

7 Supplementary Materials Overview

This supplementary material provides additional details of the paper along with supplementary results that were omitted from the main paper due to space constraints. In Section 8 we present details and additional supplementary results of the synthetic dataset experiments. In Section 9, we present details and additional supplementary results of the real dataset (COCO [19]) experiments. Finally, in Section 10, we present details and additional supplementary results of the case study in healthcare (MIMIC-CXR [14,16]).

8 Synthetic Validation

8.1 Experiments

Dataset. Each image in the dataset is described by a set of concepts describing distinct colored geometric shapes: $\{\text{red, green, blue}\} \times \{\text{circle, rectangle}\}$. Given the vector of indicator variables $\mathbf{s}_i \in \mathbb{R}^6$, we construct the image X_i by randomly placing each of the shapes in the image such that no two shapes overlap. The caption for the image is then a string describing each of the shapes in the image, separated by a comma. For instance, an image containing a red-colored circle of radius 4 centered at (5, 10) and a blue-colored rectangle with the top-left corner at (20, 30) and bottom-right corner at (50, 60) would have the caption "red circle 4 (5, 10), blue rectangle ((20, 30) (50, 60))". We show examples of real and generated images of the synthetic shapes dataset in Fig. 10. We note that, while the diffusion model does not generate the correct locations for the shapes, this does not affect downstream classification results which do not rely on shape locations.

Diffusion Model. A diffusion model initialized on Stable Diffusion [30] was fine tuned for 105000 iterations on 107,000 images with a batch size of 16 at a 256x256 resolution and a learning rate of 1.0e-4. We fine-tuned only the U-Net and text-encoder of the model.

Concept Classifier. The concept classifier g was a CNN with 5 layers, each consisting of a convolution, batch norm, ReLU, and max pooling followed by a 3-layer multilayer perceptron that made six predictions for the presence of the six shapes. The model was trained on 50,000 images for 15 epochs.

Baselines. We generated Grad-CAM [32] and LIME [28] explanations for the predicted class of each image. The class prediction was determined by thresholding model predictions that maximized the true positive rate while minimizing the false positive rate across the test set. Each GradCAM heatmap was first converted to a binary mask by thresholding at the lowest non-zero value of the Grad-CAM heatmap. 5 features were used to generate each LIME mask. For every shape in the image, we calculated the intersection over union (IOU) between the shape and the explanation. Finally, we ranked shapes by their mean IOU across the entire test set.

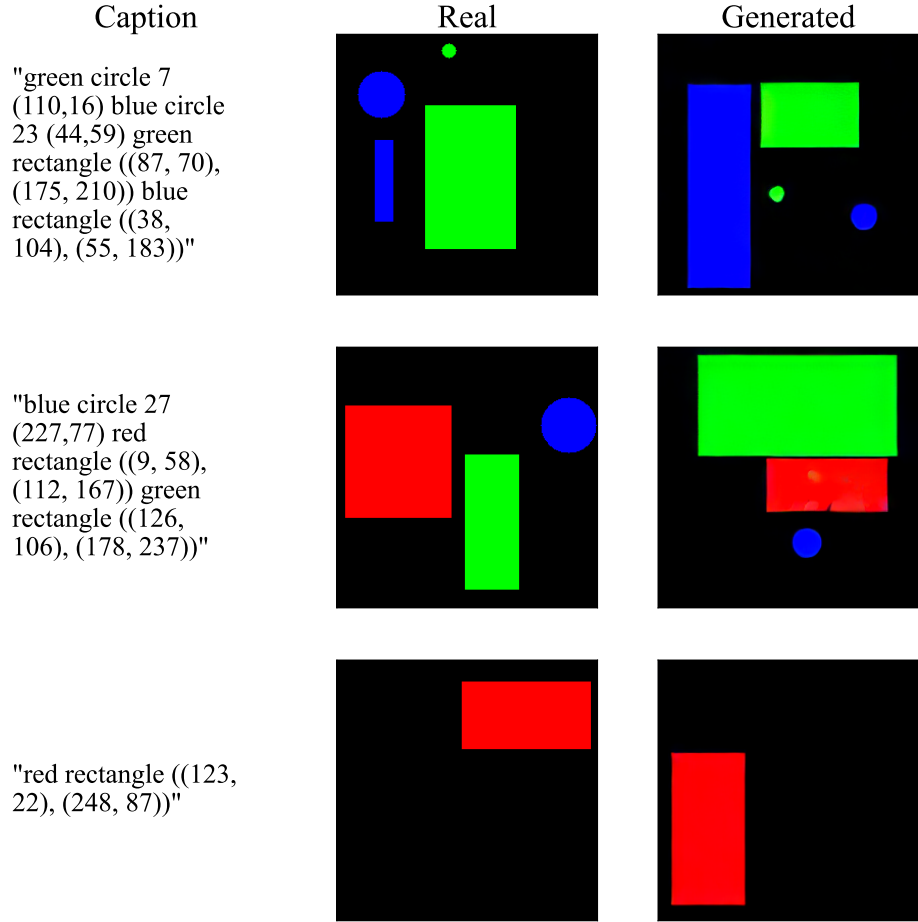


Fig. 10: Comparison between real and generated images for synthetic dataset. We compare real and generated images from the diffusion model conditioned on the original captions. We find that the generated images look realistic and reflect the shapes present in the captions.

Table 2: Effective generation validation for synthetic dataset models. We report AUROC (median, IQR) on both real and generated images for 100 target models on the synthetic dataset.

	AUROC (median, IQR)
Real Images	0.99 (0.98-1.0)
Generated Images	0.91 (0.84-0.94)

Table 3: Effective generation validation for synthetic dataset concept classifier. We show AUROC on both real and generated images for the concept classifier on the synthetic dataset across all six shapes. The concept classifier is able to detect all six shapes from the generated images with high accuracy.

	red circle	green circle	blue circle	red square	green square	blue square
Real Images	1.00	0.99	1.00	1.00	1.00	1.00
Generated Images	0.97	0.95	0.88	0.96	0.95	0.93

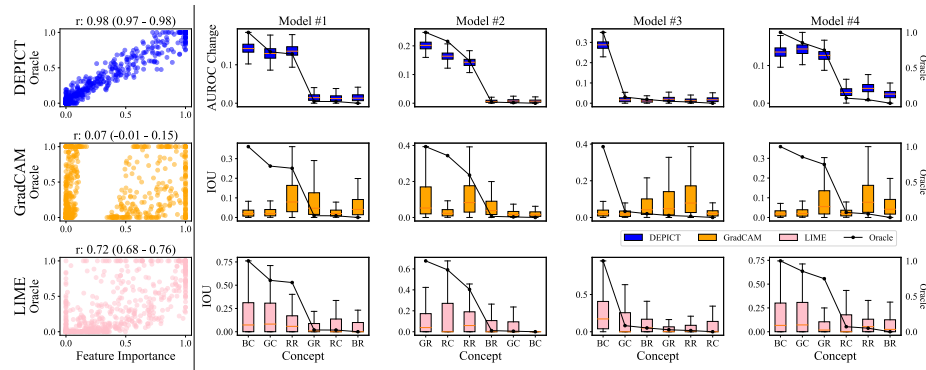


Fig. 11: Model feature importance across synthetic data models with the oracle generated by permuting concepts at the bottleneck. We compare the DEPICT ranking to GradCAM [32] and LIME [28]. Left: DEPICT has higher correlation with the standardized regression weights compared to GradCAM and LIME. Right: ranking generated for 4/100 randomly chosen classifiers. RC: red circle; BC: blue circle; GC: green circle; RR: red rectangle; BR: blue rectangle; GR: green rectangle.

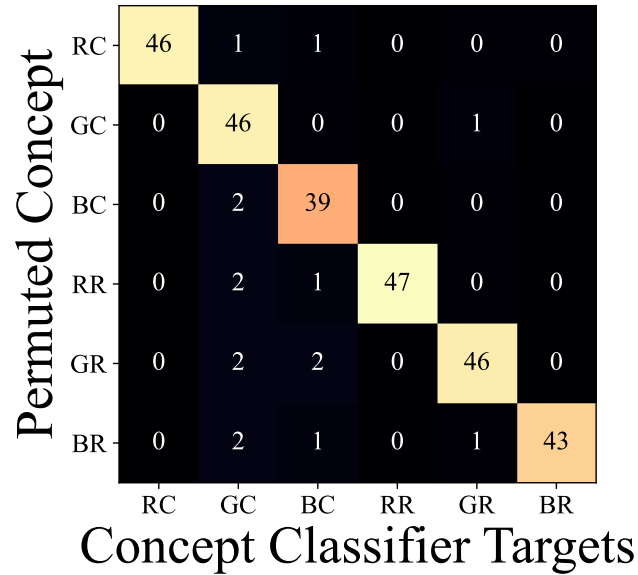



Fig. 12: Independent permutation validation for synthetic dataset. We report the average change in AUROC (unit = 0.01) of the concept classifier for the six shapes when permuting each individually for both real (oracle) and generated images. We observe permutation independence: a large change in performance when classifying permuted concepts, and minimal change in performance for unpermuted concepts. Colormap: 0  50. RC: red circle, BC: blue circle, GC: green circle, RR: red rectangle, GR: green rectangle, BR: blue rectangle.

9 COCO

9.1 Experiments

Dataset. COCO [23] contains 117k training and 4.5k validation images annotated with 80 object categories, which we consider to be concepts in the images. COCO also has 20k test images that are not labelled with object categories. Instead, we randomly sampled 10k images from the training set to use for test sets in downstream classification tasks, resulting in a final training set of 107k images. To caption each image, we disregarded the natural language captions corresponding to the images, and instead constructed new captions consisting of all the concepts in the images. E.g., if an image contained 2 persons and 1 couch, the corresponding caption is “2 person, 1 couch.” The 15 concepts used were: person, bottle, cup, bowl, chair, couch, bed, dining table, tv, laptop, remote, cell phone, oven, sink, and book. For downstream scene classification, we labelled each of the images using a ResNet trained on Places365 [43]. We mapped the scene label to one of six indoor labels from the MIT SUN Database [39]: shopping and dining, workplace (office building, factory, lab, etc.), home or hotel, transportation (vehicle interiors, stations, etc.), sports and leisure, and cultural (art, education, religion, military, law, politics, etc.).

Diffusion Model. We fine-tuned a Stable Diffusion [30] model for 1.34 million iterations with a batch size of 64 on COCO image-caption pairs at a 256x256 resolution and a learning rate of 1.0e-4. We fine-tuned only the U-Net and text-encoder of the model.

Concept Classifier. We fine-tuned a DenseNet-121 [11] pretrained on ImageNet [9] to predict the presence of the 80 objects in each image. The model was trained using stochastic gradient descent with momentum minimizing binary cross-entropy loss with a learning rate of 1.0e-1, momentum of 0.8, weight decay of 1.0e-4 and a batch size of 128. Early stopping based on validation loss with a patience of 5 was used after at least 8 training epochs. During training, images were reshaped such that their smaller axis was 256 pixels, and then center cropped along their longer axis to 256x256. Images were also randomly rotated up to 45 degrees, and vertically flipped with probability 0.3. We used ImageNet normalization across all experiments.

Primary feature models. We trained target classifiers on a binary task: *home or hotel* or *not*. We only considered images labelled with one of these two scene-level labels. Furthermore, for each of the target classifiers, we subsampled the data such that there was a 1:1 correlation between the presence of a *primary* concept (e.g., *person*) and the outcome. We trained 15 models using 15 concepts that were present in more than 5% of the data: person, bottle, cup, bowl, chair, couch, bed, dining table, tv, laptop, remote, cell, phone, oven, sink, and book. Models were trained with momentum 0.8, weight decay 1.0e-4, and learning rate of 1.0e-1. The best model was chosen as the epoch with the lowest validation loss. During training, images were reshaped such that their smaller axis was 256 pixels, and then center cropped along their longer axis to 256x256. Images were

also randomly rotated up to 45 degrees, and vertically flipped with probability 0.3. We used ImageNet normalization across all experiments.

Mixed feature models. We trained six total scene classifiers, where a model classifies if an image is one of six indoor scenes: (1) **shopping and dining**, (2) **workplace**, (3) **home or hotel**, (4) **transportation**, (5) **cultural**, and (6) **sports and leisure**. Here, we do not resample the training data to encourage the model to rely on specific concepts, but rather use the entire training set to let the model rely on any combination of concepts. Models were trained with momentum 0.8, weight decay $1.0e-4$, and a learning rate of $1.0e-1$. The best model was chosen as the epoch with the lowest validation loss. During training, images were reshaped such that their smaller axis was 256 pixels, and then center cropped along their longer axis to 256×256 . Images were also randomly rotated up to 45 degrees, and vertically flipped with probability 0.3. We used ImageNet normalization across all experiments.

Baselines. We generated Grad-CAM [32] and LIME [28] explanations for the predicted class of each image. The class prediction was determined by thresholding model predictions that maximized the true positive rate while minimizing the false positive rate of the validation set. Each GradCAM heatmap was first converted to a binary mask by thresholding at the lowest non-zero value of the Grad-CAM heatmap. 5 features were used to generate each LIME mask. For every object in the image, we calculated the intersection over union (IOU) between the object mask and the explanation. Finally, we ranked objects by their mean IOU across the entire test set.

Unconstrained primary feature models. In reality, we might want to explain a model that is not a concept bottleneck. Thus, we also trained primary feature models end-to-end. When the model is not constrained to a specific set of concepts, we want to observe that DEPICT still detects the primary feature as the most important concept in a classifier’s decisions.

9.2 Results

Unconstrained primary feature models. We compare three randomly selected unconstrained (trained end-to-end) primary feature model rankings generated by DEPICT to those generated by GradCAM and LIME in Fig. 17. DEPICT identifies the primary feature in all cases as significantly more important compared to the other concepts. While we do not have an oracle model to compare to in the unconstrained setting (as the model is not a CBM, and thus an oracle cannot be calculated), DEPICT’s results do align with the fact that we resampled the training data to encourage the models to focus on the primary feature. Note that these models use the same data as the original primary feature models, so the validation of assumptions (effective generation, independent permutation) hold for these models.

Table 4: Effective generation validation for COCO primary feature models. We show AUROC on both real and generated images for the primary feature models, each with one primary feature. The models are able to classify generated images with high AUROC and a maximum difference between the real and generated images of 0.09.

	Primary Feature Model														
	person	bottle	cup	bowl	chair	couch	bed	dining table	tv	laptop	remote	cell phone	oven	sink	book
Real Images	0.97	0.90	0.89	0.93	0.87	0.94	0.95	0.93	0.95	0.96	0.89	0.82	0.99	0.99	0.86
Generated Images	0.91	0.82	0.80	0.87	0.79	0.86	0.89	0.88	0.95	0.92	0.81	0.80	0.95	0.91	0.80

Table 5: Effective generation validation for COCO concept classifiers in primary feature models. We show AUROC on real and generated images for concept classifiers on COCO across all primary feature models and all concept classifier targets. The concept classifier is able to classify generated images with high AUROC and a maximum difference between the real and generated images of 0.07.

		Primary Feature Model									
		Person		Bottle		Cup		Bowl		Chair	
Concept Classifier	Target	Real	Gen	Real	Gen	Real	Gen	Real	Gen	Real	Gen
Person		0.95	0.92	0.97	0.97	0.97	0.96	0.97	0.97	0.97	0.97
Bottle		0.86	0.83	0.87	0.86	0.87	0.81	0.87	0.82	0.86	0.82
Cup		0.86	0.84	0.83	0.84	0.84	0.81	0.85	0.85	0.86	0.86
Bowl		0.91	0.88	0.88	0.88	0.91	0.89	0.93	0.89	0.92	0.91
Chair		0.86	0.87	0.86	0.86	0.87	0.88	0.86	0.87	0.86	0.82
Couch		0.91	0.92	0.91	0.89	0.92	0.90	0.90	0.91	0.91	0.90
Bed		0.94	0.92	0.92	0.93	0.90	0.91	0.90	0.91	0.94	0.95
Dining table		0.92	0.88	0.87	0.88	0.87	0.90	0.86	0.88	0.86	0.85
Tv		0.93	0.97	0.93	0.94	0.94	0.95	0.93	0.94	0.96	0.96
Laptop		0.97	0.96	0.97	0.97	0.97	0.96	0.97	0.96	0.96	0.97
Remote		0.86	0.86	0.91	0.89	0.91	0.87	0.89	0.86	0.91	0.87
Cell phone		0.88	0.87	0.88	0.88	0.89	0.88	0.87	0.88	0.84	0.88
Oven		0.98	0.92	0.98	0.94	0.98	0.92	0.98	0.95	0.97	0.95
Sink		0.97	0.92	0.98	0.96	0.97	0.95	0.97	0.94	0.98	0.97
Book		0.87	0.88	0.87	0.89	0.88	0.89	0.86	0.89	0.87	0.89
		Couch		Bed		Dining Table		Tv		Laptop	
Concept Classifier	Target	Real	Gen	Real	Gen	Real	Gen	Real	Gen	Real	Gen
Person		0.97	0.97	0.97	0.96	0.97	0.97	0.97	0.96	0.97	0.97
Bottle		0.87	0.82	0.86	0.80	0.86	0.83	0.87	0.82	0.88	0.82
Cup		0.87	0.84	0.85	0.86	0.86	0.86	0.86	0.86	0.86	0.87
Bowl		0.88	0.89	0.88	0.90	0.91	0.89	0.87	0.87	0.89	0.87
Chair		0.87	0.86	0.86	0.86	0.88	0.88	0.88	0.88	0.87	0.87
Couch		0.93	0.90	0.89	0.90	0.92	0.90	0.93	0.91	0.92	0.91
Bed		0.92	0.94	0.97	0.95	0.90	0.92	0.91	0.94	0.91	0.93
Dining table		0.86	0.89	0.87	0.87	0.93	0.89	0.86	0.87	0.86	0.90
Tv		0.94	0.95	0.94	0.94	0.94	0.95	0.96	0.96	0.95	0.95
Laptop		0.96	0.95	0.96	0.95	0.96	0.96	0.95	0.93	0.94	0.95
Remote		0.90	0.88	0.87	0.85	0.90	0.86	0.89	0.89	0.91	0.90
Cell phone		0.85	0.87	0.84	0.87	0.86	0.88	0.87	0.89	0.86	0.89
Oven		0.97	0.93	0.98	0.92	0.98	0.95	0.98	0.93	0.97	0.93
Sink		0.97	0.95	0.97	0.95	0.97	0.95	0.98	0.96	0.97	0.94
Book		0.87	0.89	0.86	0.88	0.87	0.88	0.88	0.87	0.88	0.90
		Remote		Cell Phone		Oven		Sink		Book	
Concept Classifier	Target	Real	Gen	Real	Gen	Real	Gen	Real	Gen	Real	Gen
Person		0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Bottle		0.87	0.82	0.89	0.83	0.88	0.83	0.87	0.80	0.88	0.83
Cup		0.88	0.87	0.87	0.87	0.86	0.85	0.86	0.82	0.87	0.86
Bowl		0.89	0.89	0.87	0.89	0.91	0.91	0.91	0.86	0.88	0.91
Chair		0.86	0.86	0.86	0.86	0.86	0.86	0.87	0.86	0.85	0.84
Couch		0.91	0.90	0.91	0.91	0.89	0.88	0.91	0.91	0.90	0.91
Bed		0.92	0.93	0.91	0.92	0.90	0.92	0.91	0.92	0.90	0.94
Dining table		0.86	0.90	0.88	0.90	0.87	0.89	0.88	0.89	0.87	0.89
Tv		0.95	0.96	0.95	0.96	0.93	0.94	0.94	0.94	0.95	0.95
Laptop		0.96	0.96	0.97	0.96	0.97	0.95	0.97	0.96	0.97	0.97
Remote		0.86	0.85	0.89	0.87	0.90	0.89	0.91	0.89	0.92	0.90
Cell phone		0.85	0.88	0.83	0.85	0.86	0.88	0.87	0.89	0.86	0.89
Oven		0.97	0.91	0.97	0.93	0.98	0.98	0.99	0.96	0.98	0.93
Sink		0.98	0.95	0.98	0.95	0.97	0.95	0.99	0.96	0.97	0.95
Book		0.88	0.88	0.87	0.89	0.86	0.86	0.88	0.87	0.85	0.85

Table 6: Effective generation validation for COCO mixed feature models.
 We show AUROC on both real and generated images for the mixed feature models.
 The differences in classification AUROC between real and generated images range from
 0.05 to 0.13 AUROC.

	Mixed Feature Model					
	shopping and dining	workplace	home or hotel	transportation	sports and leisure	cultural
Real Images	0.89	0.74	0.87	0.89	0.82	0.74
Generated Images	0.78	0.69	0.78	0.76	0.71	0.66

Table 7: Effective generation validation for COCO concept classifiers in mixed feature models. We show AUROC on real and generated images for concept classifiers on COCO for the mixed feature models and all concept classifier targets. The differences in classification AUROC between real and generated images range from 0.0 to 0.03 AUROC.

	Real Images	Generated Images
Person	0.97	0.97
Bottle	0.87	0.84
Cup	0.89	0.87
Bowl	0.91	0.90
Chair	0.89	0.88
Couch	0.94	0.93
Bed	0.97	0.96
Dining table	0.92	0.91
Tv	0.95	0.96
Laptop	0.97	0.96
Remote	0.95	0.92
Cell phone	0.89	0.88
Oven	0.98	0.96
Sink	0.98	0.95
Book	0.90	0.88

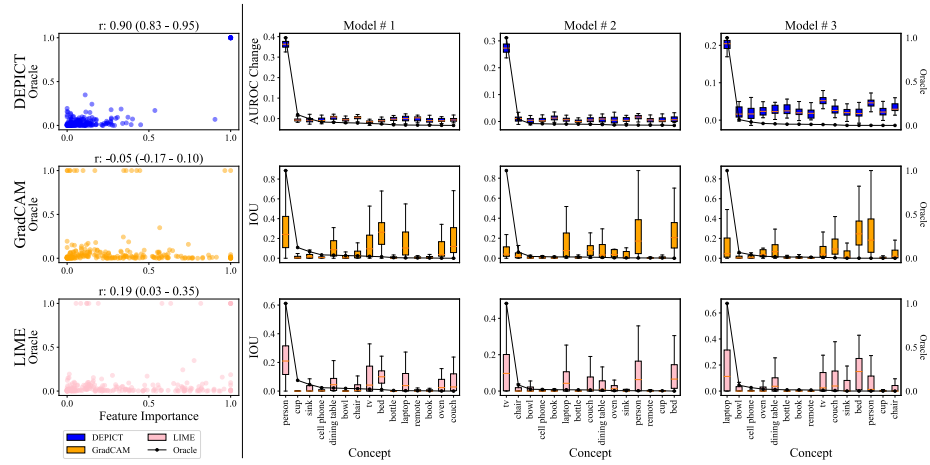


Fig. 13: Model feature importance across primary feature models with the oracle generated by permuting concepts at the bottleneck We compare the ranking produced by DEPICT to GradCAM and LIME, with the oracle generated by permuting concepts at the bottleneck. DEPICT has higher correlation with the oracle compared to LIME and GradCAM.

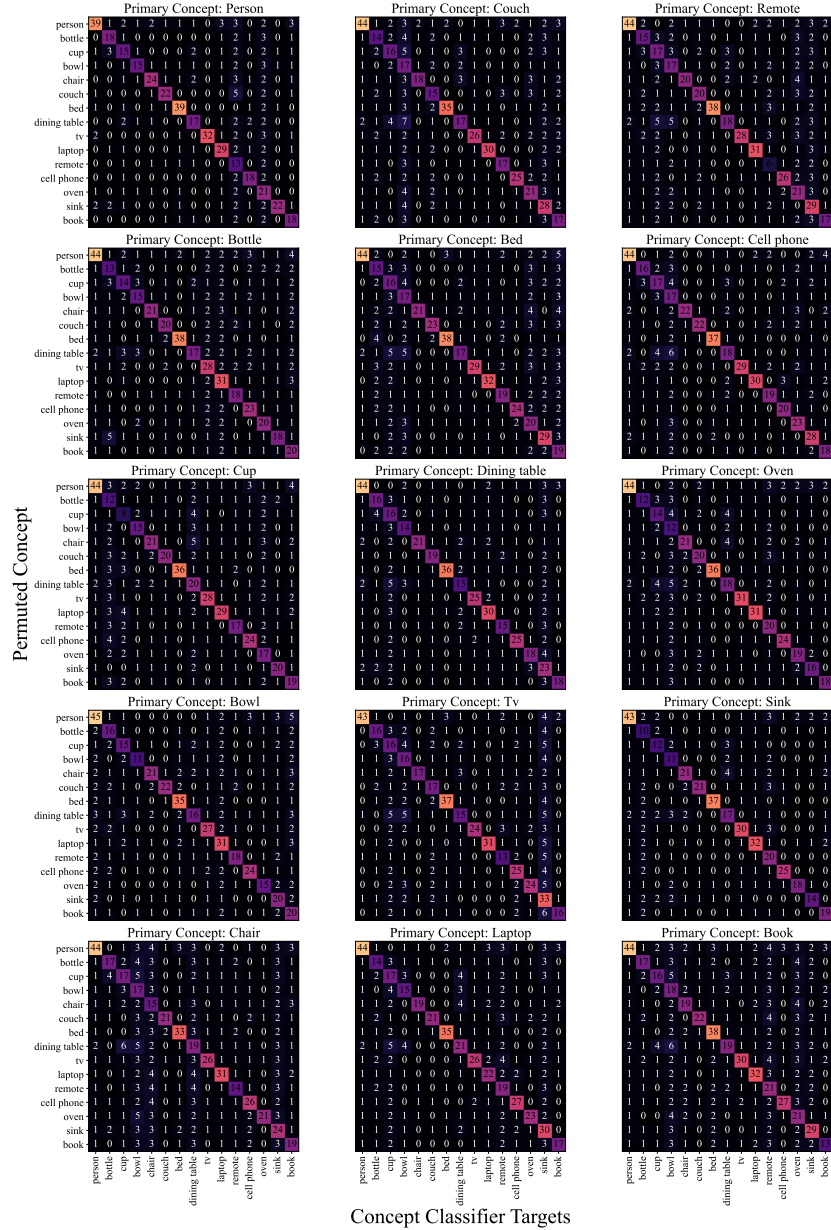


Fig. 15: Independent permutation validation for COCO primary feature models. We report the average change in AUROC (unit = 0.01) of the concept classifier for the COCO primary feature models when permuting each concept independently. We observe permutation independence: a large change in performance when classifying permuted concepts, and minimal change in performance for unpermuted concepts. Colormap: 0 50.

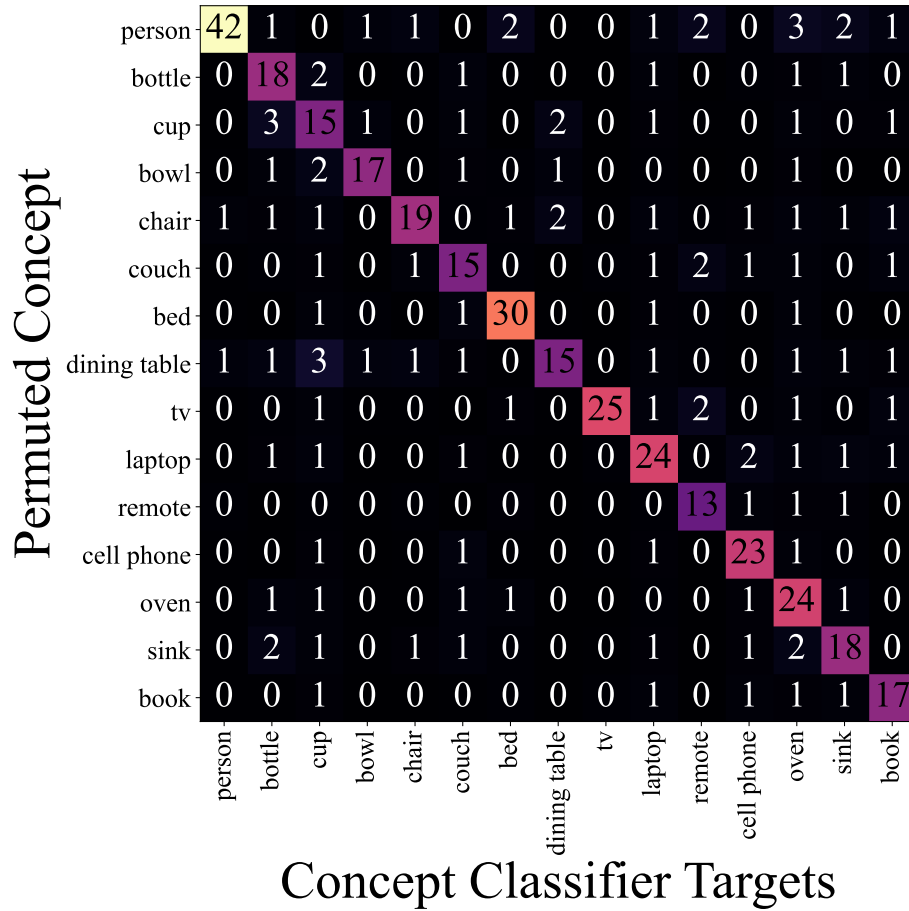


Fig. 16: Independent permutation validation for COCO mixed feature models. We report the average change in AUROC (unit = 0.01) of the concept classifier for the COCO mixed feature models when permuting each concept independently. We observe permutation independence: a large change in performance when classifying permuted concepts, and minimal change in performance for unpermuted concepts. Colormap: 0 50.

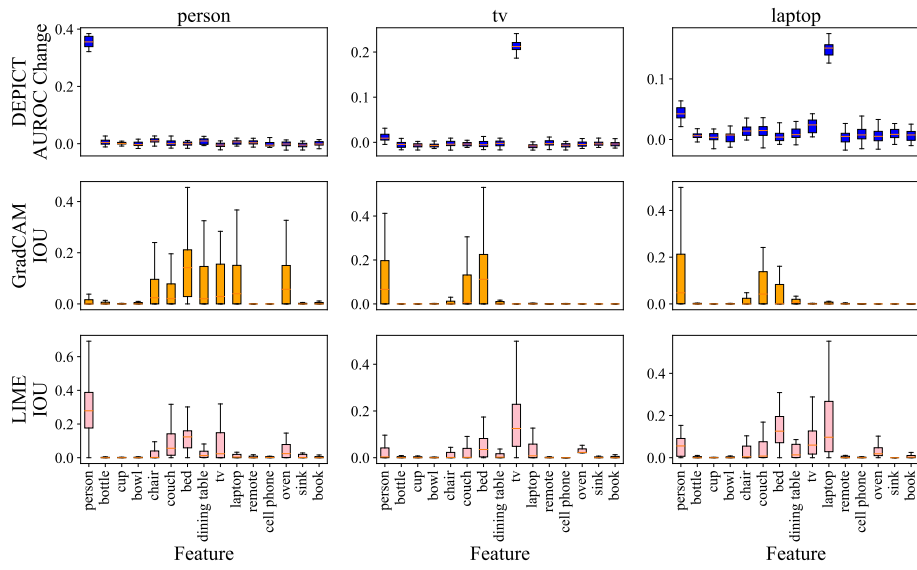


Fig. 17: Unconstrained primary feature model rankings. We compare three randomly selected unconstrained (trained end-to-end) primary feature model rankings generated by DEP ICT to those generated by GradCAM and LIME. DEP ICT identifies the primary feature in all cases as significantly more important compared to the other concepts.

10 MIMIC-CXR

10.1 Experiments

Dataset. MIMIC-CXR [14, 16] consists of 242,479 frontal chest X-rays with corresponding radiology reports. We split the data into 193706/24549/24224 images for training, validation, and test sets. To construct a final caption for each image, we extracted demographic information corresponding to the patients’ body mass index (BMI), age, and sex at the time the chest X-ray was taken, and prepended these information to the radiology report corresponding to the chest X-ray. We subsampled the data for downstream tasks where we injected a 1:1 correlation between pneumonia and each primary features: bmi, age, or sex.

Diffusion Model. A diffusion model initialized on Stable Diffusion [30] with the text encoder replaced with publicly available clinical BERT embeddings [3] was fine tuned on chest X-ray/radiology report pairs for 295569 iterations on a batch size of 16 at a 256x256 resolution with a learning rate of 1.0e-4. We fine-tuned only the U-Net and text-encoder of the model.

Target Models. We trained target classifiers to predict the presence of pneumonia. We trained the classifier on top of the concept classifier. During training, images were reshaped such that their smaller axis was 256 pixels, and then randomly cropped along their longer axis to 256x256. Images were also randomly rotated up to 15 degrees. We used ImageNet normalization across all experiments.

Concept Classifier. We fine-tuned a DenseNet-121 [11] pretrained on ImageNet [9] to learn the presence of radiological findings and the three permutable concepts: bmi, age, sex, enlarged cardiomeastinum, cardiomegaly, lung opacity, lung lesion, edema, consolidation, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, and support devices. The model was trained for three epochs using stochastic gradient descent with momentum minimizing binary cross-entropy loss with a learning rate of 1.0e-4, momentum of 0.8 and a batch size of 32. During training, images were reshaped such that their smaller axis was 256 pixels, and then randomly cropped along their longer axis to 256x256. Images were also randomly rotated up to 15 degrees. We used ImageNet normalization across all experiments.

Validation of assumptions. For effective generation, we measure the difference in target model performance between real and generated images. If the target model performs well on generated images, then we can measure concept classifier performance on specific concepts in the images that we wish to permute. For concepts that we are *not* permuting, we might not need to validate effective generation depending on the scenario: (1) *Target model does not pass the checks of effective generation.* Consider a concept that we are not permuting in text space (e.g., Pleural effusion on the chest X-ray). If the target model performance drops on the generated images, then we might want to investigate why. In this setting, it would be useful to look at granular changes in model performance via the concept classifier to know if specific concepts are not being generated well, and thus contributing to poor target model performance. In any case, since the

target model does not pass the checks of effective generation, we would not apply DEPICT since the target model does not pass the checks of effective generation. (2) *Target model does pass the checks of effective generation.* Again, consider a concept that we are not permuting in text space (e.g., Pleural effusion on the chest X-ray). If the target model passes the checks of effective generation, but the concept classifier performs poorly in detecting a specific concept that we are not permuting, we can still apply DEPICT. This is because we know the model must *not* be relying on the non-permutable concept, since the target model can still classify the generated images well. Furthermore, we do not need to generate a reference performance for the non-permutable concept since we are not permuting it, nor ranking it against other concepts.

Independent permutation. We note that it is still useful to measure independent permutation on non-permutable concepts. This way, we can check to make sure that when permuting a concept (such as age, bmi, or sex), any resulting change in target model performance is not confounded by other changes on the image.

Results. Here, we further discuss results of DEPICT on MIMIC-CXR.

Validation of assumptions. All three target models are able to accurately classify the generated images (Table 8). Similarly, the concept classifier performs well on both real and generated images for all three demographic concepts: bmi, age, and sex (Table 9). We also measured concept classifier performance on radiological findings, finding that the concept classifier performs well across most concepts in the images (Table 9), but struggles on a few such as detecting lung lesions (AUROC drop = 0.31) and pneumothorax (AUROC drop = 0.23). Again, we note that we can still apply DEPICT to these settings, as we are not permuting such concepts on the images and the target models still classify the generated images well, even without being able to detect concepts such as lung lesions and pneumothorax (target model AUROC > 0.85).

In terms of independent permutation, when permuting age, bmi, and sex, we observe some changes in concept classifier performance when detecting concepts such as lung opacity and lung lesion (Fig. 18). Thus, one must proceed with caution when interpreting the results of DEPICT. When permuting one of the three concepts, we can conclude that the model relies on each of the three primary features in some way - either directly, or by correlation with other concepts such as lung opacity and lung lesion.

Table 8: Effective generation validation for MIMIC models. We show AUROC on both real and generated images for the MIMIC models. The differences in classification AUROC between real and generated images range from 0.0 to 0.04 AUROC.

	Primary Feature		
	BMI	Age	Sex
Real Images	0.89	0.98	1.0
Generated Images	0.85	0.96	1.0

Table 9: Effective generation validation for MIMIC concept classifiers. We show AUROC on real and generated images for concept classifiers on MIMIC across all concept classifier targets.

Concept Classifier Target	Primary Feature					
	BMI		Age		Sex	
	Real	Gen	Real	Gen	Real	Gen
BMI	0.94	0.91	0.95	0.91	0.95	0.90
Age	0.98	0.95	0.98	0.97	0.98	0.96
Sex	1.00	1.00	1.00	1.00	1.00	1.00
Enlarged Cardiomeastinum	0.85	0.84	0.86	0.74	0.84	0.72
Cardiomegaly	0.91	0.85	0.92	0.86	0.91	0.82
Lung Opacity	0.70	0.66	0.79	0.71	0.69	0.60
Lung Lesion	0.88	0.68	0.92	0.78	0.95	0.64
Edema	0.95	0.82	0.95	0.84	0.93	0.83
Consolidation	0.91	0.81	0.91	0.83	0.91	0.85
Atelectasis	0.70	0.57	0.81	0.58	0.81	0.57
Pneumothorax	0.96	0.79	0.97	0.89	0.96	0.87
Fracture	0.88	0.68	0.85	0.69	0.87	0.72

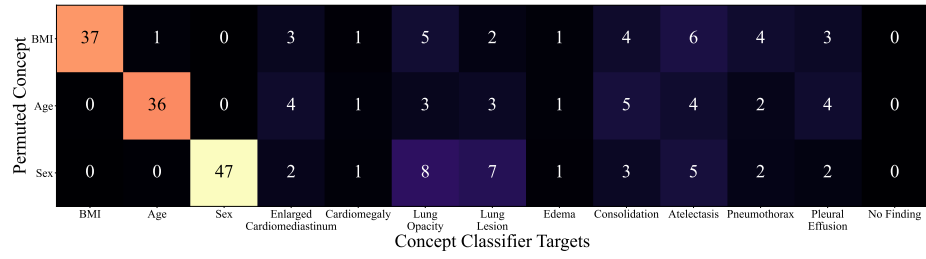



Fig. 18: Independent permutation validation for MIMIC-CXR. We report the average change in AUROC (unit = 0.01) of the concept classifier for the MIMIC concepts when permuting each one independently. When permuting bmi, age, and sex, we observe changes to the concept classifier’s ability to detect other radiological findings such as lung opacity and lung lesion. Thus, the importance of these three demographic concepts could be due, in part, to changes in the presence of other findings on the chest X-ray. Colormap: 0  50.