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

Supplementary Material

Mateusz Michalkiewicz
Rice University

Sheena Bai
Rice University

Mahsa Baktashmotlagh
The University of Queensland

Varun Jampani
Stability AI

Guha Balakrishnan
Rice University

This supplementary material includes additional experimental results that complement the findings presented in the main paper. Specifically, we provide:

- Detailed descriptions of the nine featurizers utilized in our analyses, highlighting differences in training paradigms, architectures, and datasets.
- Additional results on label overlap across nine featurizers for stable, *accidental*, and *other* viewpoints, evaluated using pairwise Intersection over Union (IoU) heatmaps.
- Cluster samples from *accidental* and *other* viewpoints for seven additional featurizers not included in the main paper. These examples highlight consistent trends and instabilities in viewpoint classification.
- Expanded qualitative examples of Visual Question Answering (VQA) using LLaVA-1.5, which demonstrate the model's strengths and limitations across stable, *accidental*, and *other* viewpoints.
- Additional single-view monocular 3D reconstruction results. These examples further illustrate the impact of viewpoint instability on reconstruction accuracy, emphasizing the challenges posed by *accidental* and *other* viewpoints.

1. Featurizer Descriptions

We conduct our analysis using nine different deep network featurizers with different training data, training paradigms, and application areas:

1. CLIP [12]: trained using contrastive learning on 400 million (image, text) pairs scraped from the Internet. We used the default public ViT-B/32 architecture. **2. DINO** [2]: self-supervised framework based on student-teacher knowledge distillation, pretrained on ImageNet (without labels) [13]. We used the default public ViT variant with 85.8M parameters. **3. DINOv2** [11]: an improved version of DINO that is trained on a curated dataset LVD-142M [11] with 142M images. We used the public ViT variant with 88.6M parameters. **4. ConvNeXt** [10]: CNN family achieving comparable performance to vision transformers. We used a public model fine-tuned on ImageNet-1k [5] with 846.5M parameters. **5. Deit III** [14]: an approach for training vision transformers that includes techniques such as using 3-Augment for data augmentation, simple random cropping, and low resolutions. We used the image classification variant with 86.6M paramters trained on Imagenet-1k [5]. **6. DreamSim** [7]: used to compute perceptual similarity between a pair of images, trained on the NIGHTS [7] dataset with 60k images with human perceptual evaluations. We used the standard configuration with ViT-B/16 backbone and 266 million parameters. **7. Masked Autoencoder (MAE)** [8]: self-supervised model that reconstructs images from partially masked inputs, pre-trained on ImageNet-1K. We used the ViT variant with 630.8 million parameters. **8. Sharpness-Aware Minimization (SAM)** [3, 6]: an optimization method that improves generalization and performance of large vision transformers. We used a ViT model with 88.2 million parameters. parameters. **9. SigLip** [15]: CLIP-based image-text model utilizing sigmoid loss, pre-trained on the WebLI dataset [4], offering enhanced 0-shot accuracy. We used the public model with 878 million parameters.

2. Featurizer Agreement on Viewpoint Labels

We present additional results on the level of agreement, measured using Intersection over Union (IoU), across the nine featurizers in identifying stable, *accidental*, and *other* viewpoints. Specifically, we provide IoU heatmaps for each category across all nine featurizers (see Fig. 1). The results indicate consistent trends: all featurizers exhibit strong agreement on stable viewpoints (IoU scores ranging from 0.98 to 1.00), moderate agreement on *accidental* viewpoints (IoU scores between 0.34 and 0.83), and very little agreement on *other* viewpoints (IoU scores ranging from 0.02 to 0.28). For *accidental* viewpoints, the moderate agreement suggests that these challenging viewpoints are generally identified similarly across featurizers, demonstrating some commonality in how models handle such instances. In contrast, the lack of agreement for *other* viewpoints, with many pairs of featurizers showing no overlap (IoU scores under 0.1), highlights the model-specific nature of viewpoint instability. This underscores the challenge of constructing datasets or methods that generalize viewpoint invariance across models. The results suggest that while *accidental* viewpoint invariance may benefit from shared strategies across models, *other* viewpoint invariance likely needs to be tailored to specific featurizers rather than assuming a universal approach.

3. Cluster Samples Across Additional Featurizers

We presents samples from *accidental* and *other* clusters for the seven featurizers not included in the main paper: DeiT III, SAM, SigLip, ConvNeXT, DINOv2, Dreamsim, and MAE. As shown in Fig. 2. The observed trends are consistent across featurizers. For *accidental* views, specific camera orientations obscure an object's true 3D structure, effectively reducing its perceived dimensionality by collapsing one axis of depth or perspective. For *other* views, we observe uncommon orientations, such as objects seen from the back or upside down, as well as instabilities caused by varying lighting conditions. The CO3D dataset, being more complex, introduces additional sources of instability, including occlusions and image blur.

4. Additional Visual Question Answering (VQA) Examples

To complement the VQA analysis presented in the main paper, we provide additional qualitative examples in Fig. 3. These examples further highlight LLaVA's capabilities in generating descriptive captions across stable, *accidental*, and *other* viewpoints. As in the main paper, we used LLaVA-1.5 [9], which leverages CLIP as its backbone. For stable viewpoints, the generated captions remain accurate and closely align with the ground truth (GT) descriptions. However, for *accidental* and *other* viewpoints, the captions frequently contain inaccuracies, including misinterpretations of objects and hallucinated details that are not present in the image. For instance, the model might describe a black table as ”a laptop sitting on a table,” despite the absence of a laptop, or misinterpret a light fixture as an umbrella. We hypothesize that this occurs because the pose of the light fixture in the image resembles a pose commonly associated with umbrellas in the training set, leading the model to incorrectly associate the two objects. In the case of *accidental* viewpoints, the 3D structure of objects is often collapsed, making it difficult or even impossible to accurately identify the object. This inherent ambiguity in the viewpoint could be acknowledged by the VQA model, which might express uncertainty about the object's identity rather than providing a confident but inaccurate description. Incorporating mechanisms to quantify or communicate uncertainty in such cases could improve the robustness and reliability of VQA models under challenging viewing conditions.

5. Additional Monocular 3D Reconstruction Results

To complement the monocular 3D reconstruction analysis presented in the main paper, we provide additional qualitative examples in Fig. 4. These examples further illustrate the significant impact of viewpoint instability on reconstruction accuracy, using Stable Fast 3D [1] with DINOv2 [11] as the image featurizer. For stable viewpoints, the model generates reconstructions with accurate geometry and well-preserved details, aligning closely with the ground truth. However, *accidental* viewpoints—characterized by insufficient depth cues due to the camera's orientation—result in collapsed or distorted reconstructions. This highlights the challenge of reconstructing 3D shapes when critical geometric information is unavailable. *Other* viewpoints present additional difficulties, as the atypical and rarely seen angles lead the model to misinterpret the object. This results in substantial inaccuracies in the reconstructed shape.

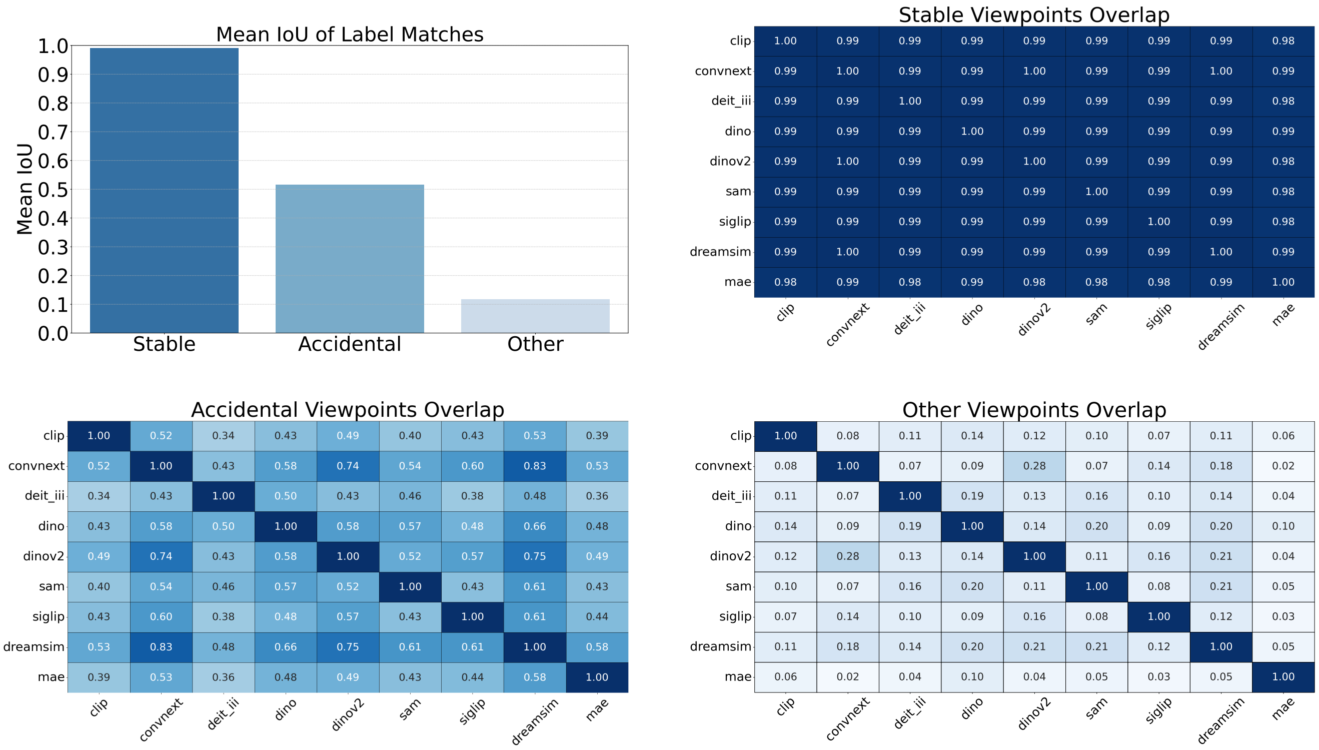


Figure 1. **Agreement levels on stable, *accidental*, and *other* viewpoints across nine featurizers, evaluated using Intersection over Union (IoU).** (i) The top-left heatmap shows the mean IoU scores across the nine featurizers for stable, *accidental*, and *other* viewpoints. (ii) The top-right heatmap displays pairwise IoU scores among the nine featurizers for stable viewpoints, demonstrating strong agreement. (iii) The bottom-left heatmap presents pairwise IoU scores for *accidental* viewpoints, indicating moderate agreement across featurizers. (iv) The bottom-right heatmap highlights pairwise IoU scores for *other* viewpoints, showing minimal agreement, with many pairs of featurizers exhibiting no overlap (IoU scores under 0.1).

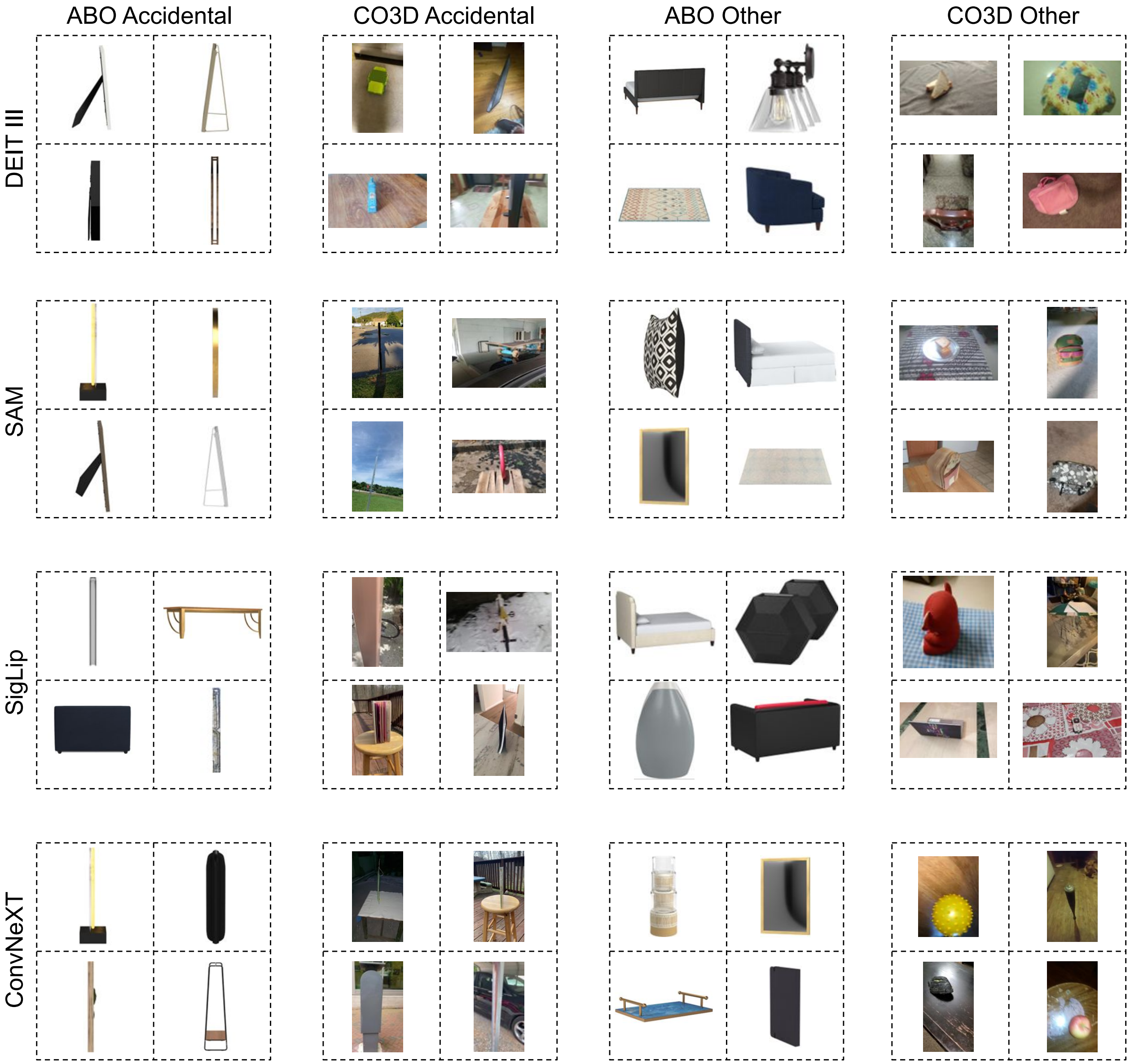


Figure continues on next page.

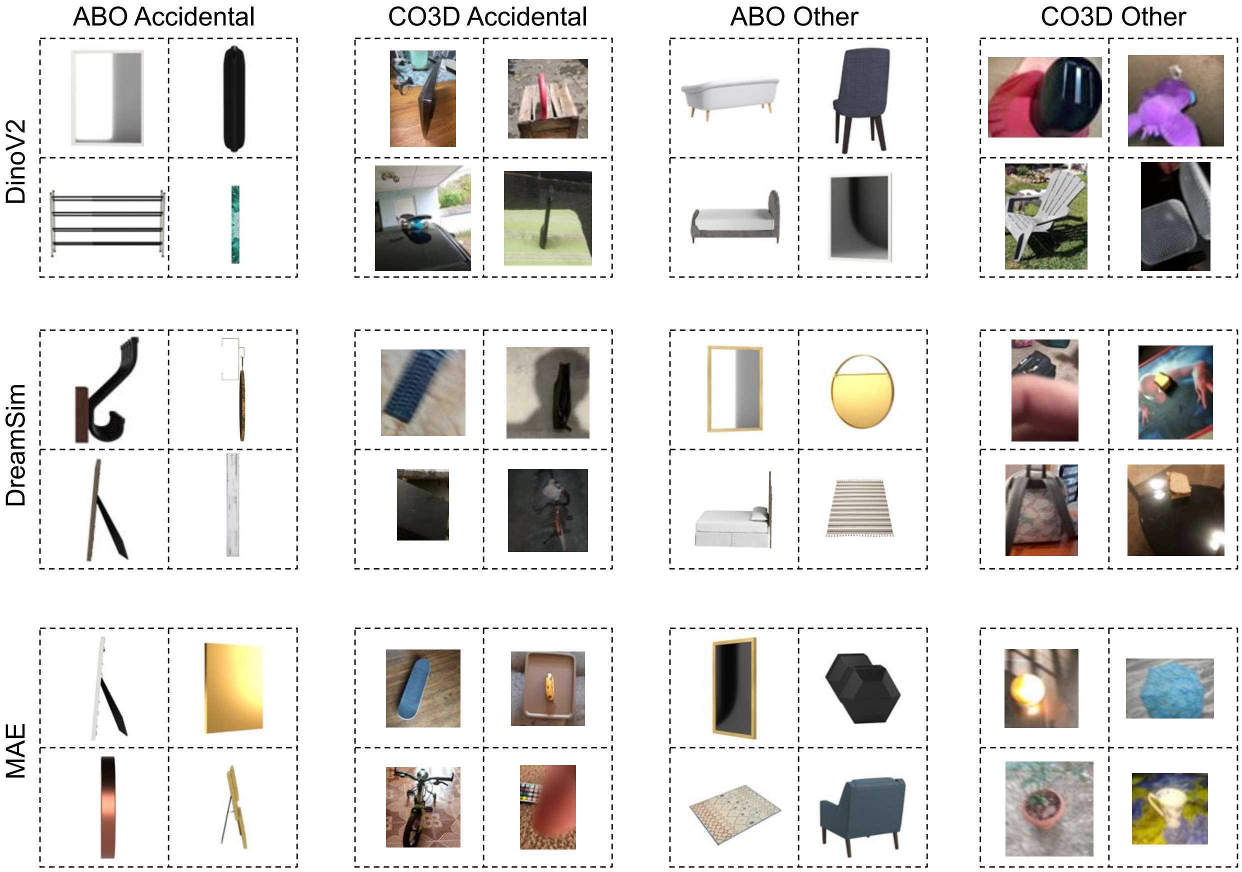
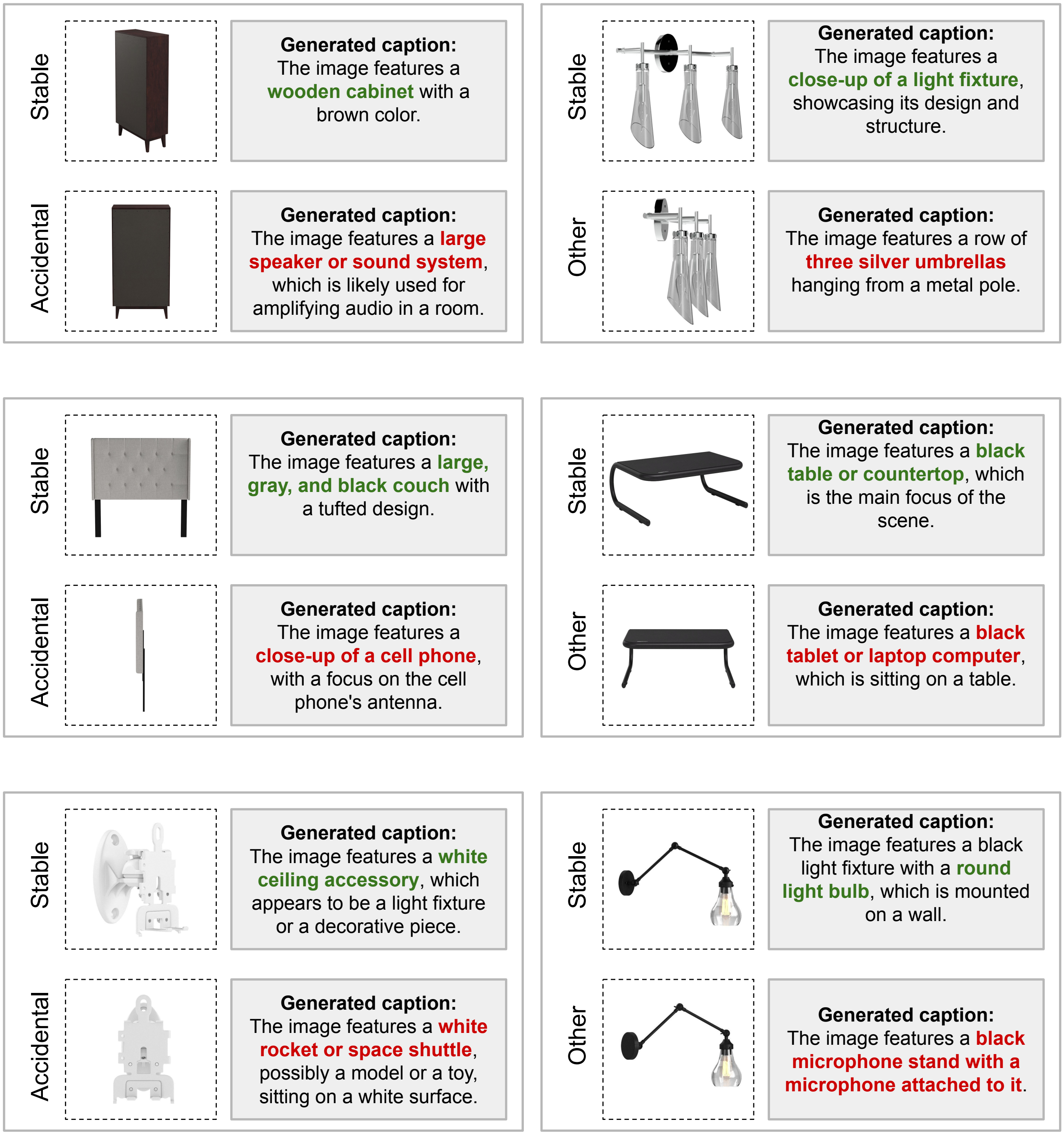


Figure 2. **Examples of *accidental* and *other* viewpoints using ConvNeXT, DeiT III, DINOv2, Dreamsim, MAE, SAM, and SigLip embeddings across the ABO and CO3D datasets.** Similar to CLIP and DINO discussed in the main paper, *accidental* views obscure an object's true structure, effectively reducing its perceived dimensionality. *Other* views present uncommon orientations, such as objects seen from the back or upside down, and are also affected by sudden lighting changes. The CO3D dataset, being more complex, introduces additional sources of instability, including occlusions and abrupt changes in focus.



Generated caption:
The image features a **wooden cabinet** with a brown color.

Generated caption:
The image features a **large speaker or sound system**, which is likely used for amplifying audio in a room.

Generated caption:
The image features a **close-up of a light fixture**, showcasing its design and structure.

Generated caption:
The image features a row of **three silver umbrellas** hanging from a metal pole.

Generated caption:
The image features a **large, gray, and black couch** with a tufted design.

Generated caption:
The image features a **close-up of a cell phone**, with a focus on the cell phone's antenna.

Generated caption:
The image features a **black table or countertop**, which is the main focus of the scene.

Generated caption:
The image features a **black tablet or laptop computer**, which is sitting on a table.

Generated caption:
The image features a **white ceiling accessory**, which appears to be a light fixture or a decorative piece.

Generated caption:
The image features a **white rocket or space shuttle**, possibly a model or a toy, sitting on a white surface.

Generated caption:
The image features a black light fixture with a **round light bulb**, which is mounted on a wall.

Generated caption:
The image features a **black microphone stand with a microphone attached to it**.

Figure 3. **Examples of generated captions for stable, *other*, and *accidental* viewpoints using LLaVA-1.5 (CLIP backbone).** Captions for stable viewpoints are accurate, while those for *accidental* and *other* viewpoints frequently contain inaccuracies, often misinterpreting objects or hallucinating details. For example, a black table might be described as ”a laptop sitting on a table,” despite no laptop being present.

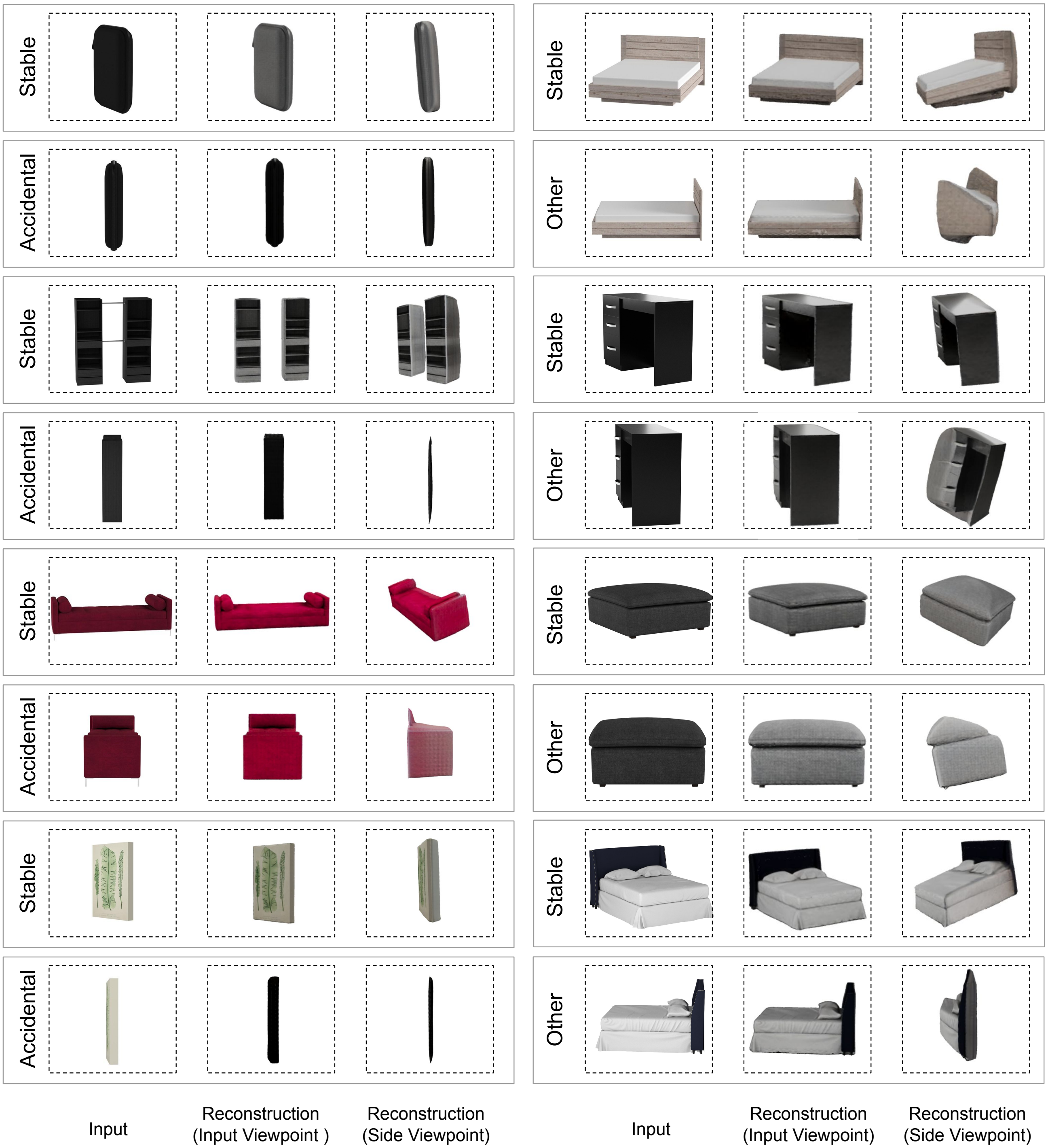

Stable
Accidental
Stable
Accidental
Stable
Accidental
Stable
Accidental
Input
Reconstruction (Input Viewpoint)
Reconstruction (Side Viewpoint)
Stable
Other
Stable
Other
Stable
Other
Stable
Other
Input
Reconstruction (Input Viewpoint)
Reconstruction (Side Viewpoint)

Figure 4. **Additional single-view reconstruction results for stable, *accidental*, and *other* viewpoints using Stable Fast 3D [1] with DINOv2 [11] as the image featurizer.** Stable viewpoints yield accurate reconstructions with well-preserved geometry. *Accidental* viewpoints, lacking sufficient depth cues, result in collapsed or distorted reconstructions. *Other* viewpoints, due to atypical angles, often lead to substantial inaccuracies.

References

[1] 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. 2, 7

[2] 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. 1

[3] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 1

[4] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 1

[7] 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. 1

[8] 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. 1

[9] 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. 2

[10] 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. 1

[11] 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. 1, 2, 7

[12] 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. 1

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[14] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022. 1

[15] 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. 1