
- **Contrast:** Similar bidirectional adjustments
- **Saturation:** Controlled adjustments to color intensity
- **Hue:** Warm color shifts (hue=0.1)

3.2.3. Clarity and Focus Transformations

- **Blur:** Gaussian blur (kernel size 5,5)
- **Sharpness:** Bidirectional modifications ($\pm$50%)

3.2.4. Geometric Distortions

- **Perspective:** Controlled shifts (distortion_scale=0.3)
- **Affine:** Shear (30), Translation (20%), Scale (20%)

3.2.5. Resolution and Size Modifications

- **Center Crop:** 224px crop from 256px images
- **Aspect Ratio:** Horizontal stretching (160$\times$256 pixels)

3.2.6. Image Processing Effects

- **Noise:** Gaussian noise ($\sigma = 0.1$)
- **Intensity:** Solarization (threshold=128), Posterization (2-bit), Equalization
- **Color Inversion:** Complete color space inversion

Each augmented image was paired with its original caption plus a description of the applied transformation.

3.3. Data Distribution

As shown in Figure 2, we implemented diverse augmentations with balanced coverage across transformation types. The categorical analysis (Figure 3) shows that color transformations constitute approximately 43.6% of all augmentations, followed by processing transformations (41.1%), distortion effects (7.8%), and clarity adjustments (7.5%).

Each category serves a specific evaluation purpose:

- **Geometric:** Tests spatial understanding (rotation, flipping)
- **Color:** Evaluates perception of color variations
- **Clarity:** Assesses recognition under different sharpness levels
- **Distortion:** Tests robustness to perspective and affine changes
- **Size:** Evaluates performance under dimensional changes
- **Processing:** Assesses robustness to image processing effects

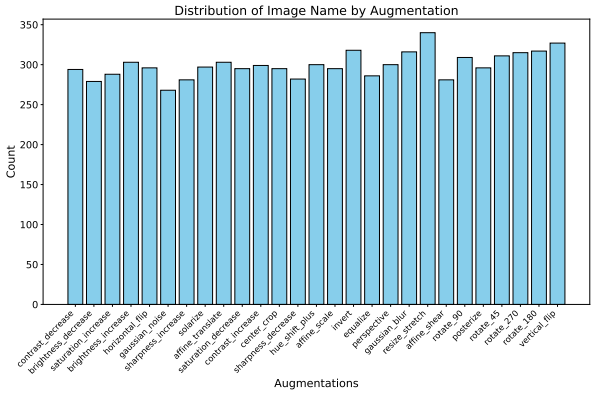

Figure 2. Distribution of individual augmentations applied to the Flickr8k dataset. The augmentations span across multiple transformation types including geometric (rotations, flips), color adjustments (brightness, contrast, saturation), clarity modifications (blur, sharpness), and various image processing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

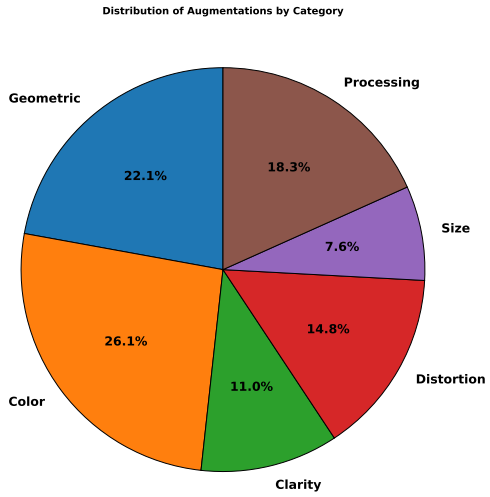

Figure 3. Distribution of augmentations applied to the dataset. The augmentations are grouped into six primary categories: Geometric (rotations and flips), Color (brightness, contrast, saturation, and hue adjustments), Clarity (blur and sharpness), Distortion (perspective and affine transformations),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

This distribution ensures comprehensive evaluation of model capabilities across different types of image modifications.

4. Evaluation of Vision Language Models

In this section, we present a comprehensive evaluation of VLMs, specifically focusing on their ability to understand and process image augmentations. Our evaluation is

structured into three key experiments: understanding augmented descriptions, matching augmented images with descriptions, and classifying image transformations. Each experiment is designed to test different aspects of the models' capabilities, providing a holistic view of their strengths and limitations. Through these experiments, we aim to uncover the extent to which VLMs can accurately interpret and respond to various image modifications, shedding light on their potential and areas for improvement.

4.1. Understanding Augmented Descriptions

We first assess the ability of VLMs to accurately associate textual descriptions of image augmentations with their corresponding modified images. This evaluation aims to determine whether the models can comprehend and link the specified transformations described in text to the visual alterations present in the images. By examining the relationship between the augmented descriptions and the visual changes, we can gauge the models' proficiency in understanding and interpreting basic image modifications.

4.1.1. 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.

4.1.2. Results

Table 1. Experiment 1 Overall Accuracy Comparison Across Models

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 shows the accuracy across different model variants.

- **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

4.2. Matching Augmented Images with Descriptions

This evaluation examines the ability of VLMs to accurately match transformed images with their corresponding augmented textual descriptions. The objective is to determine whether the models can effectively identify when an augmented image corresponds to a description that includes specific transformation details, as opposed to a description without such details. By evaluating the models' capacity to link visual modifications with the appropriate textual descriptions, we gain insights into their effectiveness in image-text alignment tasks.

4.2.1. 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:

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

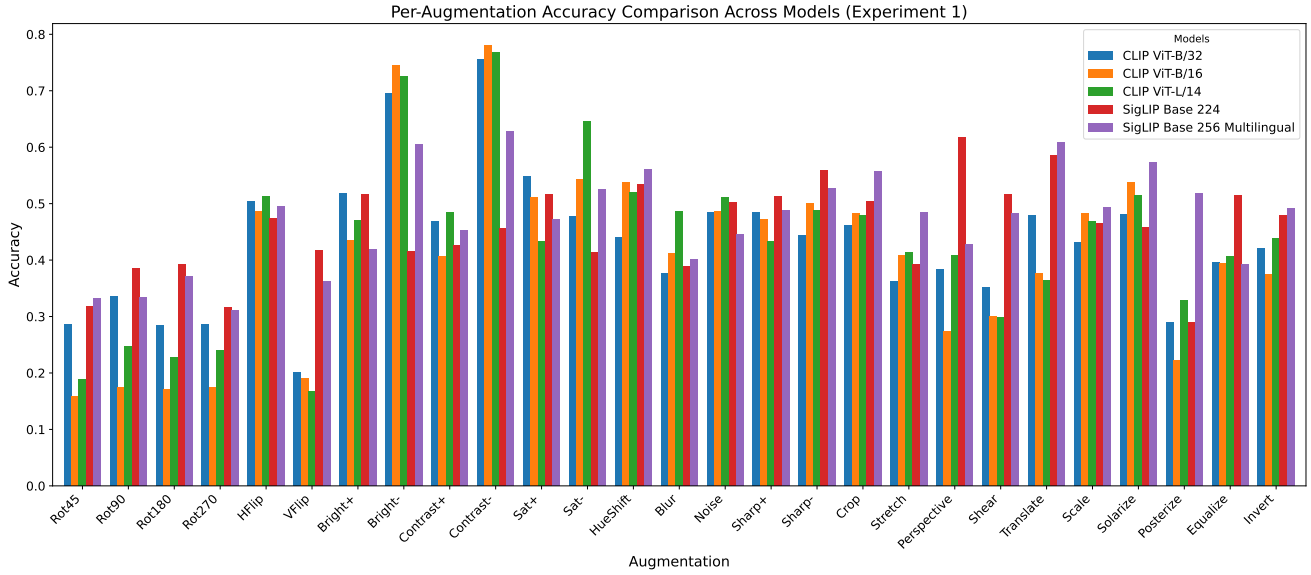

Figure 4. Accuracy comparison of model performance on augmented prompt recognition. Higher values indicate better understanding of the relationship between textual descriptions of transformations and their visual manifestations.

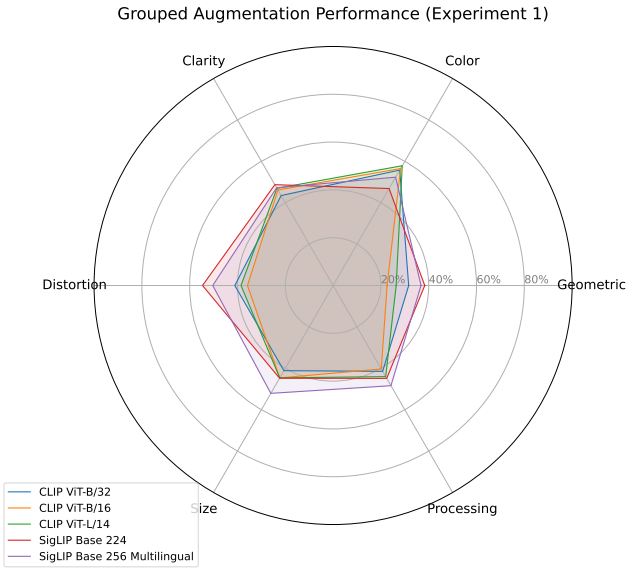

Figure 5. Comparison of model performance on augmentations grouped according to their properties.

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:

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

4.2.2. 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.

4.2.3. Per-Augmentation Analysis

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

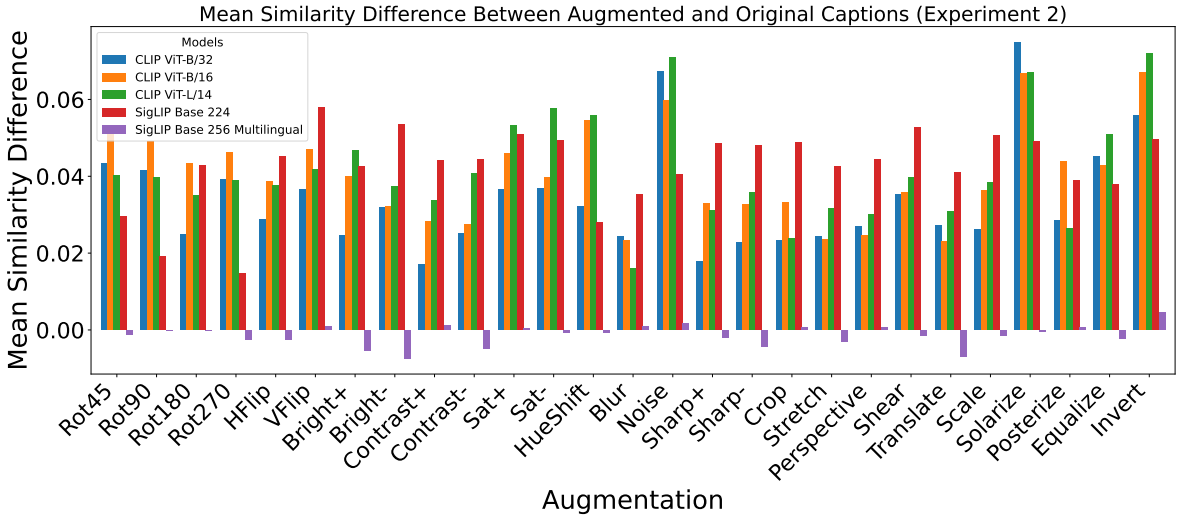

Figure 6. Mean difference between similarity of augmented image with actual prompt and augmented image with augmented prompt

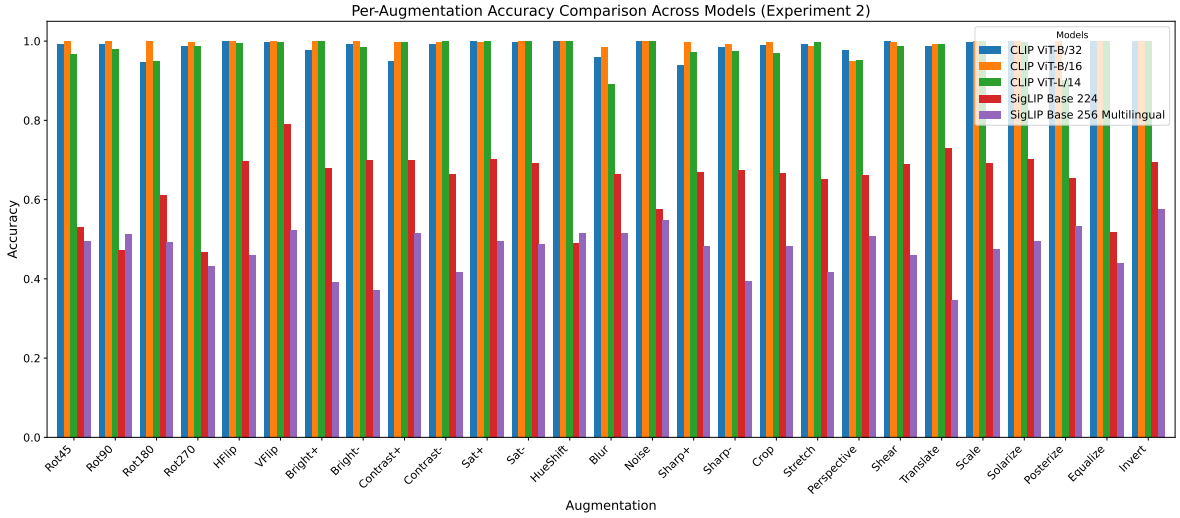

Figure 7. Per Augmentation Accuracy Experiment 2

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%

4.3. Classifying Image Transformations

This evaluation assesses the ability of VLMs to accurately identify specific image transformations from a predefined set of augmentations. Unlike the previous evaluations, which focused on pairwise comparisons, this assessment tests the models' direct classification capabilities across a comprehensive range of augmentation types. By examining how well the models can classify various image modifications, we can better understand their ability to recognize and categorize different types of visual changes.

4.3.1. 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

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

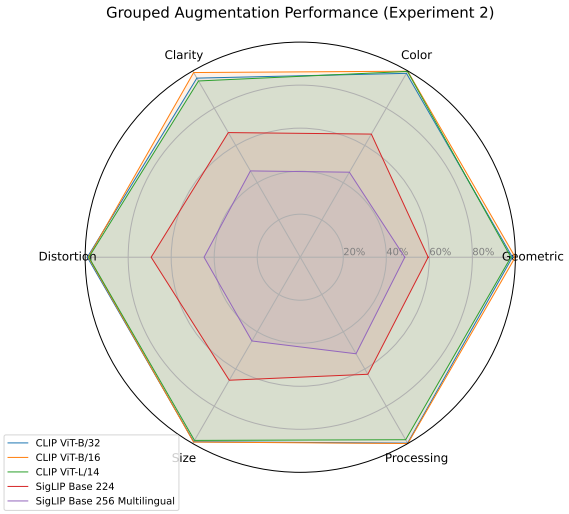

Figure 8. Per Augmentation Accuracy Experiment 2

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:

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

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.

4.3.2. Results

This experiment shows Vision Language Understanding of Augmentations where can a model associate itself with the correct Augmentation. Figure 9 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.

5. Impact on Downstream task

With the rise of AI in Image/Video Editing[31], this study reveals an important lack of understanding of the image level in vision language models. These models, predominantly built on CLIP[21] as their backbone architecture, form the foundation of numerous downstream tasks such as Image Generation[23][33], Controlled Image Generation models [37][12][35][20], Image-to-Image Editing[19][4][17] and multiple other downstream tasks. CLIP-based architectures, which align visual and textual representations through contrastive learning, have demonstrated remarkable capabilities in understanding semantic content. However, our analysis exposes a critical limitation in their spatial understanding of images. Different types of image transformation are a basic tool in traditional image editing tools such as Photoshop[1], yet modern AI systems struggle with these operations. Table 4 shows examples of common AI Image editing models, Instruct Pix2Pix[2], Dall.E 3[7], and IP Adapter[34] with the prompt **"Rotate the input image 90 degrees"**. The results demonstrate that none of these CLIP-powered models was able to understand this basic instruction and failed to generate an image with the requested transformation applied. This fundamental limitation suggests that despite their impressive semantic capabilities, current CLIP-based models lack a comprehensive understanding of image structure and spatial relationships due to their invariant nature which comes at the cost of explicit spatial understanding. This paper motivates us to think about newer training paradigms for Vision Language Models that balance invariance with explicit transformation awareness, where models can have global context, understand images at a deeper structural level beyond just semantic content, and reason about spatial manipulations when required. Addressing these limitations will help

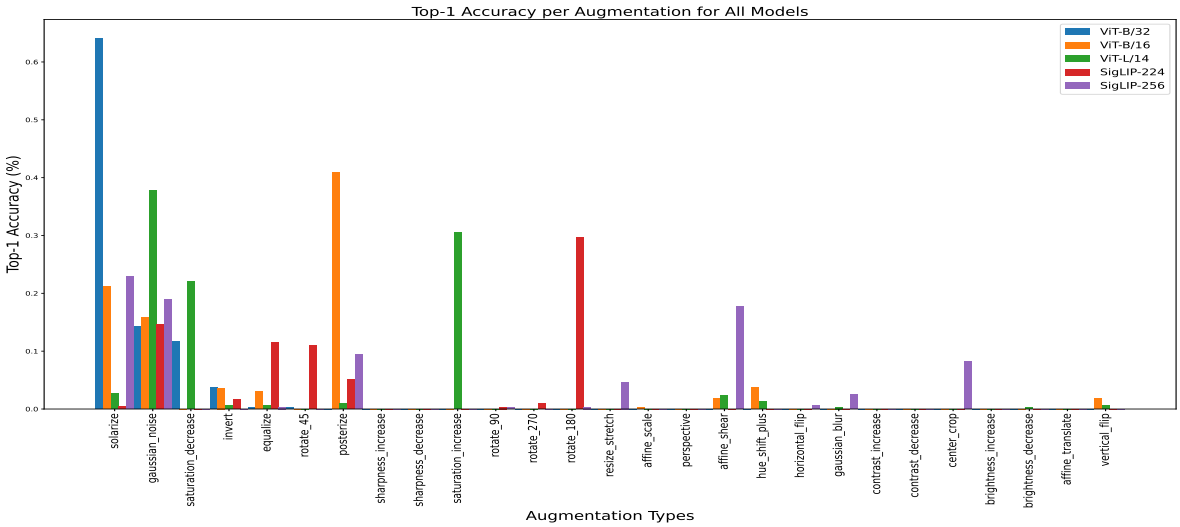

Figure 9. Top-1 Accuracy per Augmentation type for all models

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.

unlock newer capabilities in downstream tasks, potentially bridging the gap between AI-powered and traditional image editing tools.

References

[1] Adobe. Transform and rotate image selections and layers. https://helpx.adobe.com/photoshop/using/transforming-objects.html, 2024. Adobe Photoshop User Guide. 7

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 7

[3] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip, 2023. 2

[4] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models, 2023. 7

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022. 1

[6] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining, 2022. 2

[7] Gabriel Goh, James Betker, Li Jing, and Aditya Ramesh. Dall·e 3, 2024. 7

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[9] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2

[10] Clement Laroudie, Andrei Bursuc, Mai Lan Ha, and Gianni Franchi. Improving clip robustness with knowledge distillation and self-training, 2023. 2

[11] 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. 2

[12] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback, 2024. 7

[13] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning, 2023. 2

[14] Mushui Liu, Bozheng Li, and Yunlong Yu. Omniclip: Adapting clip for video recognition with spatial-temporal omni-scale feature learning, 2024. 2

[15] Yiming Ma, Victor Sanchez, and Tanaya Guha. Clip-ebc: Clip can count accurately through enhanced blockwise classification, 2024. 2

[16] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision, 2016. 3

[17] Naoki Matsunaga, Masato Ishii, Akio Hayakawa, Kenji Suzuki, and Takuya Narihira. Fine-grained image editing by pixel-wise guidance using diffusion models, 2023. 7

[18] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten, 2023. 2

[19] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023. 7

[20] 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. 7

[21] 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. 1, 7

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 7

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 1

[25] Dylan Sam, Devin Willmott, Joao D. Semedo, and J. Zico Kolter. Finetuning clip to reason about pairwise differences, 2024. 2

[26] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation, 2022. 2

[27] Cristian Sbrolli and Matteo Matteucci. No captions, no problem: Captionless 3d-clip alignment with hard negatives via clip knowledge and llms, 2024. 2

[28] Cristian Sbrolli, Paolo Cudrano, and Matteo Matteucci. Can shape-infused joint embeddings improve image-conditioned 3d diffusion?, 2024. 2

[29] 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. 2

[30] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension, 2022. 2

[31] 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. 7

[32] Weijie Tu, Weijian Deng, and Tom Gedeon. Toward a holistic evaluation of robustness in clip models, 2024. 2

[33] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition, 2021. 1, 7

[34] 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. 7

[35] Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems, 2024. 7

[36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 1

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 7

[38] 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. 1

[39] Zeliang Zhang, Zhuo Liu, Mingqian Feng, and Chenliang Xu. Can clip count stars? an empirical study on quantity bias in clip, 2024. 2