This ICCV 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.

Not all Views are Created Equal: Analyzing Viewpoint Instabilities in Vision Foundation Models

Mateusz Michalkiewicz
Rice University

Sheena Bai
Rice University

Mahsa Baktashmotlagh
The University of Queensland

Varun Jampani
Stability AI

Guha Balakrishnan
Rice University

Abstract

In this paper, we analyze the viewpoint stability of foundational models - specifically, their sensitivity to changes in viewpoint- and define instability as significant feature variations resulting from minor changes in viewing angle, leading to generalization gaps in 3D reasoning tasks. We investigate nine foundational models, focusing on their responses to viewpoint changes, including the often-overlooked accidental *viewpoints where specific camera orientations obscure an object's true 3D structure. Our methodology enables recognizing and classifying* accidental, stable *and* other *viewpoints using feature representations alone, without accessing the actual images at inference time. Our findings indicate that while foundation models consistently encode* accidental *viewpoints, they vary in their interpretation of* other *viewpoints due to inherent biases, at times leading to object misclassifications based on geometric resemblance. Through quantitative and qualitative evaluations on three downstream tasks - classification, VQA, and 3D reconstruction - we illustrate the impact of viewpoint instability and underscore the importance of feature robustness across diverse viewing conditions.*

1. Introduction

A common assumption in visual reasoning is that scene content in the 3D world is independent of camera viewpoint. While implicitly assumed in classical reconstruction and rendering algorithms [30], this assumption is not explicitly enforced in feature representation models now used in a range of computer vision applications. Popular foundation representation models, e.g., DINO [5] and CLIP [36], simply leverage large datasets and self-supervised training objectives to transform image pixels into a different feature space amenable to semantic interpretation. Although effective, this approach can introduce generalization gaps, particularly related to 3D reasoning [15, 45] and spatial context [51], resulting in poor performance when faced with unusual viewpoints or atypical test cases. Hence, while the success of foundation models are impressive, they lack certain fundamental reasoning capabilities, such as content-viewpoint disentanglement.

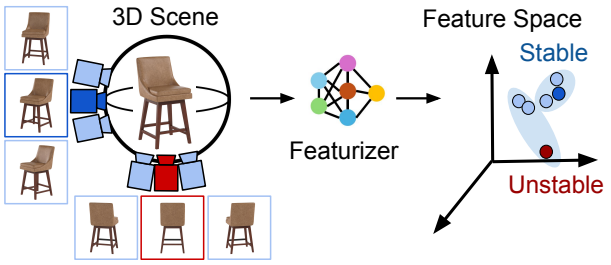


Figure 1. **Viewpoint stability of a featurizer.** The blue cameras represent small perturbations in camera space that result in correspondingly small changes in a feature space of a featurization function, indicating a stable viewpoint. In contrast, the red camera represents small perturbations in camera space that lead to large changes in feature space, signifying an unstable viewpoint.

In light of the above discussion, we take a deeper look in this study on how popular vision foundation models are affected by changes to viewpoint. In particular, we propose studying their viewpoint stability, or their rate of change of their output features with respect to changes in viewpoint. We define a model's feature output from a certain viewpoint of a scene as unstable if minor changes in the viewing angle lead to significant change in the returned features (see Fig. 1). A robust representation should ideally exhibit minimal changes in high-level scene content with varying viewpoints. Viewpoint stability may be caused by several fac-

tors. As explored in a recent study named OmniView [41], vision language models (VLMs) such as CLIP [37] and BLIP [24] struggle with out-of-distribution (OOD) viewpoints: object viewpoints that are rare in their training distributions. For instance, a mattress could be erroneously perceived as a closed laptop from a frontal viewpoint due to a lack of closed laptops in the training distribution (see Fig. 5).

However, instability is unavoidable in some cases, such as with *accidental* viewpoints, where a specific camera orientation can obscure the true 3D structure of an object [18, 46, 52]. In these cases, there is a fundamental geometric ambiguity that is not explicitly conveyed by featurizers. Understanding when a feature representation is unstable, and whether it coincides with known *accidental* viewpoints, can improve our grasp of deep learning models' limitations, suggest pathways for enhancing their robustness, and use them in more responsible ways for downstream applications. Existing works with foundation models such as OmniView notably miss critical discussions on *accidental* viewpoints - a topic generally neglected since the foundational studies of the late 20th century [52].

In this study, we introduce a methodology for recognizing viewpoint instabilities directly from the output feature representations of foundational models at inference time, without needing direct access to the input images. Specifically, our instability classifier is trained on labeled viewpoints from images, but once trained, it operates solely on individual feature vectors extracted from these models, enabling its application in privacy-sensitive scenarios where image data access is restricted.

We also propose a method to further classify unstable viewpoints into *accidental* versus *other* varieties from their features. Through extensive experiments on downstream tasks such as classification, VQA, and monocular 3D reconstruction, we illustrate the influence of viewpoint stability on model performance and identify specific conditions under which model robustness may falter due to viewpoint-related biases. Our findings first reveal that foundation models, regardless of their structural differences, exhibit a notable consistency in encoding *accidental* unstable viewpoints. However, their interpretations of *other* viewpoints are variable, which we attribute to inherent model biases. This underscores the need for robust feature representations to support stable performance across diverse viewing conditions.

2. Related Works

Viewpoint-invariant reasoning is a fundamental challenge in computer vision. Classical algorithms to address viewpoint variability depend on explicit geometric constraints or multiple view representations in the image space [2, 18, 19, 55]. Of particular relevance to our study are *accidental* viewpoints [18, 46, 52], or unstable viewpoints caused by the camera orientation obscuring an object's true 3D structure [52], which present significant barriers to robust recognition. In contrast to classical approaches, this work forms the first comprehensive study of modern foundation models for viewpoint instability, rather than from images cues directly.

2.1. Viewpoint Biases in Neural Networks

Deep neural network classifiers have been known to have viewpoint biases, i.e., performance gaps depending on viewpoint. Popular CNNs used for classification tasks exhibit a noticeable drop in performance for images taken from underrepresented viewpoints [1, 11, 28]. Other studies [13, 58] investigated the biases of models towards canonical viewpoints and introduced datasets and techniques for evaluating viewpoint robustness, focusing on uncommon (but not explicitly *accidental*) viewpoints. Specifically, ViewFool [13] proposed a technique to generate adversarial viewpoints that push the models towards poor performance, effectively revealing their weaknesses in handling OOD viewpoints. Another study demonstrates predicting OOD viewpoint instabilities of deep networks and achieving viewpoint-invariant visual recognition through adversarial training [40]. Several other studies also propose network training paradigms [6, 21, 22, 31, 44, 53] to encourage viewpoint-agnostic learning representations for specific applications.

Though inspired by these studies, our work differs in several ways. First, we focus our attention on a range of modern foundation models rather than traditional classifiers. Second, we aim to predict viewpoint instability directly from features outputted by networks, which offers potential means to flag uncertain responses from these algorithms. Third, we provide a distinct emphasis on understanding how networks react to *accidental* viewpoints.

2.2. Robustness Analysis of Foundation Models

Foundation models, such as CLIP [37], DINOv2 [33], and Llama [48], are now widely used in computer vision due to their state-of-the-art generalization capabilities. As a result, their robustness has also been under scrutiny in a range of contexts, including adversarial attacks [42], dataset content [32], compositional reasoning [14], and visual anomalies [49]. Several recent works analyze foundation model features with respect to 3D awareness [15, 29, 43, 54], and Omniview [41] found that the robustness of vision-language models (VLMs) varies significantly based on viewpoints in their training data.

While Omniview is the closest in spirit to our analysis, they did not explicitly study unstable viewpoints (and the further exploration of *accidental* cases), and only evaluated two foundation models (CLIP and BLIP). In contrast,

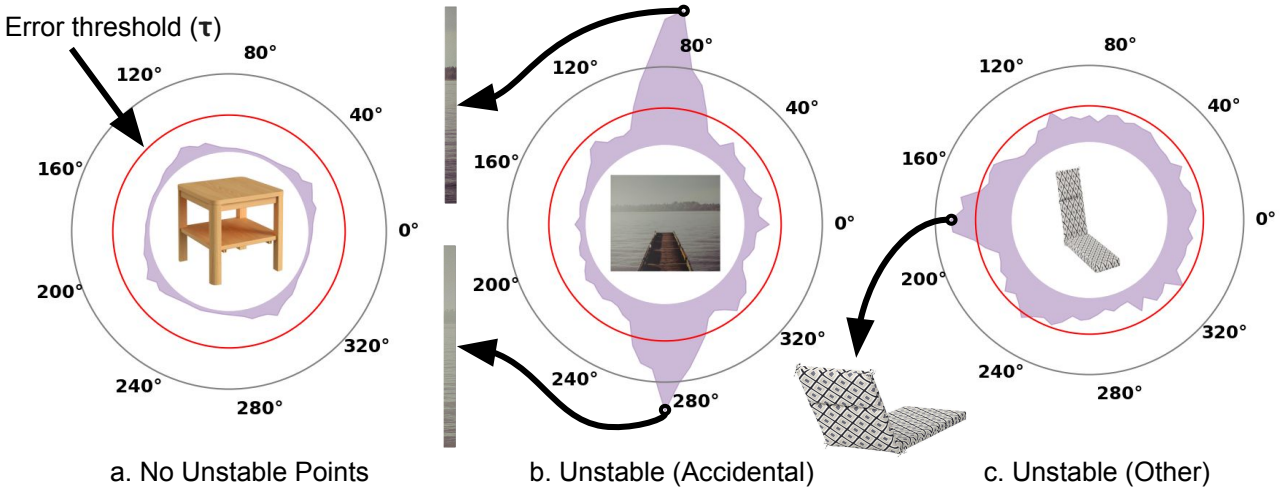


Figure 2. **Instability scores of CLIP [37] features with respect to viewpoints for three scenes in the ABO [9] dataset.** (Left) A wooden side table with no instabilities. (Center) a framed painting with two *accidental* viewpoints, where a side view obscures its structure, showing only the frame's edge and making its identity unclear. (Right) High-back cushion with an *other* viewpoint.

we systematically evaluate nine foundation models. Therefore, our work highlights fundamental gaps in robustness that have not been addressed in prior literature.

3. Experimental Analysis

We present various analyses of foundation models with respect to viewpoint stability in this section. First, we describe the data and models we use in Sec. 3.1. Second, we describe our method for quantifying viewpoint instability types in Sec. 3.2. Third, we explore predicting viewpoint instabilities from featurizer outputs alone in Sec. 3.3. Finally, we demonstrate that viewpoint instabilities have significant consequences in various downstream applications using these foundation models.

3.1. Datasets and Featurizers

We used two datasets for our analysis. The first is **Amazon-Berkeley Objects (ABO)** [9], consisting of product catalog images, metadata, and artist-created 3D models of real household objects. We specifically used the "Spins" subset of ABO, consisting of 8,200 objects distributed across 103 categories, e.g., sofas, wall art, pillows, dressers, and lamps. Each object has 72 image renderings captured at 5-degree azimuth intervals with 360-degree coverage, with randomized elevation angles and varying lighting conditions. In total, this yields 590,400 unique viewpoints across all objects with which we conduct our analysis. Our second dataset is **Common Objects in 3D (CO3D)** [38], consisting of 1.5 million real-world multi-view images for roughly 19,000 objects coming from 50 MS COCO [25] common object categories. The images are collected by Amazon Mechanical Turk workers, who are instructed to take a 360-degree video of an object from a specified category. For our analyses, we utilized nine different foundation model featurizers: CLIP [37], ConvNeXt [27], Deit III [47], DINO [4], DINOv2 [33], DreamSim [20], MAE [23], SAM [7, 17], and SigLip [56]. Detailed descriptions of these featurizers are provided in the supplementary material.

3.2. Viewpoint Instabilities in Foundation Models and Their Types

Given an image $I_v \in \mathbb{R}^{H\times W}$ of a 3D scene taken at view v (3D position and angle), and an arbitrary model $f(\cdot) : \mathbb{R}^{H\times W} \to \mathbb{R}^d$ that embeds the image into a d-dimensional vector, we aim to measure the model's viewpoint stability at v, i.e., how the model's output features $f(I_v)$ respond to a small perturbation in v. If such a perturbation causes a large change, we consider these features unstable.

We define two distance functions: $d_v(\cdot,\cdot)$ which measures the distance between two orientations v_1 and v_2, and $d_f(\cdot,\cdot)$ which measures the distance between two feature vectors $f(v_1)$ and $f(v_2)$. In our experiments, we let d_v encode Euclidean distance based on 3D position and angle, and let d_f encode cosine distance (the most commonly used distance for features of modern foundation models). Further, let $N_v(v_i, r)$ denote the set of all nearest neighbors to v_i such that if $u \in N(v_i, r)$, $d_v(v_i, u) \leq r$. We then define the instability score at v_i:

$$\text{ins}_f(v_i) = \frac{1}{|N_v(v_i,r)|} \sum_{u\in N(v_i,r)} d_f(f(v_i), f(u)). \quad (1)$$

We then label v_i as unstable if $\text{ins}_f(v_i) > \tau$, for some fixed threshold τ.

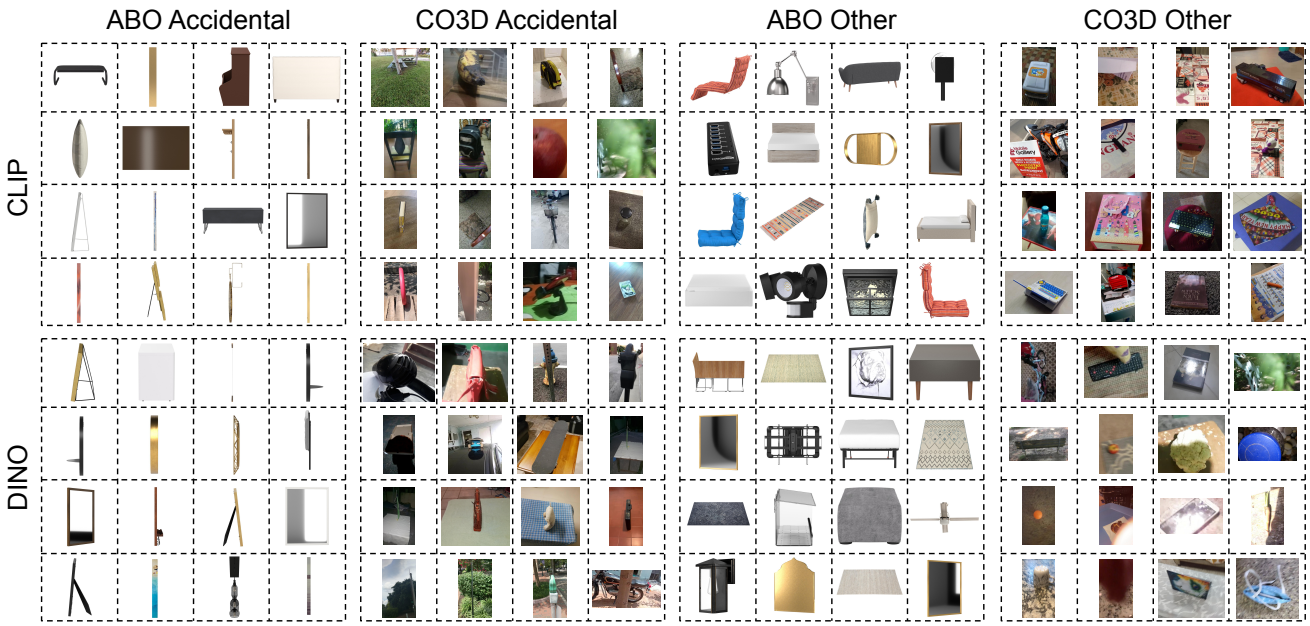


Figure 3. **Examples of *accidental* and *other* viewpoints using CLIP and DINO embeddings across the ABO and CO3D datasets.** *Accidental* views obscure object's true structure, while *other* views present uncommon orientations rarely or never seen during training. In the CO3D dataset, additional sources of instability include occlusions, image blur, and objects displayed upside down.

Fig. 2 illustrates stability scores and unstable viewpoints detected for three examples from the Amazon-Berkeley Objects (ABO) dataset [9]. Some objects, such as the side table (Fig. 2-left), have no unstable viewpoints. Others, such as the framed painting (Fig. 2-center), have *accidental* viewpoints where the camera angle reveals only the edge of the frame, obscuring the painting itself. And yet other objects, such as the cushion (Fig. 2-right) have instabilities which do not appear *accidental*, but rather are caused by other biases or artifacts, which we call *other*.

3.3. Predicting Instabilities from Features

Classifying Stable and Unstable Points. In most real applications (e.g. classification from a single image or monocular 3D reconstruction), we will not have access to a dense set of neighboring images to a given view. In these cases, we will be unable to apply Eq. 1 to determine whether $f(v_i)$ are stable features or not. Hence, we explore predicting instability from $f(v_i)$ directly. To do so, we trained SVM binary classifiers [10] with RBF kernels for all featurizers on both datasets to predict view instability, using Eq. 1 to produce ground truth binary labels. For each (featurizer,dataset) combination, we set the threshold τ to the 97th percentile of instability scores computed across all viewpoints. We split the datasets into 80/20 train/test splits along scenes, such that images from the same scene do not span both groups. To reduce redundant examples, we applied non-maximal suppression to only choose maximally unstable points in local neighborhoods. Table 1 presents classification results. We obtained accuracies of [81-96]% across featurizers on ABO, and [71-77]% on CO3D, demonstrating the surprising finding that view stability is largely predictable from features alone. MAE is generally the hardest feature space in which to separate its stable and unstable views.

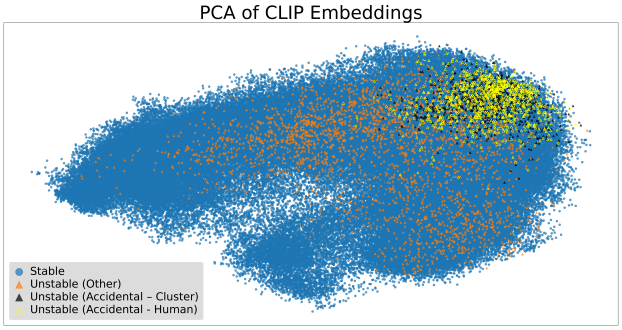


Figure 4. **PCA visualization of stable (blue circles), unstable other (orange triangles), clustered accidental (black triangles), and human-annotated accidental viewpoints (empty yellow triangles).** Automatic clustering shows strong overlap with human labels of accidental viewpoints. Please refer to sample images from the *accidental* and *other* in Fig. 3.

Predicting *Accidental* and *Other* Unstable Subtypes. We

Table 1. **Classification results for separating stable and unstable viewpoints in various feature spaces.** We obtain stable/unstable labels per viewpoint using Eq. 1. We obtain surprisingly high classification scores across all featurizers, with better performance on ABO due to less variabilities (backgrounds, lighting, artifacts). Rows are ordered using the Borda count ranking method [50] across both datasets.

Model	ABO	CO3D
DINO	94.08%	77.08%
DreamSim	96.60%	73.37%
ConvNeXt	90.73%	76.18%
DINOv2	92.79%	73.78%
DeiT III	88.05%	76.85%
SigLip	87.97%	74.65%
SAM	89.53%	72.33%
CLIP	88.50%	72.37%
MAE	81.68%	71.52%

further hypothesized that unstable view subtypes of *accidental* and *other* may also be predictable from features alone. Fig. 4 shows a 2D projection of CLIP features on the ABO dataset using Principal Component Analysis (PCA), with stable points displayed in blue, and two clusters of unstable points obtained with k-means shown in orange and black. Upon visual inspection of samples, we find the black cluster largely corresponds to *accidental* views, while the orange are *other* views. Interestingly, the black cluster is also "tighter," which agrees intuitively with the idea that these views share similar collapsed 1D/2D structures. We therefore explore separating *accidental* from *other* unstable viewpoints in a feature space by fitting $k = 2$ clusters on unstable points, and assigning points that fall in the cluster with highest silhouette score [39] as *accidental*.

Using this approach, we obtained *accidental* and *other* clusters for all featurizers, and present random samples from the CLIP and Dino models in Fig. 3. The CO3D dataset, which includes more complex backgrounds and in-the-wild images, also reveals *accidental* viewpoints that arise when objects are temporarily occluded. For CO3D *other* views, we often observe objects that appear upside down or temporarily blurry, likely due to changes in focus along the camera's trajectory. We found similar trends for all *other* featurizers (see Supplementary).

We next analyzed whether images from the predicted *accidental* clusters agree with human perception of *accidental* viewpoints. Using 6 human annotators, each labeled 2,000 images (including both stable and unstable views) as either *accidental* or non-*accidental*. Annotators received a clear definition—*accidental* viewpoints are those exhibiting ambiguous or collapsed 3D structure—along with 20 exemplar images (10 *accidental* and 10 non-*accidental*) prior to labeling. We observed an approximate 79% overlap between the human labels and the automatically clustered *accidental* views. Notably, only about 10% of stable views were labeled as *accidental*, whereas approximately 90% of unstable views matched the *accidental* labels, confirming that *accidental* views are rarely stable. Moreover, when merging all unstable views (both *accidental* and *other*) into a single *accidental* category to account for clustering ambiguities, the overlap increases to 85%. Under this categorization, DreamSim achieves the highest agreement (96.5%), while CLIP exhibits the lowest agreement (74.9%). This indicates that general instability is far easier to predict in feature space compared to unstable subtypes and that methods fine-tuned for alignment with human visual similarity - as in DreamSim - can yield stronger agreement with human annotations.

Finding *accidental* views is crucial for using featurizers responsibly in downstream applications, as these views inherently pose challenges that may never be fully resolved. Recognizing when an *accidental* viewpoint occurs allows models to account for these limitations. We compute the probability of a viewpoint being accidental as: $P(\text{accidental}) = P(\text{unstable}) \cdot P(\text{accidental} \mid \text{unstable})$, where the probabilities are obtained using Platt scaling [35], ensuring interpretable confidence measures. Indeed, applying this approach, we obtain an average confidence of approximately $\sim 90\%$ when predicting *accidental* viewpoints across the nine models. In contrast, CLIP's standard confidence estimation - computing softmax over a set of label scores - exhibits near-zero correlation (0.02) with *accidental* viewpoint labels. Notably, directly clustering *all* views into *accidental* vs. non-*accidental* results in $\sim 65\%$ false positives, whereas filtering out stable views beforehand reduces this to $\sim 2.5\%$.

Finally, we assessed the level of agreement across featurizers on identifying *accidental* and *other* viewpoints. Using Intersection over Union (IoU) to measure overlap in viewpoint classification, we observed a high level of agreement for stable viewpoints (mean IoU of 0.99) and moderate agreement for *accidental* viewpoints (0.52), indicating that these viewpoints are generally identified similarly across featurizers. However, agreement on *other* viewpoints was markedly low, with a mean IoU of only 0.12. We provide pairwise IoU heatmaps for each label category (stable, *accidental*, and *other*) across all 9 featurizers in Supplementary.

3.4. Downstream Applications

The above analyses highlight viewpoint instabilities in the feature spaces of popular foundation models. To understand the practical impact of these instabilities on real-world applications, we next investigated three downstream tasks: classification (zero-shot and linear probing), visual question answering (VQA), and monocular 3D reconstruction.

Zero-Shot Classification. In this experiment, we evalu-

Table 2. **Zero-shot classification accuracy of CLIP for stable and unstable viewpoints at different Top-K values on the ABO and CO3D datasets.** For the ABO dataset, Acc@5 is marked in bold as the most suitable measure due to overlapping object categories, while for the CO3D dataset, Acc@1 is marked in bold given the distinct categories. The model shows a decline in performance for *other* and *accidental* viewpoints across both datasets.

ABO Dataset			
	Stable	Other	Accidental
Acc @1	40.23	22.47	0.84
Acc @5	**68.74**	42.56	3.00
Acc @10	79.97	56.05	4.85
CO3D Dataset			
	Stable	Other	Accidental
Acc @1	**78.58**	68.45	48.36
Acc @5	93.77	90.45	75.49

ated CLIP's zero-shot classification performance on both ABO and CO3D datasets. Leveraging CLIP's shared feature space, we encode ground truth (GT) category labels as text embeddings and compute feature embeddings for each image, assigning the label with the minimum cosine distance to each image embedding. We report classification performance using accuracy@k, which reflects the percentage of cases where the GT label is among the top k closest predicted labels, with k chosen from $\{1, 5, 10\}$. The results in Table 2 show a sharp drop in accuracy for unstable viewpoints relative to stable ones, particularly for *other*, and even more so for *accidental*. This illustrates the model's challenge in maintaining reliable classification under less common or ambiguous viewing conditions. The effect is particularly pronounced in ABO, which, being semi-synthetic, lacks shadows and background distractors. This absence of neighboring scene context makes *accidental* viewpoints more pronounced than in CO3D, where surrounding elements provide additional cues to aid classification.

We further investigate whether recently proposed viewpoint-invariant methods can effectively mitigate viewpoint instabilities. Specifically, we analyzed two representative methods: OmniView-tuned OpenCLIP (OVT-OpenCLIP) [41] and CLIP fine-tuned for multiview equivariance (CLIP-FT) [54]. Table 3 summarizes their zero-shot classification performances relative to standard CLIP. OmniView significantly improves performance on *unstable other* viewpoints, demonstrating its effective regularization for viewpoint invariance, while CLIP-FT only marginally outperforms the baseline. Both methods show limited improvements on *accidental* viewpoints, underscoring the inherent geometric ambiguity in these cases that training alone cannot fully resolve.

Linear Probe Classification. While CLIP's shared image-

Table 3. **Zero-shot classification comparison between CLIP and viewpoint-invariant OmniView (OVT-OpenCLIP) and CLIP fine-tuned for multiview equivariance (CLIP-FT).** For the ABO dataset, we report Acc@5 due to overlapping categories, and Acc@1 for the CO3D dataset with distinct categories. OmniView notably improves accuracy for *unstable other* viewpoints, while CLIP-FT provides minimal gains. Both methods only marginally enhance accuracy for inherently ambiguous *accidental* viewpoints.

ABO Dataset (Acc@5)			
Model	**Stable**	**Other**	**Accidental**
CLIP [37]	68.74	42.56	3.00
OVT-OpenCLIP [41]	69.98	56.13	3.34
CLIP-FT [54]	69.03	43.17	3.12
CO3D Dataset (Acc@1)			
Model	**Stable**	**Other**	**Accidental**
CLIP [37]	78.58	68.45	48.36
OVT-OpenCLIP [41]	79.01	77.92	50.12
CLIP-FT [54]	79.21	69.52	49.81

text space allows for zero-shot classification, most other foundation models lack this capability. Linear probing – or simply training an extra linear prediction layer using features outputted by a backbone network – is a common strategy used to evaluate the usefulness of a generic feature space. We trained a single linear layer on top of each model's embeddings for 100 epochs, using the Adam optimizer [12], and a learning rate of 1e-3. We applied an 80-20 train-test split, ensuring no overlap of objects between the sets to avoid data leakage. The results, shown in Table 4, indicate a substantial accuracy decline for unstable viewpoints, underscoring the limitations of current models in handling viewpoint variability and highlighting the need for more viewpoint-robust representations.

Visual Question Answering (VQA). We next conducted a VQA experiment using LLaVA-1.5 [26], which leverages CLIP as its backbone, to evaluate how well it generates descriptive captions for images across stable, *accidental*, and *other* viewpoints. Prompted with a generic question, *What is the content of the image?*, we assessed the model's accuracy in describing the content under different viewing conditions. Qualitative examples in Fig. 5 show that for stable viewpoints, LLaVA-1.5 generates accurate captions, whereas for *accidental* and *other* viewpoints, the model often fails, misidentifying objects or adding irrelevant details. To quantify performance, we compared generated captions to ground truth (GT) captions. GT captions were created using LLaVA-1.5 with an informative prompt, *What's the content of the image knowing that the main object is* $\{label\}$, where $\{$label$\}$ represents the known category from the ABO dataset. We measured the percentage of generated captions achieving an F1 score above 0.5 (F1@0.5) using BERTScore [57] to assess similarity to GT captions.

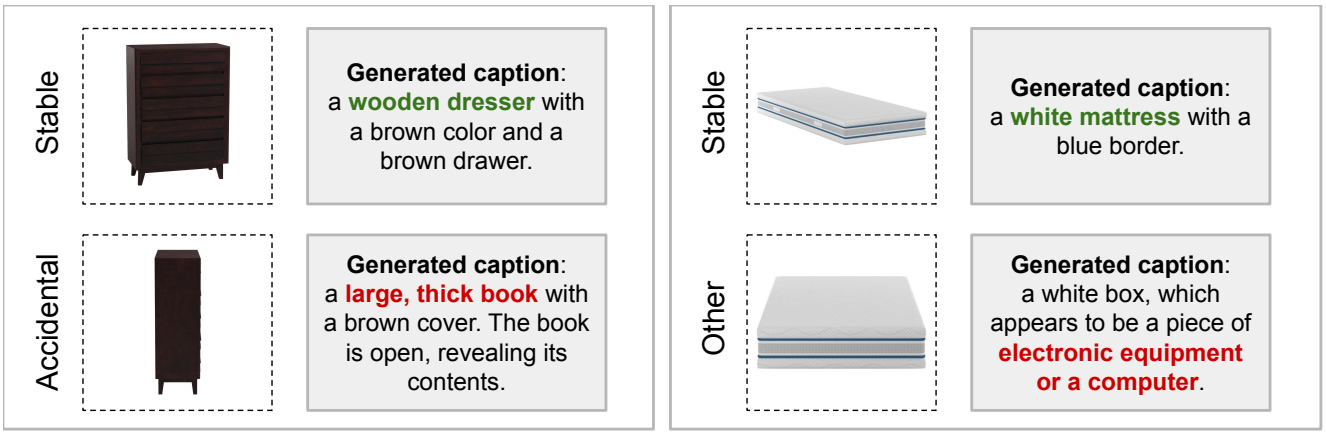


Figure 5. **Examples of generated captions for stable, *other*, and *accidental* viewpoints using LLaVA-1.5 [26] (CLIP [37] backbone).** Captions for stable viewpoints are factual, while those for *accidental* and *other* viewpoints often contain inaccuracies. Here, an accidental viewpoint causes LlaVA to describe a dresser as a book, and an *other* viewpoint causes a misinterpretation of a mattress as a computer.

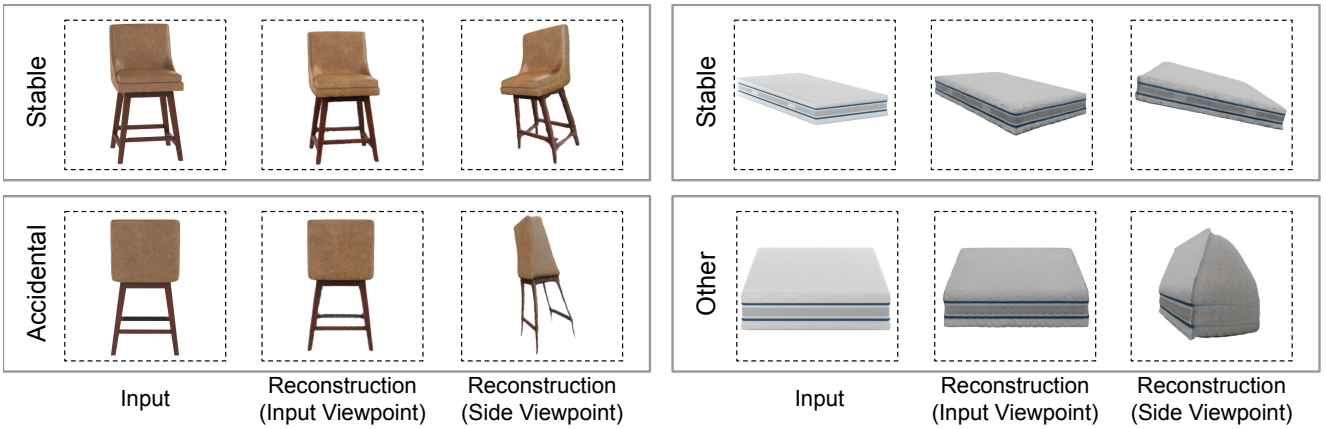


Figure 6. **Single-view reconstruction results for stable and unstable viewpoints from the ABO dataset, using Stable Fast 3D [3] with DINOv2 [33] as the image featurizer.** Stable viewpoints yield reconstructions with accurate geometry, while *accidental* and *other* viewpoints lead to distortions.

Results in Tab. 5 show significantly higher F1@0.5 for stable viewpoints compared to *accidental*, underscoring the challenges that viewpoint instability poses for accurate content description.

Monocular 3D Reconstruction. To examine the effects of viewpoint instability on 3D reconstruction, we conducted experiments using Stable Fast3D [3], a state-of-the-art monocular 3D reconstruction model that uses DINOv2 as its image featurizer. We generated single-view reconstructions for stable, *accidental*, and *other* viewpoints. After reconstruction, we aligned the predicted shapes with ground truth (GT) models using RANSAC [16] followed by ICP [8] to ensure optimal alignment. Figures 6 illustrate the differences in reconstruction quality across viewpoint types. Stable viewpoints yield accurate reconstructions with well-preserved geometry.

In contrast, *accidental* viewpoints—where depth perception is almost absent due to insufficient depth cues—cause the model to incorrectly infer the 3D shape, resulting in collapsed or distorted reconstructions. *Other* viewpoints similarly degrade performance, as the atypical angles can cause the model to interpret the object as belonging to an entirely different category, leading to substantial inaccuracies in the reconstruction. Quantitative results in Tab. 5 confirm these observations, with stable viewpoints achieving higher F1 scores relative to *other* and *accidental* viewpoints. These findings underscore the impact of viewpoint instability on reconstruction accuracy and highlight the need for models capable of adapting to challenging and atypical viewing conditions.

Table 4. **Classification accuracy using linear probe for stable and unstable viewpoints on the ABO and CO3D datasets.** The results show a significant decline in accuracy for both *other* and *accidental* viewpoints across models.

ABO Dataset			
	Stable	Other	Accidental
CLIP	89.62%	73.97%	63.19%
ConvNeXt	99.08%	88.35%	79.29%
Deit III	91.43%	73.00%	63.80%
DINO	92.37%	78.41%	69.35%
DINOv2	95.26%	76.39%	72.18%
DreamSim	97.96%	86.93%	79.89%
MAE	93.97%	89.72%	70.93%
SAM	91.79%	75.49%	63.77%
SigLip	95.95%	81.82%	73.77%
CO3D Dataset			
	Stable	Other	Accidental
CLIP	98.84%	95.22%	77.46%
ConvNeXt	99.74%	96.51%	87.34%
Deit III	96.08%	74.57%	73.45%
DINO	96.71%	83.41%	70.19%
DINOv2	98.94%	84.01%	77.49%
DreamSim	98.75%	93.12%	76.96%
MAE	93.97%	90.26%	66.49%
SAM	93.92%	73.81%	64.12%
SigLip	99.85%	99.10%	81.21%

Table 5. **F1 scores (F1@0.5) for 3D reconstruction and VQA tasks across stable, *other*, and *accidental* viewpoints on the ABO dataset.** For VQA, the F1 score is computed using BERTScore between generated and GT captions. Stable viewpoints achieve higher scores compared to *accidental* viewpoints, indicating that viewpoint instability affects both reconstruction and caption generation accuracy.

Task	**Stable**	**Other**	**Accidental**
3D Reconstruction	82.86%	65.00%	36.67%
VQA	56.42%	50.95%	29.86%

4. Discussion and Conclusion

In this study, we investigated the viewpoint stability of popular vision foundation models and introduced methodologies for recognizing and classifying instabilities directly from their feature representations. Our experiments first demonstrate that unstable viewpoints are to some degree separable from stable viewpoints in the feature spaces of all the studied foundation models. Prior works demonstrate that styles [34] and even 3D concepts [15] are separable in feature space. While viewpoint stability itself is primarily a geometric property rather than a purely semantic one, we recognize that unstable—particularly accidental—viewpoints often correspond to semantic anomalies, as objects appear as ambiguous or non-existent configurations from these angles. Thus, our findings suggest a nuanced connection between geometric ambiguity and semantic anomalies, providing deeper insight into the visual features encoded by foundation models.

All models exhibited some level of view instability. We found that CLIP, possibly the most popular of all the foundation models we studied, was less stable compared to alternatives like ConvNeXt and DreamSim, particularly in relation to *accidental* viewpoints. Interestingly, DreamSim uses an ensemble of CLIP, OpenCLIP, and DINOv2, which may explain its relative stability compared to the other algorithms. Foundation models generally agreed on *accidental* viewpoints, but varied on what they considered *other*. This is reasonable since *accidental* views are directly defined by the scene geometry and camera angle, while *other* views are also affected by the datasets, architectures, and training paradigms of each network.

Our analyses also demonstrate that viewpoint instabilities of featurizers lead to tangible dropoffs in performance of downstream models that use them for various applications. Example tasks we explored include classification variants, VQA, and 3D reconstruction. We used two datasets in our analyses: ABO and CO3D. Due to its synthetic nature, ABO yielded far cleaner separation of stable and unstable points in feature spaces. In contrast, due to backgrounds, occlusions, lighting variations, and camera acquisition factors, CO3D yielded less clear separability. And yet, even in these natural conditions, we observed significant adverse effects on downstream tasks for detected unstable views.

These analyses raise important questions on how to prevent the unsafe use of foundation models with respect to viewpoint instabilities. One natural follow-up to the results of this study is to equip the foundation models with our proposed confidence measures along with their features. Downstream models can use such confidence measures to signal when their resulting predictions might be unreliable, for instance in VQA applications, rather than feigning certainty. While unstable-*other* viewpoints—essentially out-of-distribution views—can potentially be mitigated by viewpoint-invariant methods such as OmniView [41], the inherent geometric ambiguity of *accidental* viewpoints presents a fundamental challenge that cannot be fully resolved by training alone. In these scenarios, where the geometry effectively collapses, instability is inevitable. Thus, rather than pursuing the unrealistic goal of entirely eliminating instability, we propose explicitly identifying *accidental* viewpoints, allowing downstream tasks to proactively recognize uncertainty, request additional views, or abstain from unreliable predictions.

Acknowledgement Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0076. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4845–4854, 2019. 2

[2] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987. 2

[3] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*, 2024. 7

[4] 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 IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[5] 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 IEEE/CVF International Conference on Computer Vision*, 2021. 1

[6] Wei-Da Chen and Shan-Hung Brandon Wu. Cnn^{2}: Viewpoint generalization via a binocular vision. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[7] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 3

[8] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 7

[9] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 3, 4

[10] Corinna Cortes. Support-vector networks. *Machine Learning*, 1995. 4

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2

[12] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014. 6

[13] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *Advances in Neural Information Processing Systems*, 35: 36789–36803, 2022. 2

[14] Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150, 2023. 2

[15] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 1, 2, 8

[16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 7

[17] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 3

[18] William T Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545, 1994. 2

[19] William T Freeman. Exploiting the generic viewpoint assumption. *International Journal of Computer Vision*, 20(3): 243–261, 1996. 2

[20] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 3

[21] Abdullah Hamdi, Faisal AlZahrani, Silvio Giancola, and Bernard Ghanem. Mvtn: Learning multi-view transformations for 3d understanding. *International Journal of Computer Vision*, pages 1–30, 2024. 2

[22] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 160–177. Springer, 2016. 2

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, pages 16000–16009, 2022. 3

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 6, 7

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3

[28] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how convolutional neural networks generalize to out-of-distribution category–viewpoint combinations. *Nature Machine Intelligence*, 4(2):146–153, 2022. 2

[29] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022. 2

[30] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3), 2006. 1

[31] Jiyoun Moon, Hanjun Kim, and Beomhee Lee. View-point invariant 3d classification for mobile robots using a convolutional neural network. *International Journal of Control, Automation and Systems*, 16(6):2888–2895, 2018. 2

[32] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip. *Advances in Neural Information Processing Systems*, 35:21455–21469, 2022. 2

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 7

[34] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. 8

[35] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 5

[36] 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. In *International Conference on Machine Learning*. PMLR, 2021. 1

[37] 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. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6, 7

[38] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 3

[39] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 5

[40] Shouwei Ruan, Yinpeng Dong, Hang Su, Jianteng Peng, Ning Chen, and Xingxing Wei. Towards viewpoint-invariant visual recognition via adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4709–4719, 2023. 2

[41] Shouwei Ruan, Yinpeng Dong, Hanqing Liu, Yao Huang, Hang Su, and Xingxing Wei. Omniview-tuning: Boosting viewpoint invariance of vision-language pre-training models. *arXiv preprint arXiv:2404.12139*, 2024. 2, 6, 8

[42] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023. 2

[43] 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. *arXiv preprint arXiv:2402.12336*, 2024. 2

[44] Jinghuan Shang, Srijan Das, and Michael Ryoo. Learning viewpoint-agnostic visual representations by recovering tokens in 3d space. *Advances in Neural Information Processing Systems*, 35:31031–31044, 2022. 2

[45] Jie Tang, Jing Zhang, Liming Yao, Juanzi Li, Li Zhang, and Zhong Su. From big data to knowledge in ai. *Frontiers of Information Technology & Electronic Engineering*, 18(1), 2017. 1

[46] Antonio Torralba and William T Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2012. 2

[47] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022. 3

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al.

Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[49] Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (clip). *Advances in Neural Information Processing Systems*, 36, 2024. 2

[50] Merijn Van Erp and Lambert Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. In *7th International Workshop on frontiers in handwriting recognition*, pages 443–452. International Unipen Foundation, 2000. 5

[51] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3647, 2024. 1

[52] Andrew P Witkin and Jay M Tenenbaum. On the role of structure in vision. In *Human and machine vision*. Elsevier, 1983. 2

[53] Ziyan Wu, Yang Li, and Richard J Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE transactions on pattern analysis and machine intelligence*, 37(5): 1095–1108, 2014. 2

[54] Yang You, Yixin Li, Congyue Deng, Yue Wang, and Leonidas Guibas. Multiview equivariance improves 3d correspondence understanding with minimal feature finetuning. *arXiv preprint arXiv:2411.19458*, 2024. 2, 6

[55] Richard Zemel and Geoffrey E Hinton. Discovering viewpoint-invariant relationships that characterize objects. *Advances in neural information processing systems*, 3, 1990. 2

[56] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 3

[57] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 6

[58] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. In *ICML 2022 Shift Happens Workshop*, 2022. 2