Why MLLMs Struggle to Determine Object Orientations

Supplementary Material

7. Hypothesis: CLIP Fails to Encode Object Orientation Information

7.1. Collage of Images for LLaVA 1.5 and 1.6

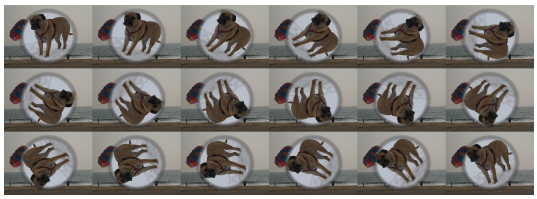


Figure 8. Collage of every 20th image from the images with the dog foreground (biggest foreground)

7.2. Plots showing Regression comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct

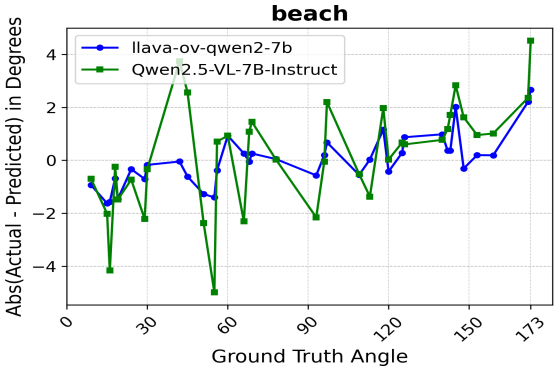


Figure 9. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with dog for 36 randomly selected images

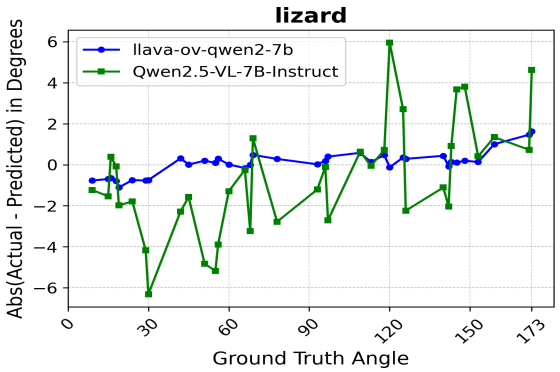


Figure 10. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with lizard for 36 randomly selected images

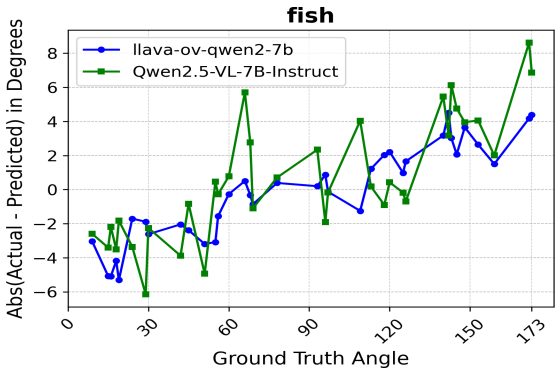


Figure 11. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with fish for 36 randomly selected images

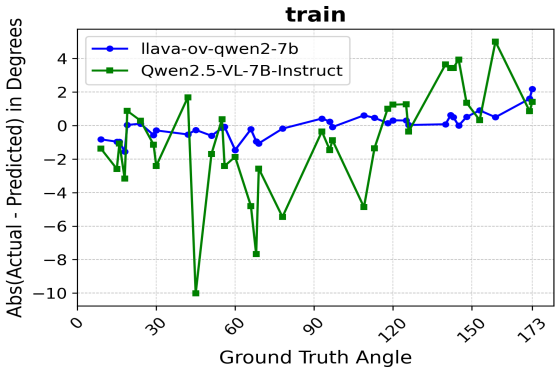


Figure 12. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with train for 36 randomly selected images

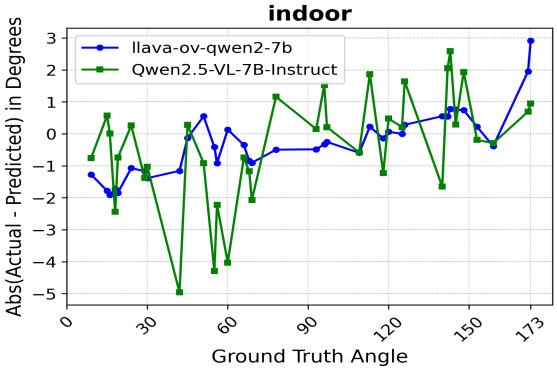


Figure 13. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with indoor for 36 randomly selected images

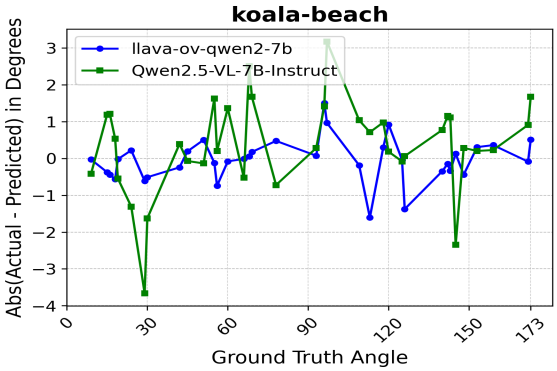


Figure 14. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with koala for 36 randomly selected images

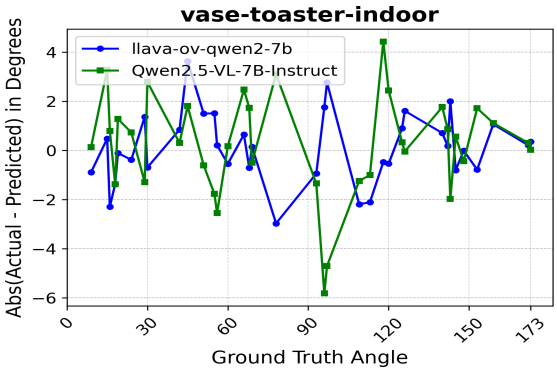


Figure 16. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with vase and toaster for 36 randomly selected images

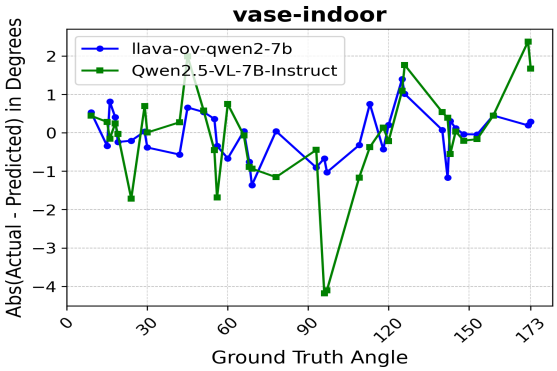


Figure 15. 2D orientation estimation performance comparison between LLaVA OneVision and Qwen2.5-VL-7B-Instruct on the images with vase for 36 randomly selected images

7.3. Plots showing Regression comparison between LLaVA 1.5 and 1.6

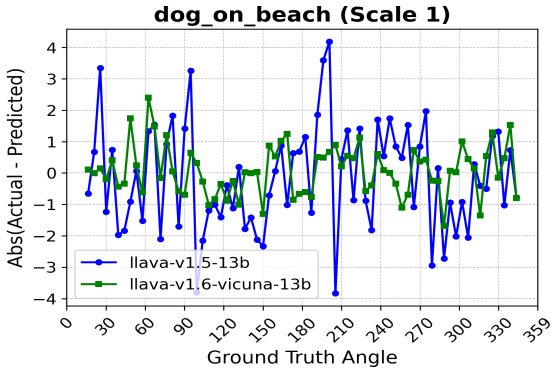


Figure 17. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with dog foregrounds (scale 1) for 72 randomly selected images

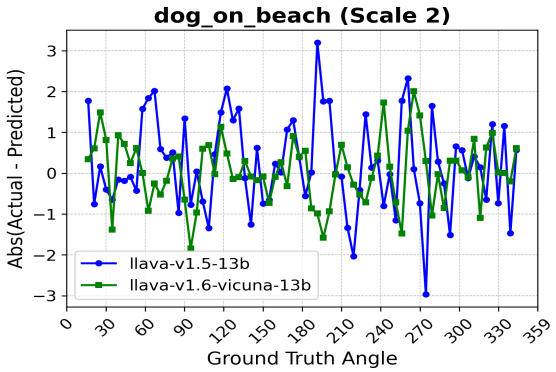


Figure 18. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with dog foregrounds (scale 2) for 72 randomly selected images

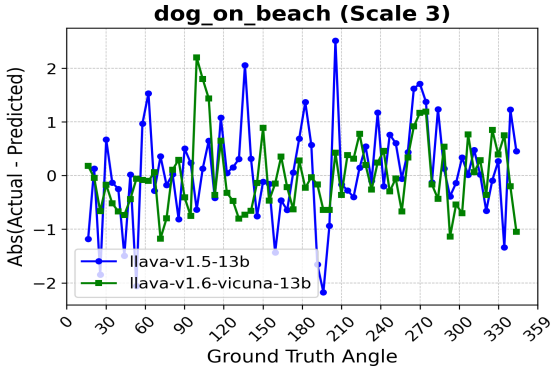


Figure 19. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with dog foregrounds (scale 3) for 72 randomly selected images

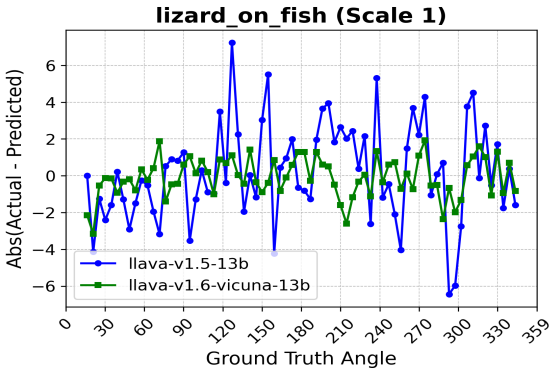


Figure 20. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with lizard foregrounds (scale 1) for 72 randomly selected images

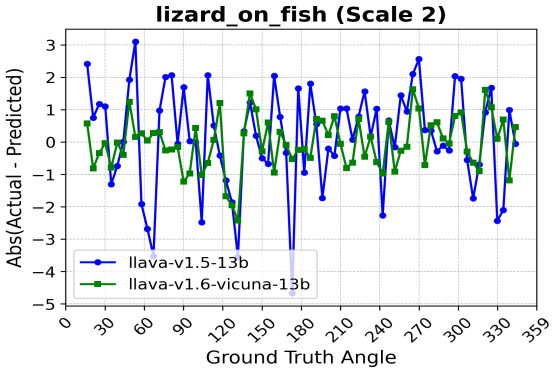


Figure 21. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with lizard foregrounds (scale 2) for 72 randomly selected images

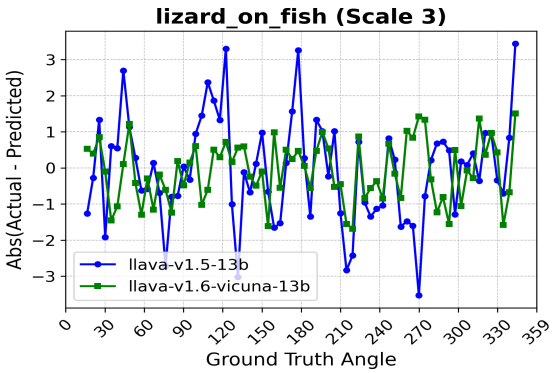


Figure 22. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with lizard foregrounds (scale 3) for 72 randomly selected images

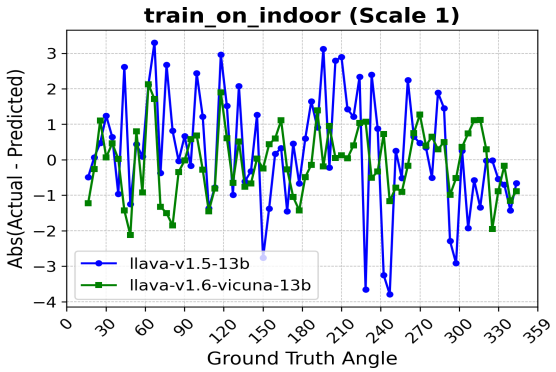


Figure 23. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with train foregrounds (scale 1) for 72 randomly selected images

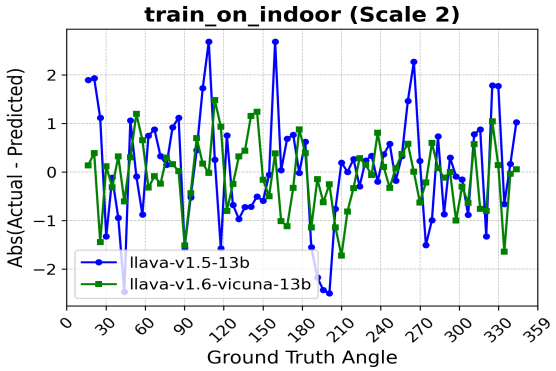


Figure 24. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with train foregrounds (scale 2) for 72 randomly selected images

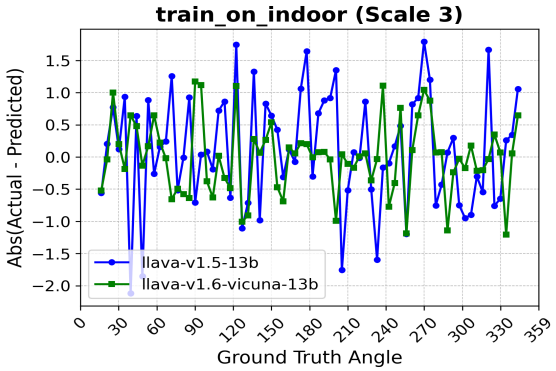


Figure 25. 2D orientation estimation performance comparison between LLaVA 1.5 and 1.6 on the images with train foregrounds (scale 3) for 72 randomly selected images

7.4. Plots Showing Statistical Analysis for LLaVA-OneVision and Qwen2.5-VL-7B-Instruct

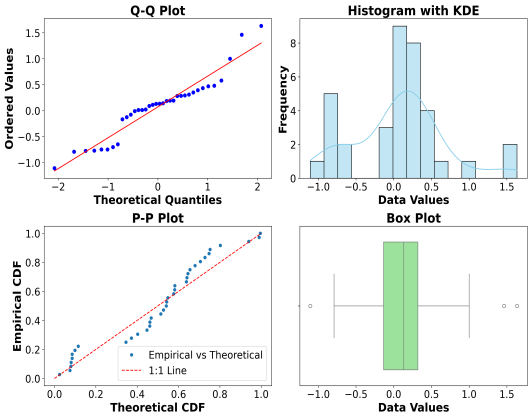


Figure 26. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with lizard. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

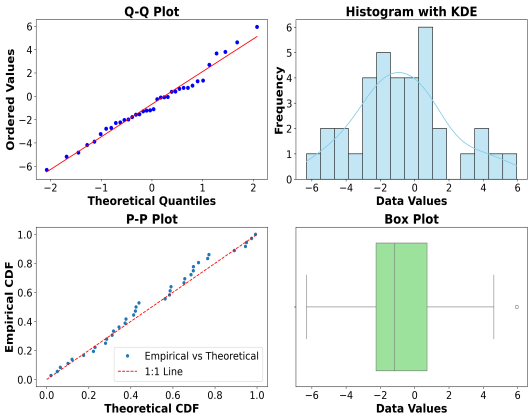


Figure 27. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with lizard. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

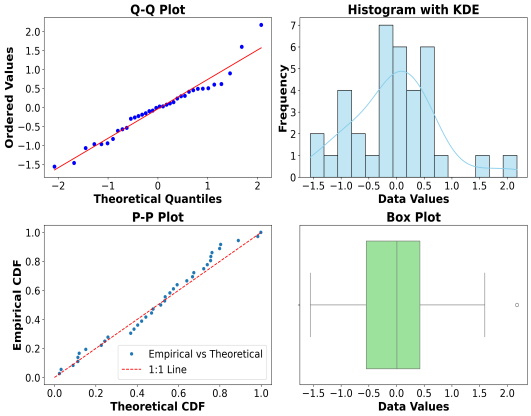


Figure 28. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with train. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

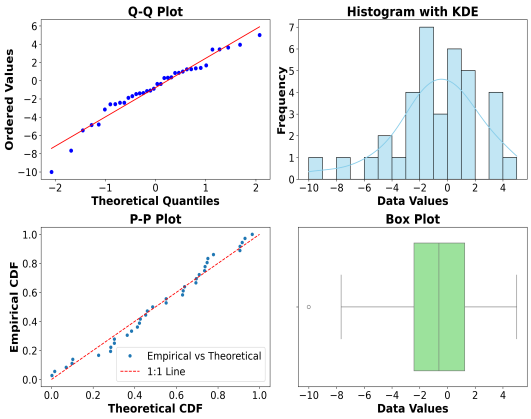


Figure 29. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with train. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

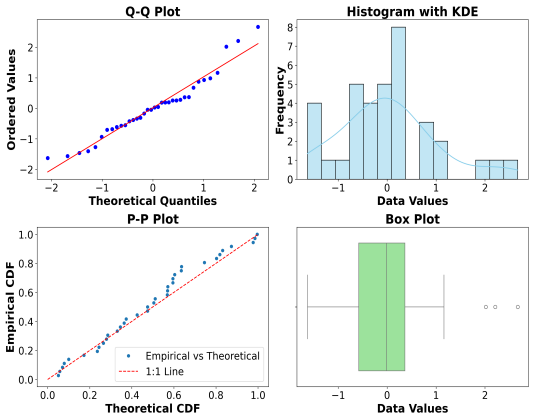


Figure 30. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with beach. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

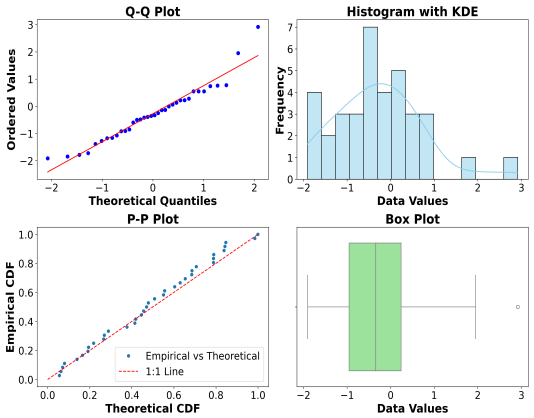


Figure 32. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with indoor. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

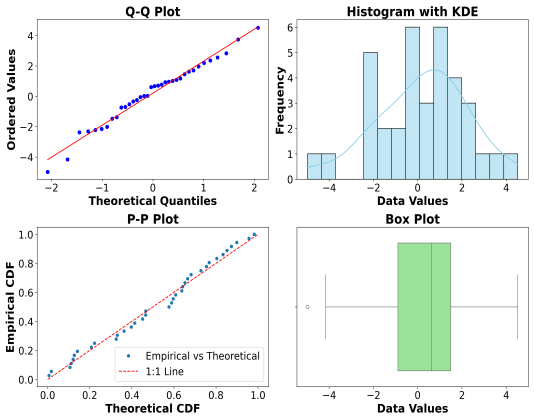


Figure 31. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with beach. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

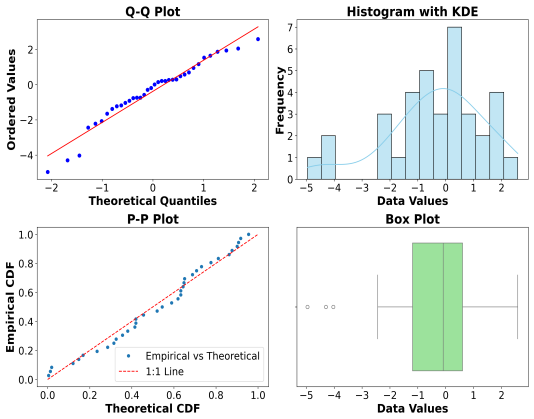


Figure 33. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with indoor. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

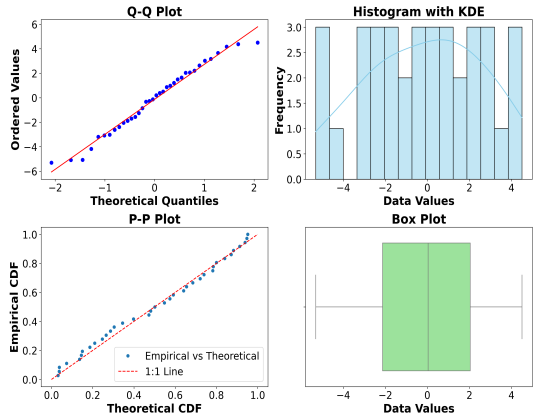


Figure 34. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with fish. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

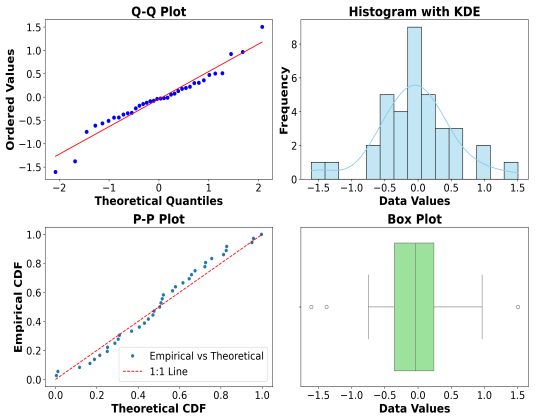


Figure 36. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with koala-beach. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

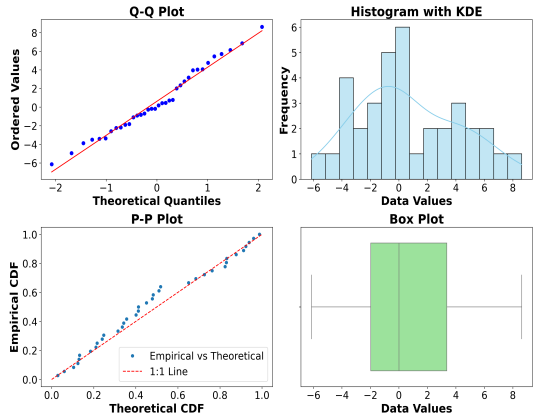


Figure 35. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with fish. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

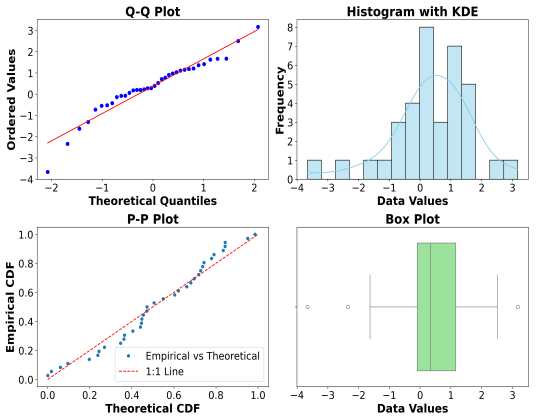


Figure 37. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with koala-beach. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

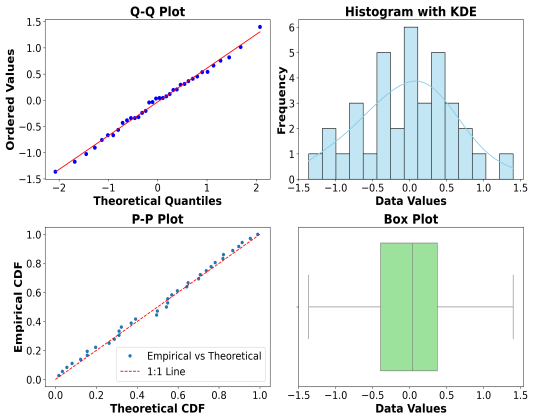


Figure 38. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with vase-indoor. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

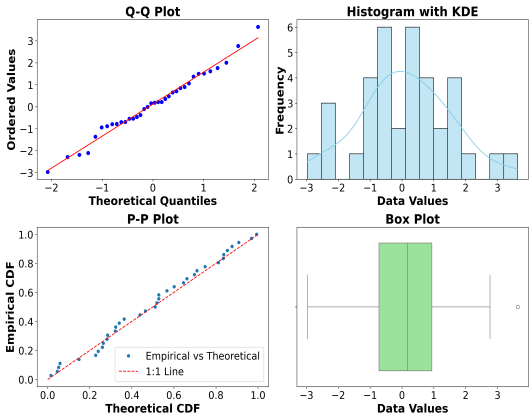


Figure 40. Statistical Analysis using visual plots for LLaVA-OneVision - results for images with vase-toaster-indoor. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

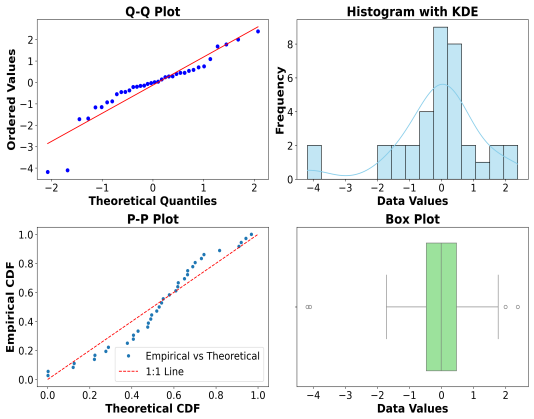


Figure 39. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with vase-indoor. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

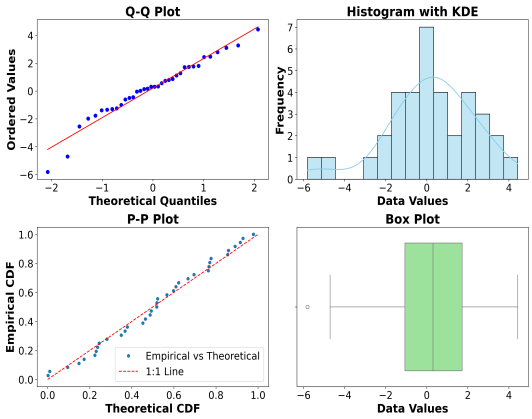


Figure 41. Statistical Analysis using visual plots for Qwen2.5-VL-7B-Instruct - results for images with vase-toaster-indoor. Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

7.5. Plots Showing Statistical Analysis for LLaVA 1.5 and 1.6

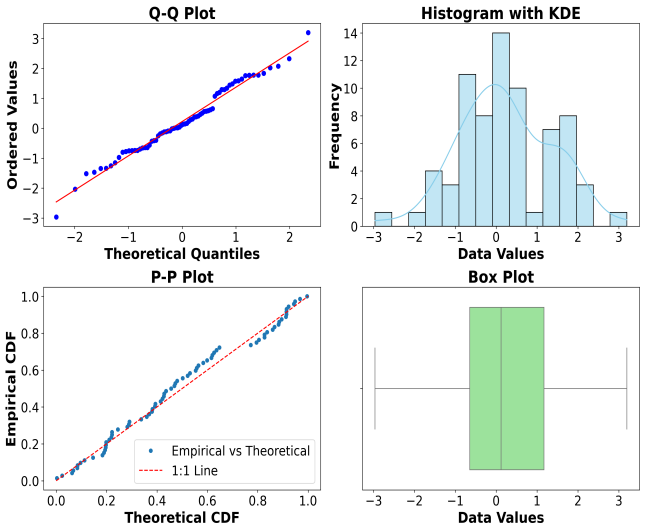


Figure 42. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with dog foregrounds (Scale 2). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

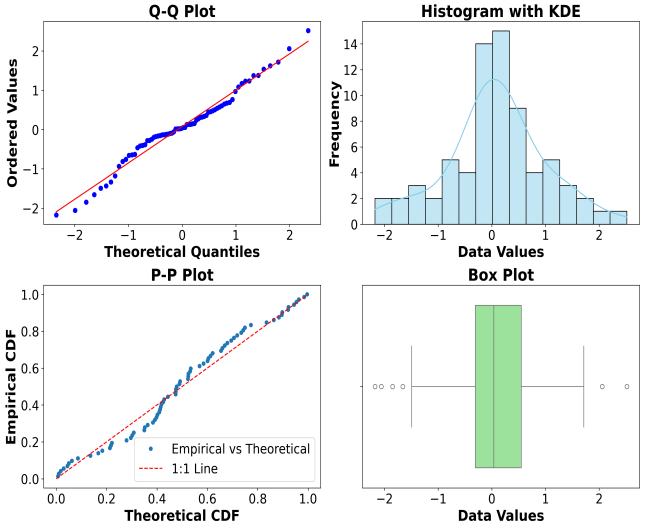


Figure 43. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with dog foregrounds (Scale 3). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

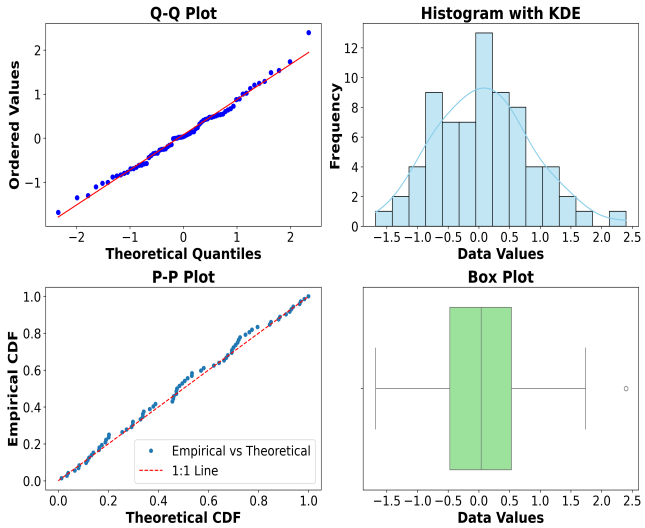


Figure 44. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with dog foregrounds (Scale 1). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

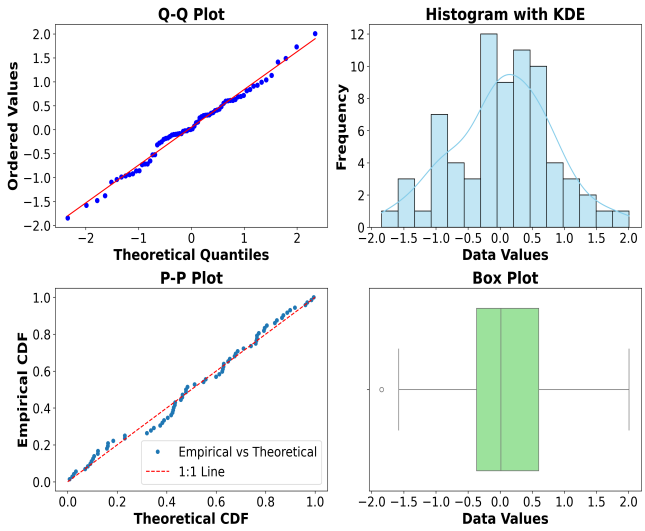


Figure 45. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with dog foregrounds (Scale 2). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

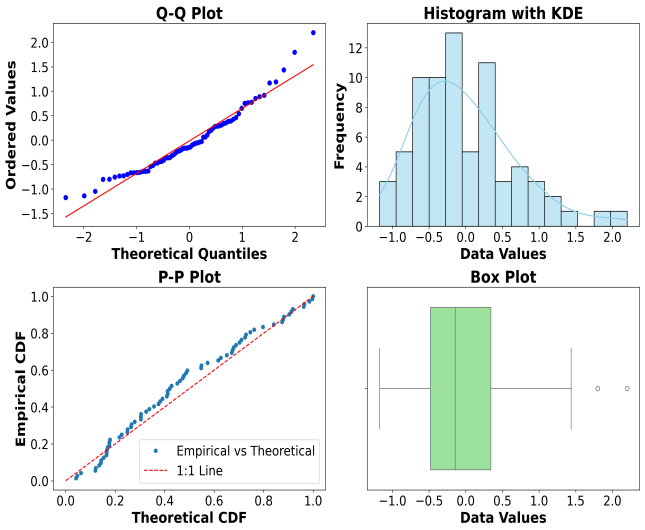


Figure 46. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with dog foregrounds (Scale 3). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

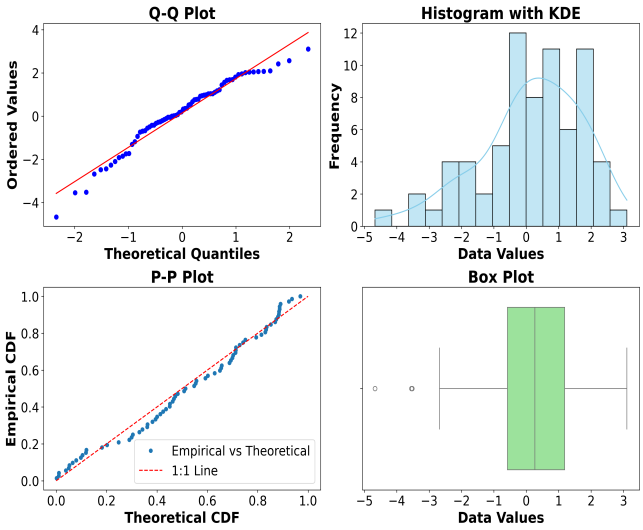


Figure 48. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with lizard foregrounds (Scale 2). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

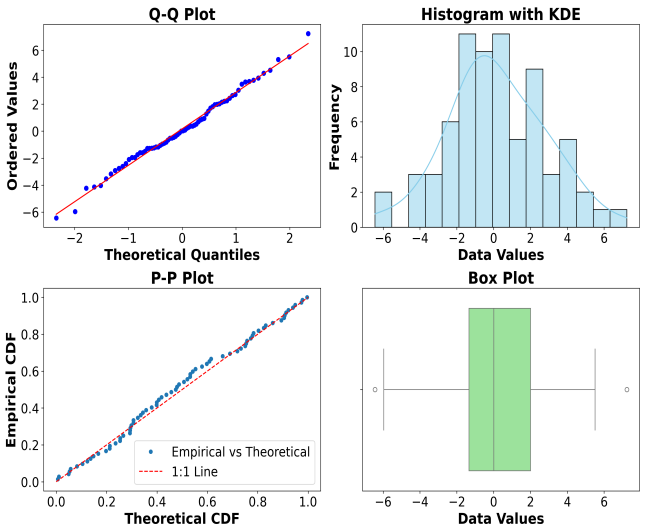


Figure 47. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with lizard foregrounds (Scale 1). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

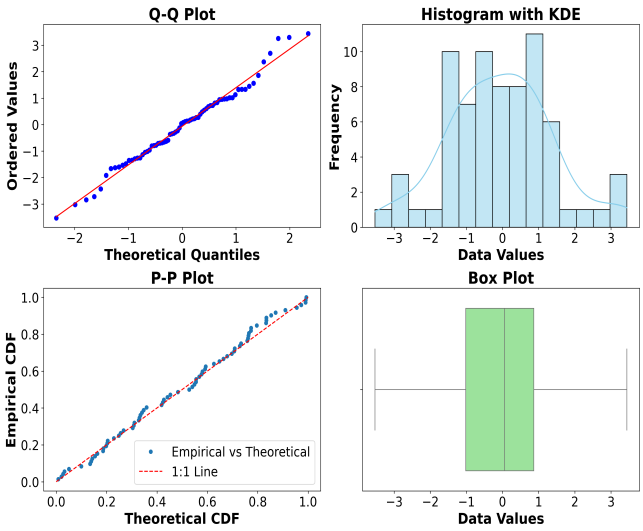


Figure 49. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with lizard foregrounds (Scale 3). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

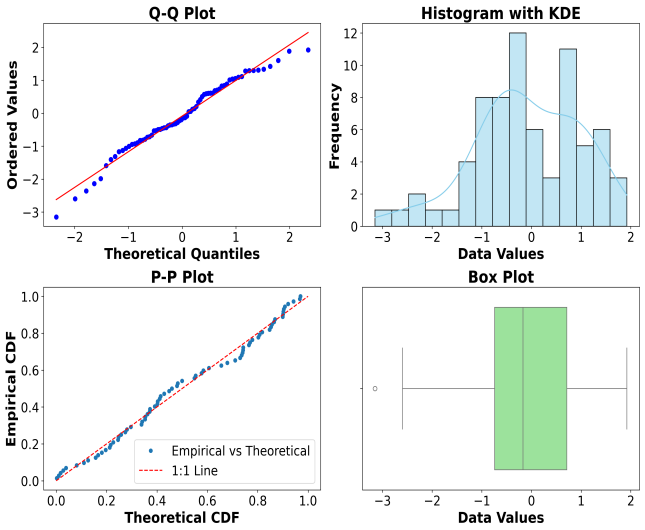


Figure 50. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with lizard foregrounds (Scale 1). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

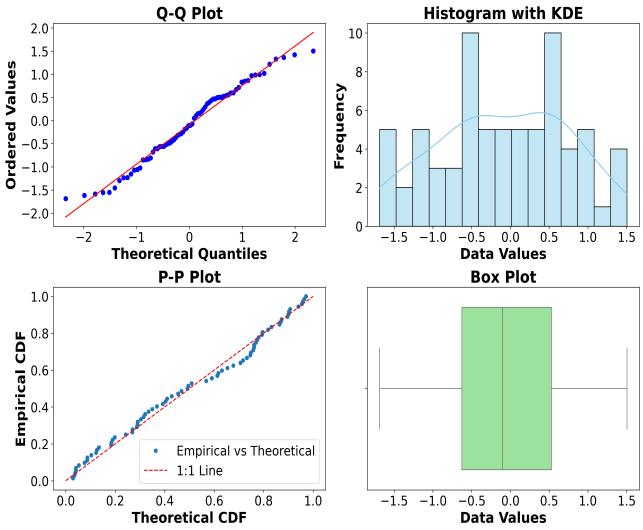


Figure 52. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with lizard foregrounds (Scale 3). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

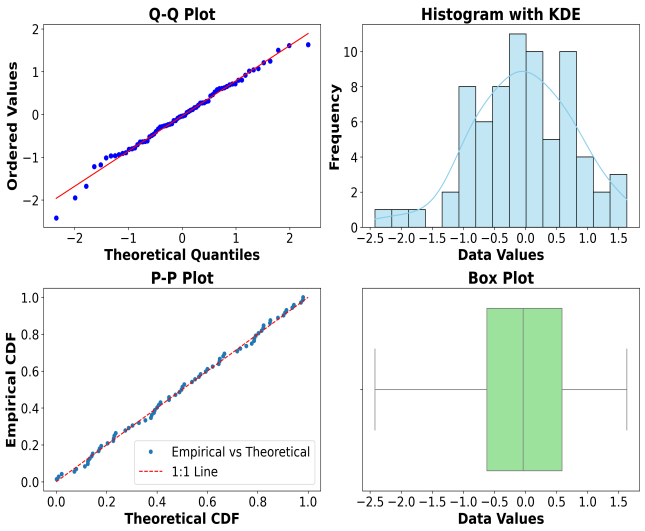


Figure 51. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with lizard foregrounds (Scale 2). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

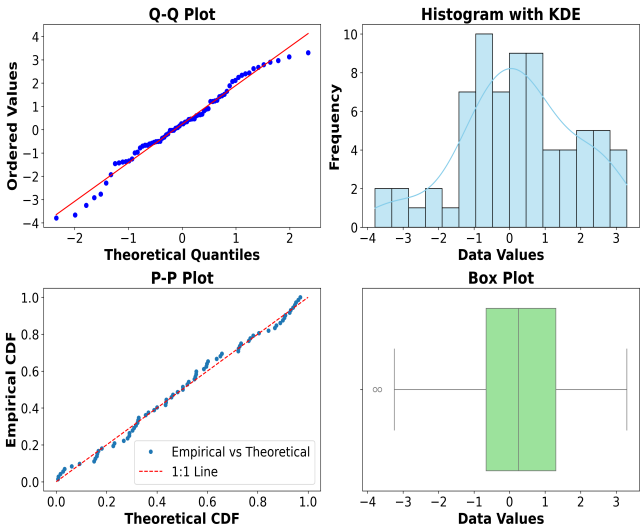


Figure 53. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with train foregrounds (Scale 1). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

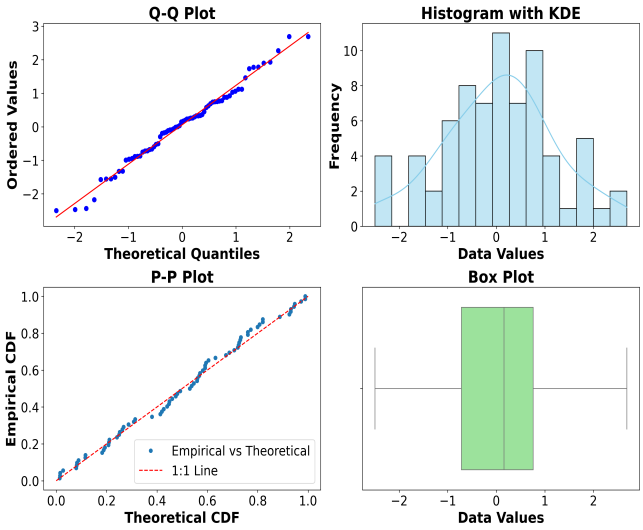


Figure 54. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with train foregrounds (Scale 2). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

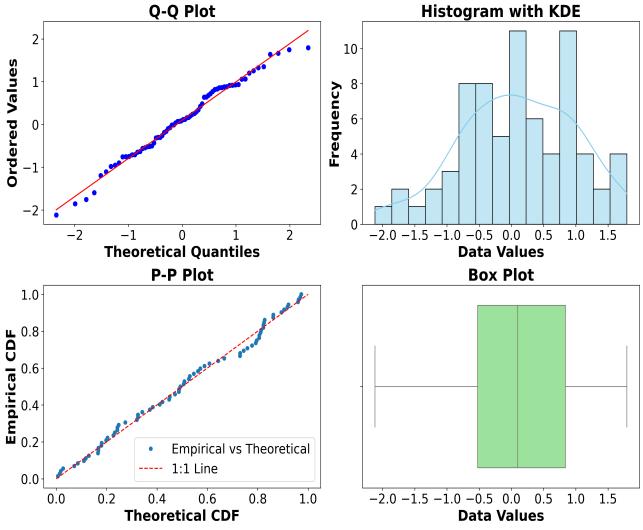


Figure 55. Statistical Analysis using visual plots for LLaVA 1.5 - results for images with train foregrounds (Scale 3). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

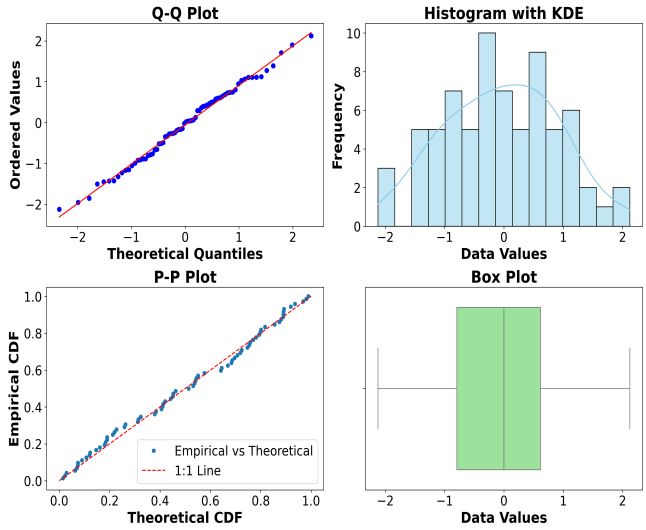


Figure 56. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with train foregrounds (Scale 1). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

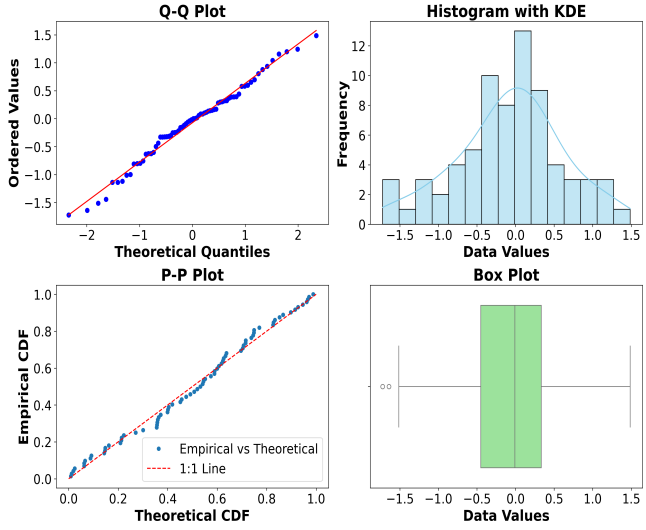


Figure 57. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with train foregrounds (Scale 2). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

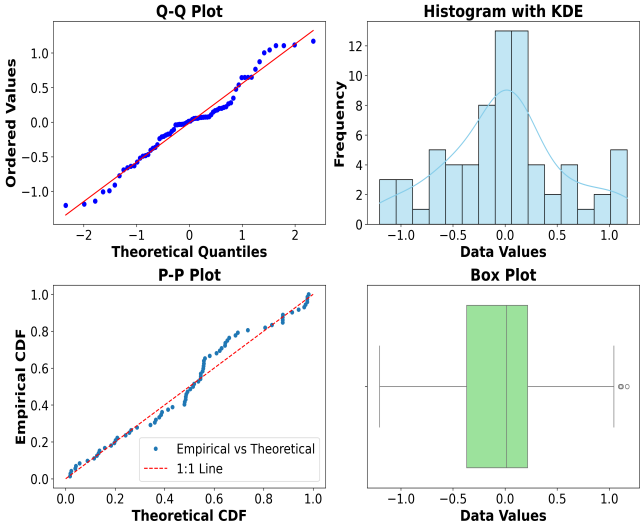


Figure 58. Statistical Analysis using visual plots for LLaVA 1.6 - results for images with train foregrounds (Scale 3). Together with the numerical results in Table 3 and the visual plots in this Figure, we can conclude that the residual error distribution (Table 2) is random/Gaussian

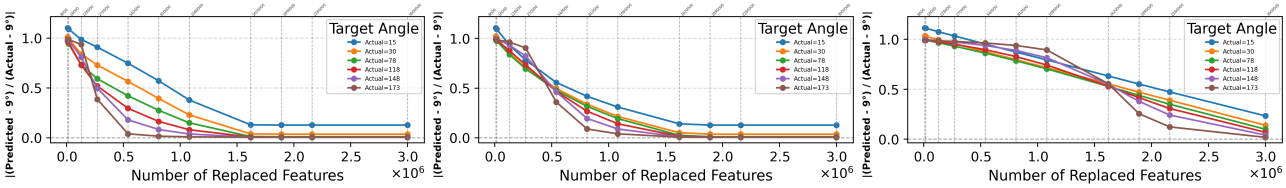
7.6. LLaVA 1.5 and 1.6 query responses

LLaVA 1.5		LLaVA 1.6	
Angle (°)	**Count**	**Angle (°)**	**Count**
90	25	90	40
not possible	10	not possible	17
45	9	0	6
0	8	no answer	4
no answer	4	180	2
Summary Performance		**LLaVA 1.5**	**LLaVA 1.6**
Correct Answers ($\|diff\| \leq 5$)		3	2
Incorrect Answers ($\|diff\| > 5$)		55	49
Correct Answers ($\|diff\| \leq 20$)		7	7
Incorrect Answers ($\|diff\| > 20$)		51	44
Correct Answers ($\|diff\| \leq 45$)		16	17
Incorrect Answers ($\|diff\| > 45$)		42	34

Table 5. LLaVA-LLaMA results for the 72 test samples with the beach scene in the background. LLaMA frequently responds that the 2D orientation is 90° or "not possible to determine". The number of correct answers (among valid responses) is very low, even with high thresholds

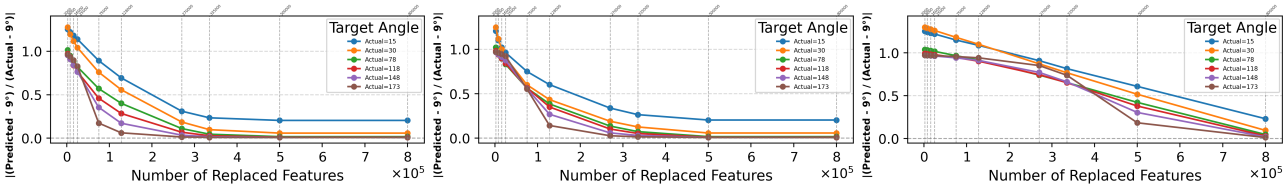
8. Orientation Encoding Properties

8.1. Feature Substitution Plots for LLaVA-OneVision and Qwen2.5-VL-7B-Instruct



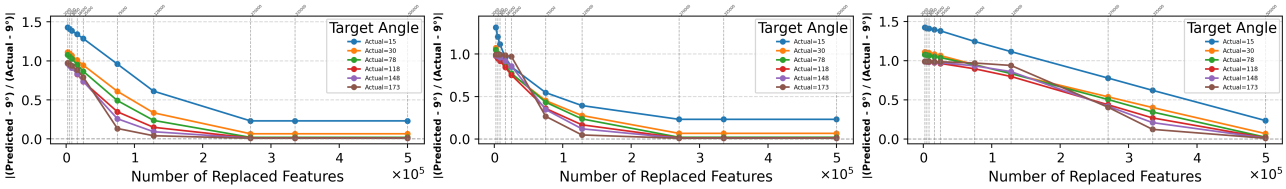
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 59. Incremental feature substitution for LLaVA-OneVision on images with the lizard scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 540,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



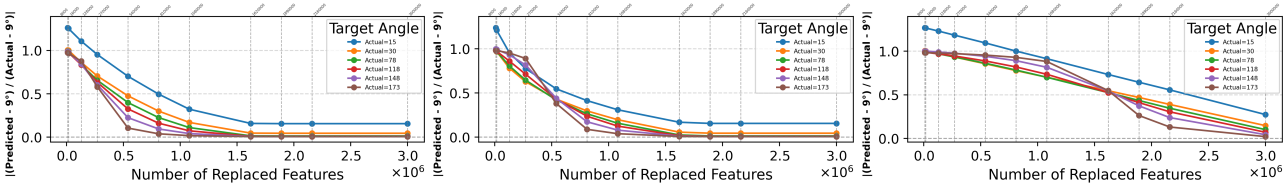
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 60. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the lizard scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



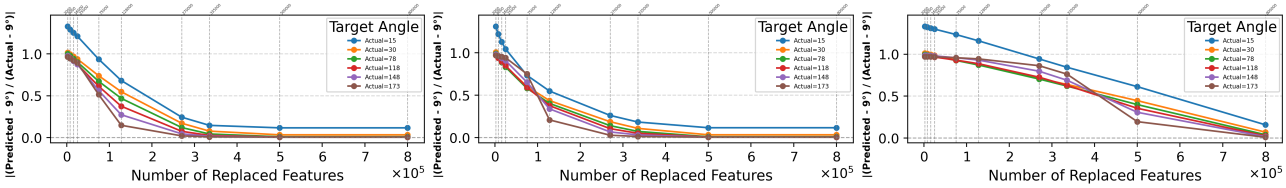
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 62. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the train scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



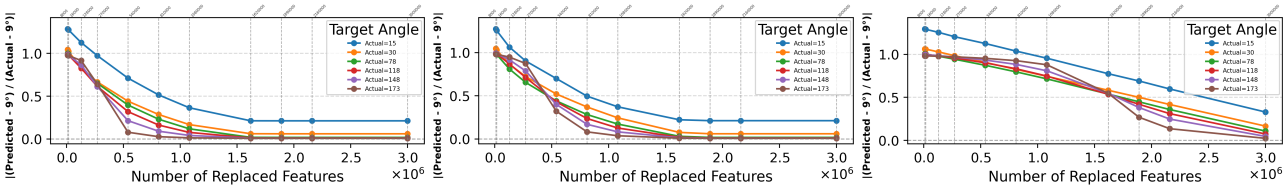
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 63. Incremental feature substitution for LLaVA-OneVision on images with the beach scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 540,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



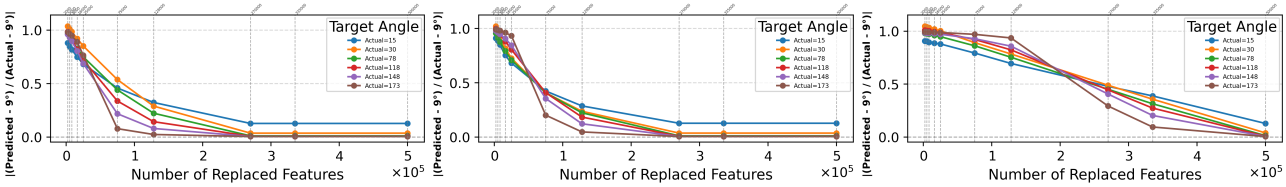
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 64. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the beach scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



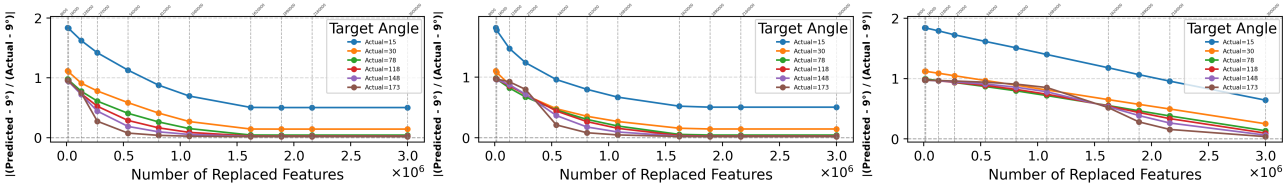
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 65. Incremental feature substitution for LLaVA-OneVision on images with the indoor scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 540,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



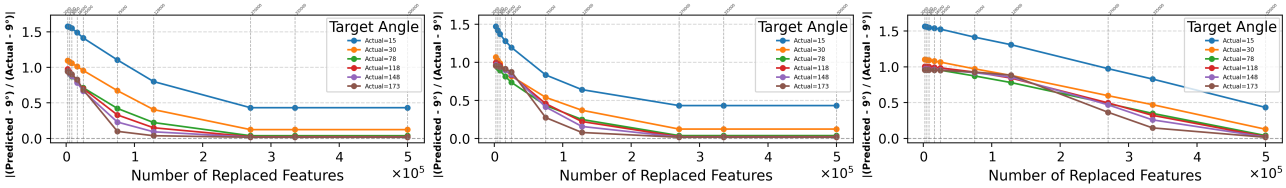
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 66. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the indoor scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



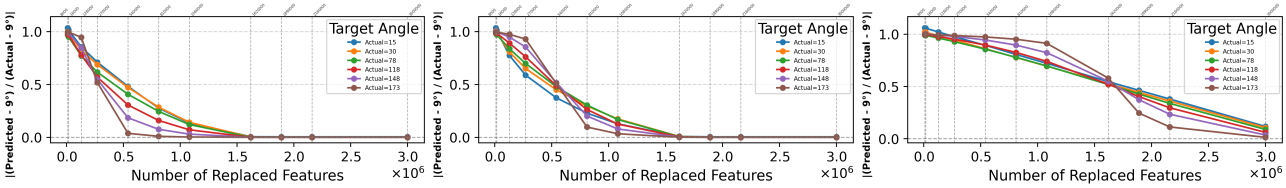
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 67. Incremental feature substitution for LLaVA-OneVision on images with the fish scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 540,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



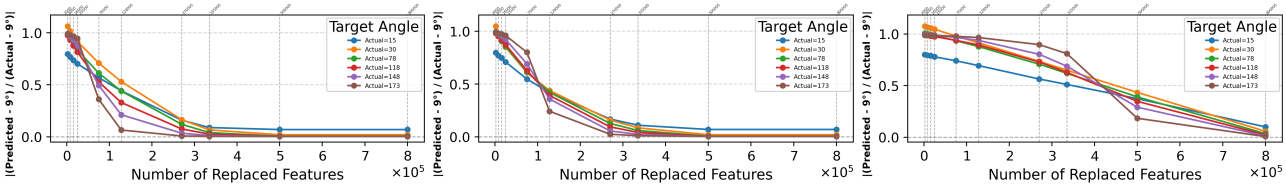
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 68. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the fish scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



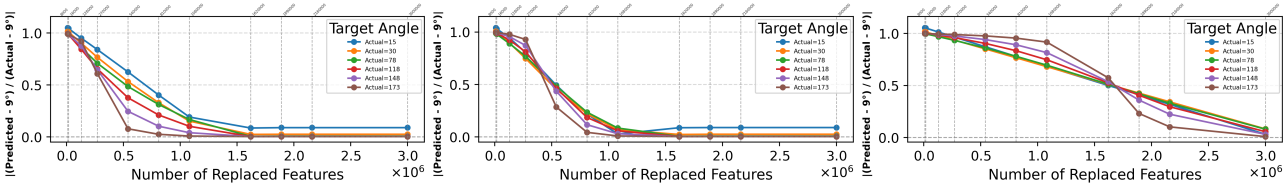
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 69. Incremental feature substitution for LLaVA-OneVision on images with the koala-beach scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 540,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



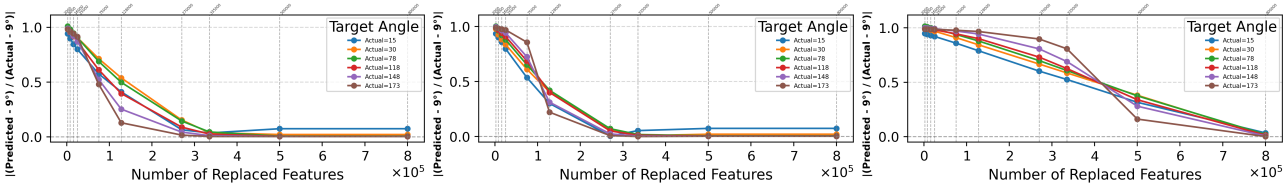
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 70. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the koala-beach scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



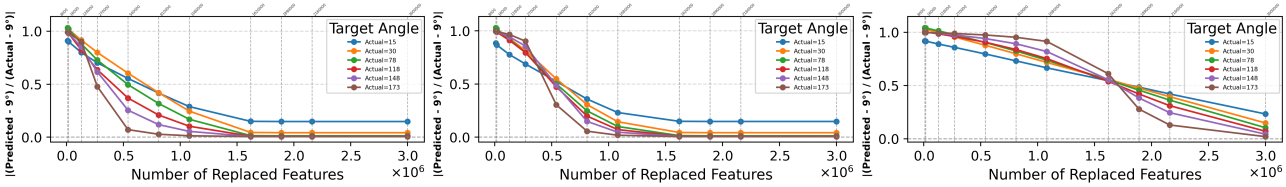
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 71. Incremental feature substitution for LLaVA-OneVision on images with the vase-indoor scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 540,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



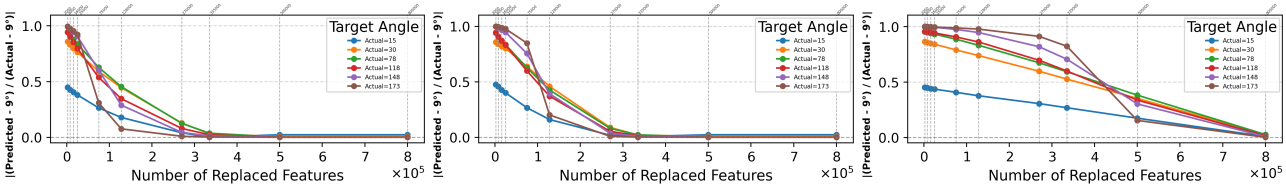
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 72. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the vase-indoor scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

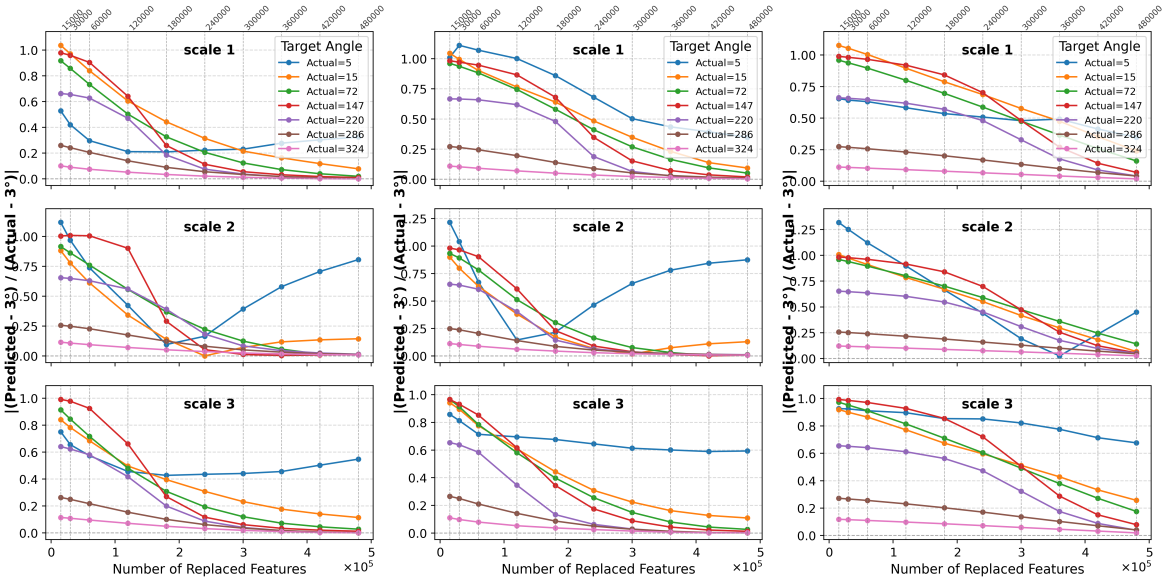
Figure 73. Incremental feature substitution for LLaVA-OneVision on images with the vase-toaster-indoor scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 540,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

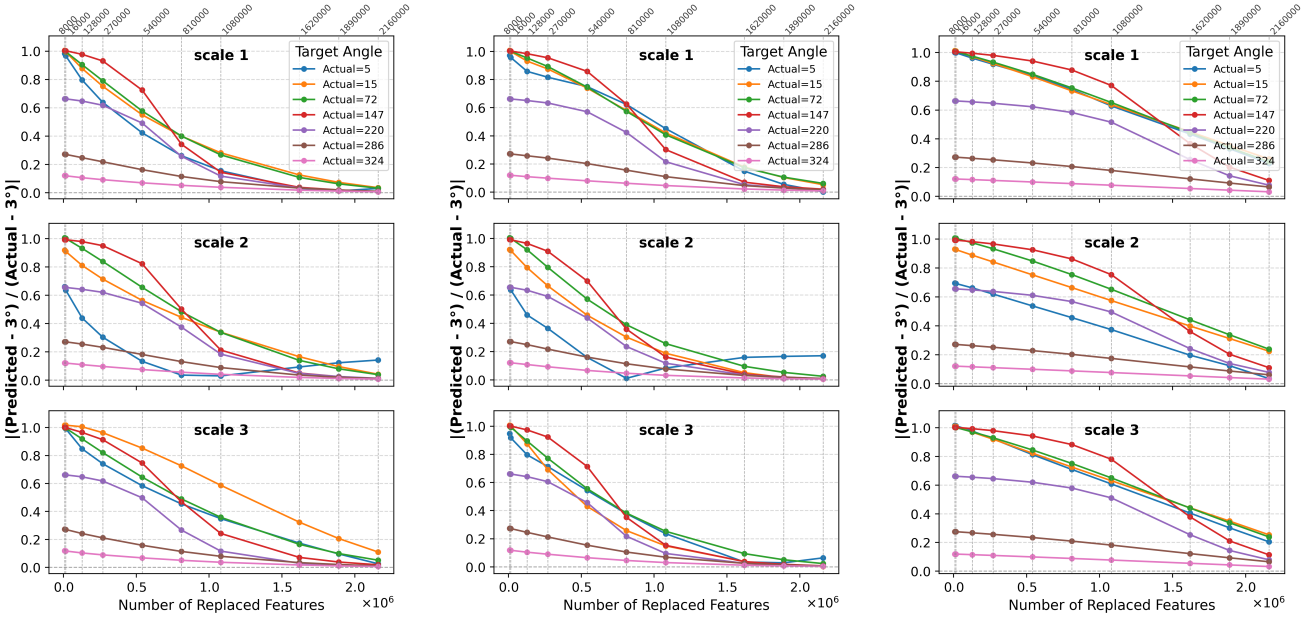
Figure 74. Incremental feature substitution for Qwen2.5-VL-7B-Instruct on images with the vase-toaster-indoor scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 128,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.

8.2. Feature Substitution Plots for LLaVA 1.5 and 1.6



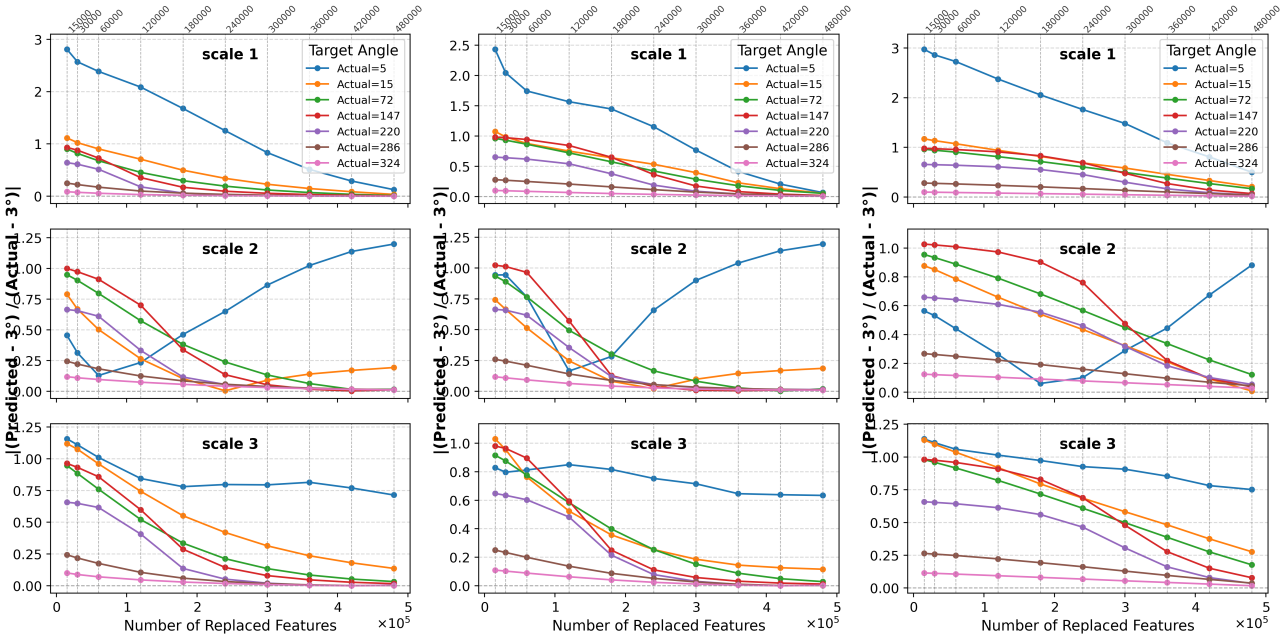
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 75. Incremental feature substitution for LLaVA 1.5 on images with the beach background scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 20,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



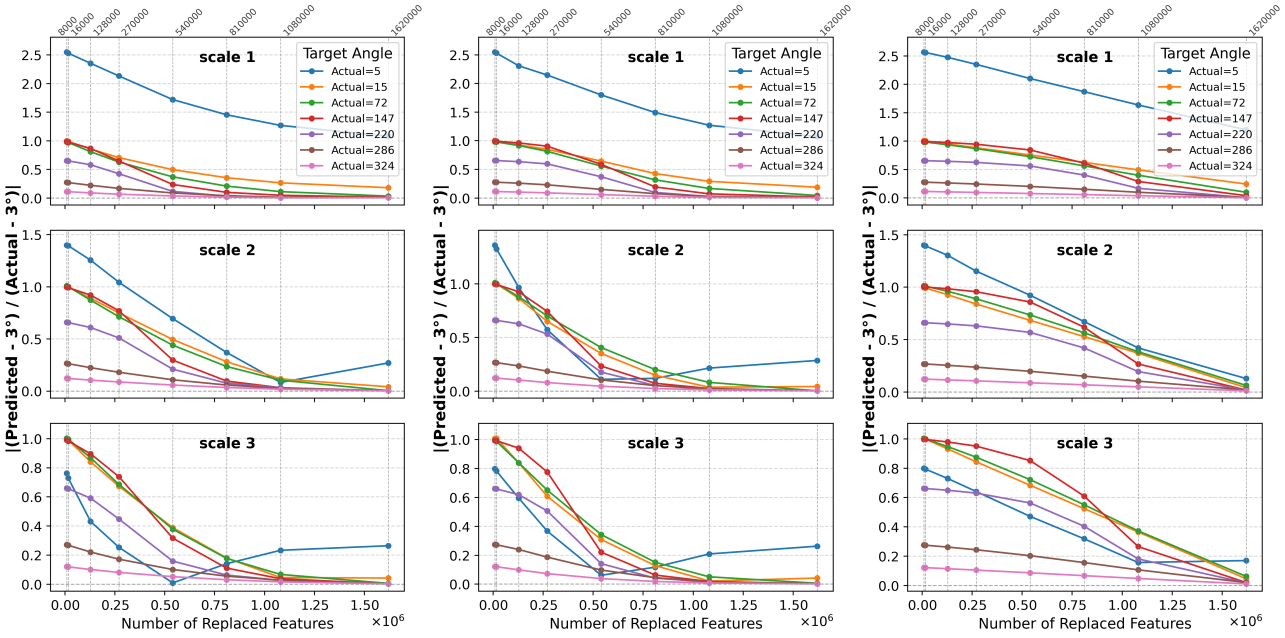
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 76. Incremental feature substitution for LLaVA 1.6 on images with the beach background scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 16,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



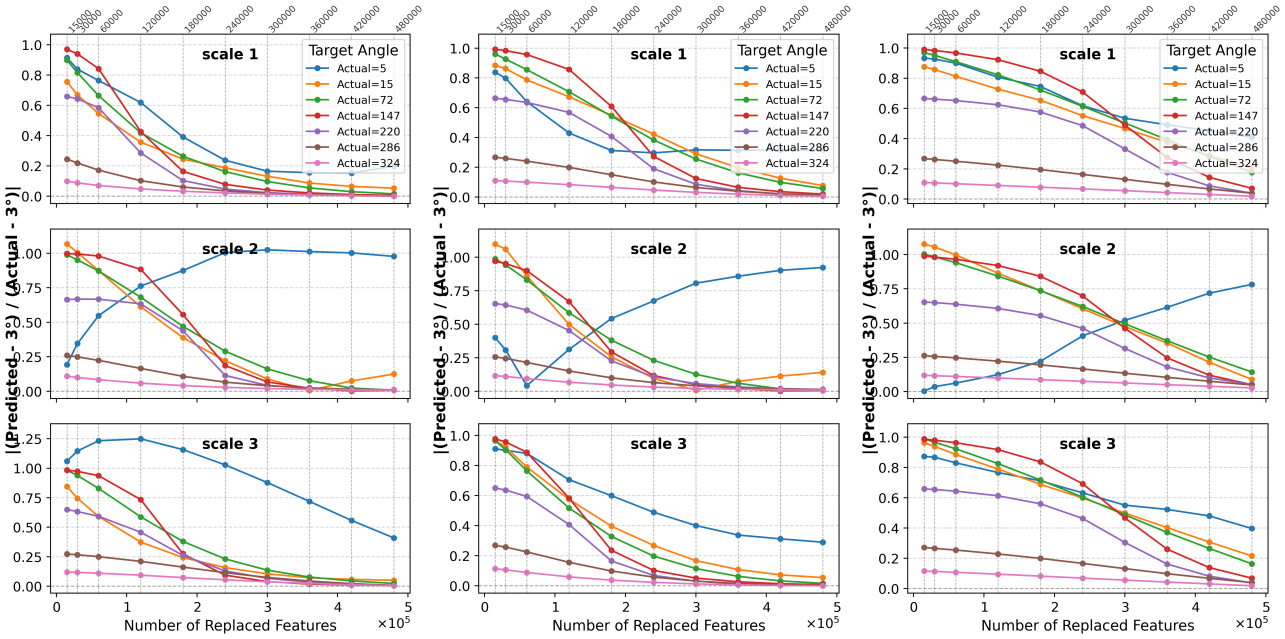
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 77. Incremental feature substitution for LLaVA 1.5 on images with the fish background scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 20,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



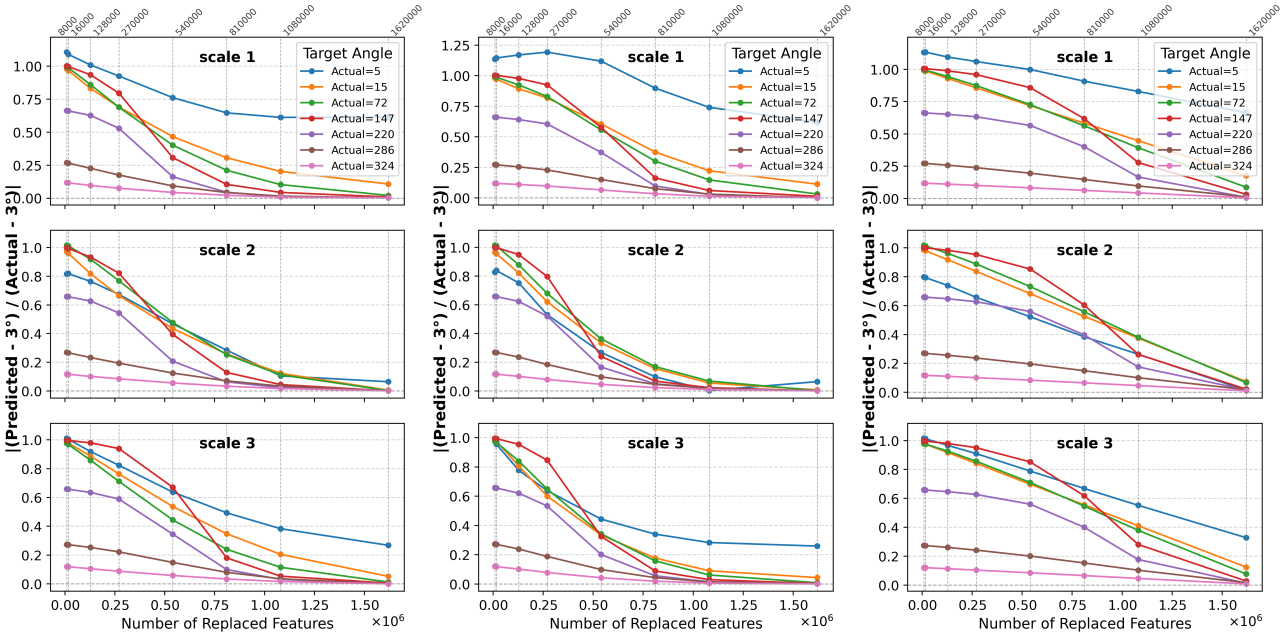
(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 78. Incremental feature substitution for LLaVA 1.6 on images with the fish background scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 16,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.



(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 79. Incremental feature substitution for LLaVA 1.5 on images with the indoor background scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 20,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^5.) This implies the orientation information is highly diffuse.



(a) Ordered By Model Weight (b) Ordered By Value Difference (c) Picked Randomly

Figure 80. Incremental feature substitution for LLaVA 1.6 on images with the indoor background scene. No matter how the features are selected (according to the magnitude of the weights in the regressor or the absolute difference between anchor and target feature values, or randomly). 16,000 features or more must be replaced to fool the predictor. (Note that the x-axis is the number of feature substitutions times 10^6.) This implies the orientation information is highly diffuse.

8.3. Background vs. Foreground Rotation

In trying to better understand how orientation information is embedded by visual encoders, we looked by at Table 2 and noticed something small but odd: when estimating foreground orientations, the MAE does not get larger as the foreground patch gets smaller. In fact, although the effect is small, predictions are better for smaller image patches. This led us to investigate the relationship between backgrounds and the estimated orientation of the foreground.

We evaluated the model trained on only foreground rotated images using two variant image sets - (1) background rotated and foreground static, and (2) both background and foreground rotated. The results for LLaVA 1.5 on images with dog foregrounds (scale 1) are shown in Figures 81 - 87. Both experiments fared poorly with an MAE upwards of 80°.

To understand why the model is unable to predict the foreground orientation when the background is rotated, we repeat the experiments on an image set with synthetic backgrounds (see Figure 82) with horizontal and vertical lines. Our hypothesis is that accurate foreground orientation is dependent on the background being in its canonical orientation, so this experiment gauges the sensitivity of the foreground orientation estimation to edges in the background. Results are in Tables 6-8. Our experiments show that accurately estimating the orientation of the foreground is dependent on the orientation of the background. If the training set contains background and foreground rotations, then the test with both background and foreground rotations perform very well. But this is not the case when the training and test sets contain only background rotations. This requires further investigation.

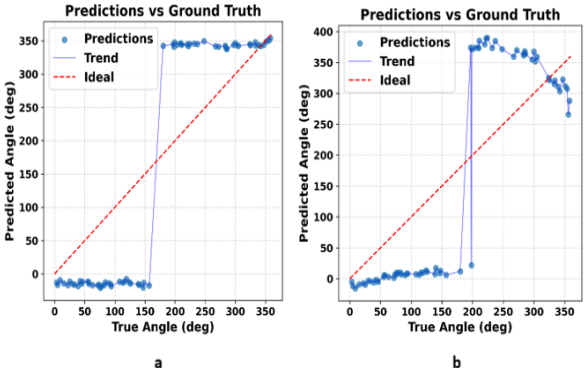


Figure 81. Results of foreground orientation estimation by LLaVA 1.5 for dog images (scale 1) when (a) only BG (b) BG and FG are rotated in the test images and training set consists of images with only FG rotated.

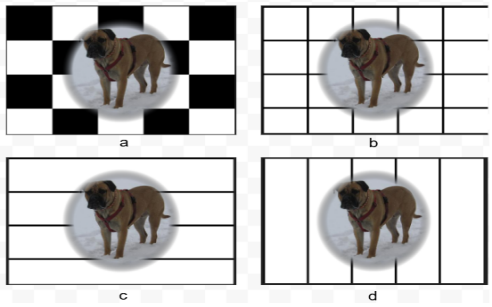


Figure 82. Images with synthetic backgrounds used to test the impact of background rotations on foreground orientation estimation.

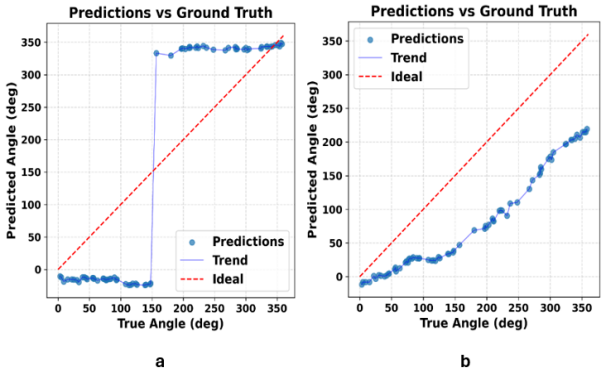


Figure 83. Results of foreground orientation estimation by LLaVA 1.6 for dog images (scale 1) when (a) only BG (b) BG and FG are rotated in the test images and training set consists of images with only FG rotated.

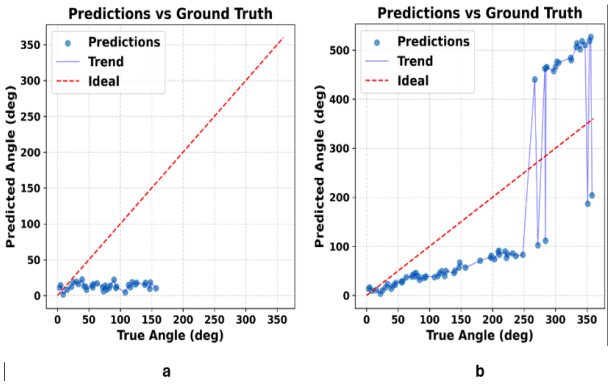


Figure 84. Results of foreground orientation estimation by LLaVA 1.5 for lizard images (scale 1) when (a) only BG (b) BG and FG are rotated in the test images and training set consists of images with only FG rotated.

S/N	Train Set	Test set	MAE (degrees)	
			LLaVA 1.5	LLaVA 1.6
1	dog on chessboard (FG rotated)	baseline (FG rotated)	1.1	0.71
		BG + FG rotated	96.5	92.34
		BG rotated	80.09	80.48
	dog on grid (FG rotated)	baseline (FG rotated)	1.28	0.68
		BG + FG rotated	92.13	88.14
		BG rotated	81.77	91.29
	dog on horizontal lines (FG rotated)	baseline (FG rotated)	1.35	0.7
		BG + FG rotated	86.2	89.45
		BG rotated	75.03	80.59
	dog on vertical lines (FG rotated)	baseline (FG rotated)	1.4	0.64
		BG + FG rotated	84.2	86.67
		BG rotated	92.06	81.28
2	dog on grid lines - BG + FG rot.	dog on vertical lines - BG + FG rot.	1.87	1.37
	dog on grid lines - BG + FG rot.	dog on horizontal lines - BG + FG rot.	1.75	2.14
3	dog on chessboard - BG + FG rot.	dog on vertical lines - BG + FG rot.	2.73	3.33
	dog on chessboard - BG + FG rot.	dog on horizontal lines - BG + FG rot.	2.73	3.65
4	dog on grid lines - BG rot.	dog on horizontal lines - BG rot.	29.95	21.67
	dog on grid lines - BG rot.	dog on vertical lines - BG rot.	58.98	115.78

Table 6. Impact of background (BG) image rotation on foreground (FG) rotation (rot.) estimation for dog foreground images (Scale 1) using LLaVA 1.5 and LLaVA 1.6 - Mean Absolute Error (MAE) for synthetic background image sets under different rotation conditions. When background is rotated, performance: (1) degrades sharply when trained on only FG rot. images, (2) and (3) improves significantly when trained on BG+FG rot., (4) improves moderately when trained on only BG rot.

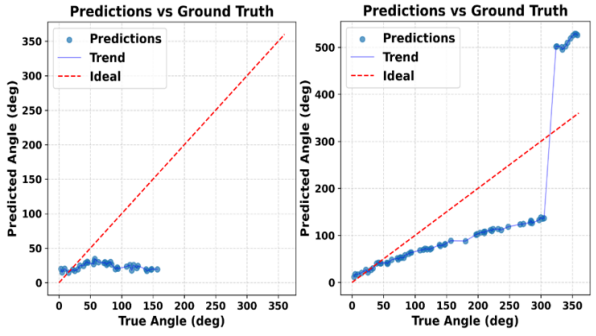


Figure 85. Results of foreground orientation estimation by LLaVA 1.6 for lizard images (scale 1) when (a) only BG (b) BG and FG are rotated in the test images and training set consists of images with only FG rotated.

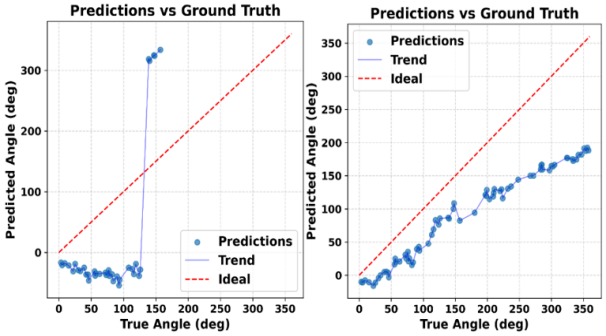


Figure 86. Results of foreground orientation estimation by LLaVA 1.5 for train images (scale 1) when (a) only BG (b) BG and FG are rotated in the test images and training set consists of images with only FG rotated.

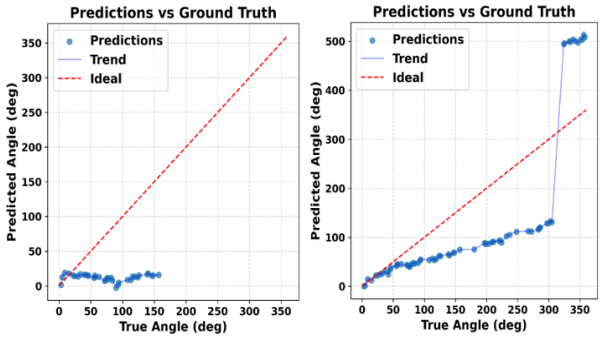


Figure 87. Results of foreground orientation estimation by LLaVA 1.6 for train images (scale 1) when (a) only BG (b) BG and FG are rotated in the test images and training set consists of images with only FG rotated.

S/N	Train Set	Test set	MAE (degrees)	
			LLaVA 1.5	LLaVA 1.6
1	lizard on chessboard (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	1.81 85.2 80.2	1.36 85.55 79.93
	lizard on grid (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	1.87 84.1 60.18	1.16 87.22 67.97
	lizard on horizontal lines (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	2.23 87.47 70.57	1.24 84.32 69.18
	lizard on vertical lines (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	1.68 94.57 72.29	1.38 85.71 70.37
2	lizard on grid lines - BG + FG rot.	lizard on vertical lines - BG + FG rot.	1.36	1.37
	lizard on grid lines - BG + FG rot.	lizard on horizontal lines - BG + FG rot.	1.75	1.61
3	lizard on chessboard - BG + FG rot.	lizard on vertical lines - BG + FG rot.	2.49	3.88
	lizard on chessboard - BG + FG rot.	lizard on horizontal lines - BG + FG rot.	2.8	3.75
4	lizard on grid lines - BG rot.	lizard on horizontal lines - BG rot.	37.96	29.28
	lizard on grid lines - BG rot.	lizard on vertical lines - BG rot.	94.99	81.97

Table 7. Impact of background (BG) image rotation on foreground (FG) rotation (rot.) estimation for lizard foreground images (Scale 1) using LLaVA 1.5 and LLaVA 1.6 - Mean Absolute Error (MAE) for synthetic background image sets under different rotation conditions. When background is rotated, performance: (1) degrades sharply when trained on only FG rot. images, (2) and (3) improves significantly when trained on BG+FG rot., (4) improves moderately when trained on only BG rot.

S/N	Train Set	Test set	MAE (degrees)	
			LLaVA 1.5	LLaVA 1.6
1	train on chessboard (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	1.2 86.59 80.46	0.77 82.96 81
	train on grid (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	1.3 87.4 68.76	0.94 87.96 79.58
	train on horizontal lines (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	1.43 82.36 72.68	0.85 88.84 78.66
	train on vertical lines (FG rotated)	baseline (FG rotated) BG + FG rotated BG rotated	1.53 77.62 62.08	0.92 84.03 80.15
2	train on grid lines - BG + FG rot.	train on vertical lines - BG + FG rot.	1.52	1.51
	train on grid lines - BG + FG rot.	train on horizontal lines - BG + FG rot.	1.31	1.25
3	train on chessboard - BG + FG rot.	train on vertical lines - BG + FG rot.	2.61	3.51
	train on chessboard - BG + FG rot.	train on horizontal lines - BG + FG rot.	2.44	3.1
4	train on grid lines - BG rot.	train on horizontal lines - BG rot.	55.09	25.01
	train on grid lines - BG rot.	train on vertical lines - BG rot.	102.92	88.76

Table 8. Impact of background (BG) image rotation on foreground (FG) rotation (rot.) estimation for train foreground images (Scale 1) using LLaVA 1.5 and LLaVA 1.6 - Mean Absolute Error (MAE) for synthetic background image sets under different rotation conditions. When background is rotated, performance: (1) degrades sharply when trained on only FG rot. images, (2) and (3) improves significantly when trained on BG+FG rot., (4) improves moderately when trained on only BG rot.