

427 APPENDIX

428 A Loss Functions

429 A.1 DDPM optimization

430 The loss function for the diffusion model consists of multiple components. First, we train the
 431 discriminator \mathcal{D} to distinguish between the real low-resolution ground truth image-mask pair \mathbf{x}_{lr}^{gt} and
 432 the denoised image-mask pair $\hat{\mathbf{x}}_t$. The objective for the discriminator is to maximize the expectation
 433 of $\mathcal{D}(\mathbf{x}_{lr}^{gt}) = 1$ and $\mathcal{D}(\hat{\mathbf{x}}_t) = 0$. Once trained, the discriminator is frozen. The diffusion model’s
 434 objective is to maximize the expectation of the clean image mask pair \mathbf{x}_0 given the noisy image
 435 mask pair \mathbf{x}_t , the text embedding \mathbf{T}_{emb} , the condition class embedding \mathbf{C}_{emb} , and the timestep t . The
 436 standard diffusion loss is:

$$\mathcal{L}_{diff} = \|\epsilon - \epsilon_\theta(x_t, t, \mathbf{T}_{emb}, \mathbf{C}_{emb})\|^2 \quad (4)$$

437 where $\theta = [\theta_T, \theta_{CE}, \theta_U]$. Additionally, we include a triplet loss to ensure that the positive condition
 438 class embedding \mathbf{C}_{emb} is brought closer to the anchor timestep embedding \mathbf{T}_{emb} , while a random
 439 permutation $\tilde{\mathbf{C}}_{emb}$ of \mathbf{C}_{emb} is considered negative and is pushed away from the anchor by a margin of
 440 1. The triplet loss can be expressed as:

$$\mathcal{L}_{trip} = \max(0, \|\mathbf{T}_{emb} - \mathbf{C}_{emb}\|^2 - \|\mathbf{T}_{emb} - \tilde{\mathbf{C}}_{emb}\|^2 + 1) \quad (5)$$

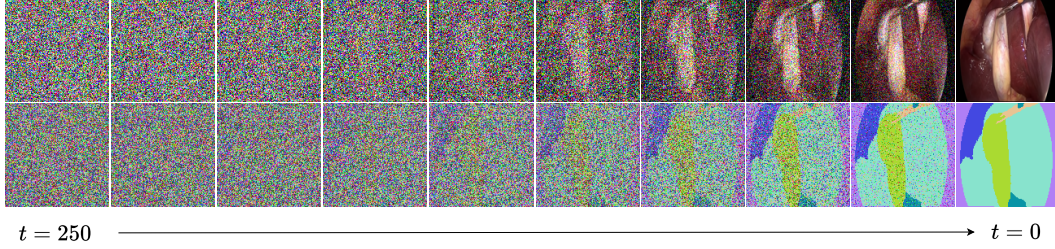


Figure 7: Predicted outputs during reverse diffusion process using CholecSeg8k dataset.

441 A.2 Adversarial learning

442 A discriminator, denoted as \mathcal{D} , is used to regularize the generation of image-mask pairs by distin-
 443 guishing between real and denoised pairs. The discriminator is trained on real image-mask pairs
 444 versus denoised ones at a sampled timestep t .

445 Let $\hat{\mathbf{X}}_t$ represent the denoised image-mask pair at timestep t . The discriminator \mathcal{D} is a convolution-
 446 linear model, which processes input pairs of dimension $\mathbb{R}^{h \times w}$ and produces an output $\mathcal{D}(\hat{\mathbf{x}}_t) \in \mathbb{R}^{1 \times 1}$,
 447 where the output range is constrained to $(0, 1)$ such that $\mathcal{D} : \mathbb{R}^{h \times w} \rightarrow (0, 1)$. The training objective
 448 for \mathcal{D} is to maximize the likelihood of correctly classifying real image-mask pairs as 1 and denoised
 449 pairs as 0. The adversarial loss encourages the denoised image-mask pair $\hat{\mathbf{x}}_t$ to be classified as real
 450 by the discriminator:

$$\mathcal{L}_{adv} = 1 - \mathcal{D}(\hat{\mathbf{x}}_t) \quad (6)$$

451 **Combined loss:** The total loss for the diffusion model is the sum of the diffusion loss, the triplet loss,
 452 and the adversarial loss.

453 We multiply the adversarial loss by a regularization factor $\beta = 0.1$ to control its influence:

$$\mathcal{L}_{total} = \mathcal{L}_{diff} + \mathcal{L}_{triplet} + \beta \cdot \mathcal{L}_{adv} \quad (7)$$

454 Thus the combined loss in 7 optimizes the parameters in θ for conditional image generation using 4,
 455 text and class alignment using 5. The loss is also regularised using a regularized factor of 6.

A.3 Super resolution optimization

The super-resolution model \mathcal{SR} is optimized by minimizing a combination of Mean Squared Error (MSE) loss and perceptual loss on the predicted and ground truth image-mask pairs. Let $\mathbf{X}_{\text{HR}}^{\text{gt}}$ denote the high-resolution ground truth image-mask pair, and $\hat{\mathbf{X}}_{\text{HR}}$ denote the high-resolution prediction from the super-resolution model.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N \left(\mathbf{X}_{\text{HR}}^{\text{gt}} - \hat{\mathbf{X}}_{\text{HR}} \right)^2 \quad (8)$$

In addition to mean squared error (MSE) from 8, we incorporate a perceptual loss to capture high-level visual features, improving the quality of texture and structure in the generated high-resolution output. We compute the perceptual loss $\mathcal{L}_{\text{perc}}$ by measuring the similarity between the VGG-encoded feature maps of the ground truth and predicted images at level 6 using 9:

$$\mathcal{L}_{\text{perc}} = \|\text{VGG}_6(\mathbf{X}_{\text{HR}}^{\text{gt}}) - \text{VGG}_6(\hat{\mathbf{X}}_{\text{HR}})\|^2 \quad (9)$$

The total loss in 10 is used for optimizing the super-resolution model \mathcal{S} is the weighted sum of MSE and perceptual losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{perc}} \quad (10)$$

where λ is a weighting factor to balance the contribution of the perceptual loss.

B Semantic FID Metrics for Evaluating Class and Boundary Alignments

To evaluate the semantic consistency and spatial alignment between generated images and their associated segmentation masks, we adopt the Semantic Fréchet Inception Distance (sFID) [3], originally proposed for assessing anatomical correctness of image generation in surgical scenes. We generalize this metric to evaluate class-aware visual fidelity across diverse domains including COCO, BTCV, MBRSC, CholecSeg8K, Cityscapes, and Pascal VOC. Unlike traditional FID, which assesses global distribution similarity between real and generated images, sFID computes class-wise distances by isolating and comparing specific semantic regions, providing a more fine-grained evaluation.

Motivation. In image-label generation tasks, especially in dense or multi-class scenes, it is critical not only that images look realistic, but also that each semantic entity (e.g., liver, gallbladder, road, vehicle, etc.) is correctly rendered in its location and appearance. sFID addresses this by using segmentation masks to isolate semantic regions and computing the FID per region. This allows us to evaluate:

- *Boundary alignment*: whether the visual features inside a region match across real and generated domains.
- *Class assignment correctness*: whether semantic content appears in the correct region.
- *Entity presence*: whether the conditioned classes actually appear in the output.

This allows sFID to explicitly penalize boundary misalignments, class confusion, or missing object classes — challenges that are often underreported by global fidelity metrics.

Mathematical Formulation. Let $C = \{c_1, c_2, \dots, c_K\}$ be the set of semantic classes present in the dataset. For each class $c_k \in C$, we extract real and generated image regions using their respective binary segmentation masks:

$$I_r^{(k)} = I_r \odot M_r^{(k)}, \quad I_g^{(k)} = \hat{I}_g \odot \hat{M}_g^{(k)},$$

where $M_r^{(k)}$ and $\hat{M}_g^{(k)}$ are binary masks for class c_k in real and generated images, respectively, and \odot denotes pixel-wise multiplication. This isolates the region of interest corresponding to c_k .

These class-specific patches are passed through a pre-trained visual encoder $\phi(\cdot)$ (typically Inception-V3 or VGG16) to obtain feature representations. For class c_k , let the features from real and generated samples be:

$$\mathcal{F}_r^{(k)} = \{\phi(I_{r,1}^{(k)}), \dots, \phi(I_{r,n_k}^{(k)})\}, \quad \mathcal{F}_g^{(k)} = \{\phi(I_{g,1}^{(k)}), \dots, \phi(I_{g,m_k}^{(k)})\}.$$

We compute the empirical means $\mu_r^{(k)}, \mu_g^{(k)}$ and covariances $\Sigma_r^{(k)}, \Sigma_g^{(k)}$ for each class c_k , and define the class-wise Fréchet Distance as:

$$\text{FID}^{(k)} = \left\| \mu_r^{(k)} - \mu_g^{(k)} \right\|_2^2 + \text{Tr} \left(\Sigma_r^{(k)} + \Sigma_g^{(k)} - 2 \left(\Sigma_r^{(k)} \Sigma_g^{(k)} \right)^{1/2} \right).$$

The overall Semantic FID is obtained by averaging valid class-wise scores:

$$\text{sFID} = \frac{1}{K'} \sum_{k=1}^{K'} \text{FID}^{(k)},$$

where $K' \leq K$ is the number of classes present in both the generated and real masks (classes absent in either domain are excluded to maintain stability).

Implementation Details.

- Masks are applied directly to RGB images using pixel-wise multiplication, optionally followed by cropping to reduce background influence.
- Extracted patches are resized (e.g., to 299×299) before feature extraction.
- Classes with fewer than a minimum number of samples (e.g., $N < 10$) may be ignored to avoid unstable statistics.
- For text-conditioned generation, the evaluation can be restricted to classes explicitly mentioned in the prompt.

Generalization Across Domains. Though sFID was originally introduced for evaluating synthetic surgical content, the metric is domain-agnostic. In our work, we apply sFID to datasets with diverse semantic complexity—from anatomical regions (BTCV, MBRSC) and surgical instruments (CholecSeg8K) to everyday objects and urban scenes (COCO, Cityscapes, Pascal VOC)—offering a unified protocol for assessing image-mask generation quality across tasks.

C Ablation Studies

C.1 Effect of Triplet and Discriminator Losses

CoSimGen utilizes two key strategies in model training: (1) a triplet loss that enforces spatio-spectral conditioning aligned with text embeddings, and (2) a discriminator loss that regularizes the output distribution. To assess their individual and combined contributions, we conduct an ablation study using the MBRSC dataset. Results in Fig. 8 show that the triplet loss improves class-conditioned FID by encouraging better alignment between textual and visual representations. The discriminator loss further enhances fidelity by ensuring generated outputs resemble real data distributions. When combined, these two losses yield the best performance, demonstrating their complementary effects on semantic alignment and realism.

C.2 Latent Space Visualization

We further visualize the latent space behavior under different loss configurations. In Fig. 8, blue and orange dots represent class and text conditions, respectively. When both triplet and discriminator losses are active, class vectors form semantically meaningful clusters guided by textual embeddings. This reflects improved alignment between class and text modalities, enabling the model to learn coherent associations and generate contextually consistent outputs.

C.3 Class-Condition Co-occurrence Analysis

We analyze the statistical relationship between class labels and conditional vectors using co-occurrence matrices shown in Figs. 9 and 10. High co-occurrence values indicate the inherent joint presence of certain classes under specific conditions.

CholecSeg8k (Fig. 9a) presents a dense co-occurrence matrix, as organs such as liver and gallbladder dominate large image regions. Consequently, visual fidelity offers a more informative metric than

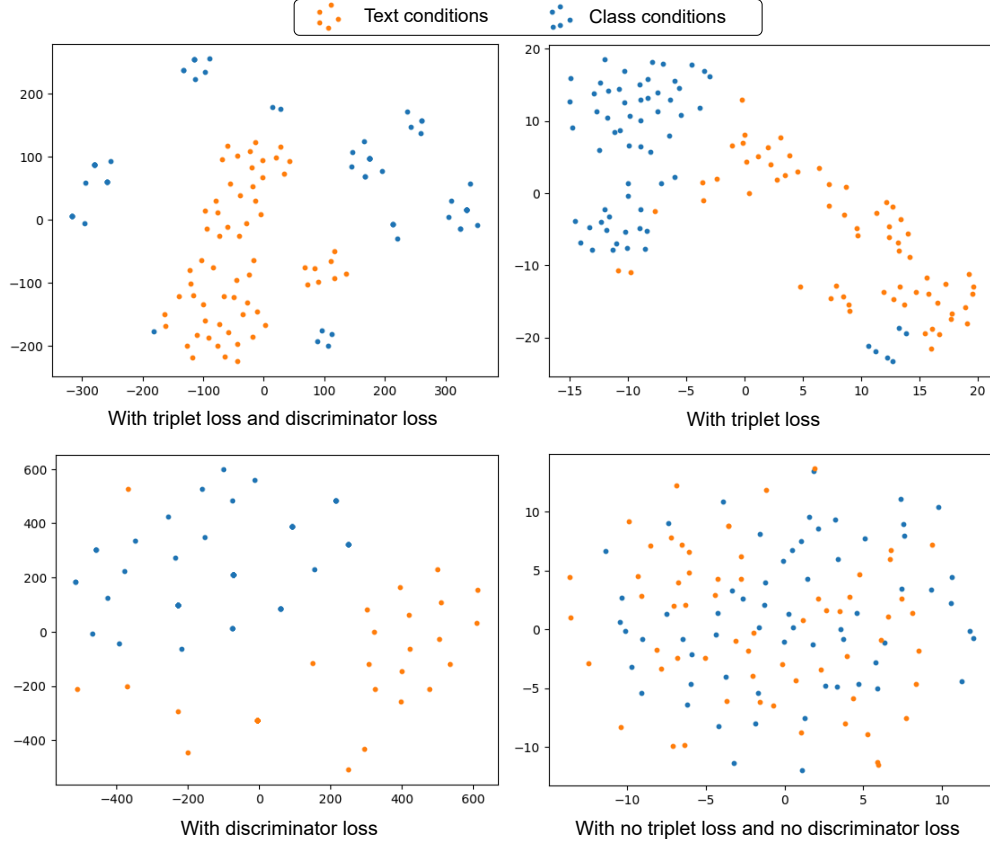


Figure 8: Ablation on the contributions of using discriminator loss and triplet loss on the MBRSC dataset.

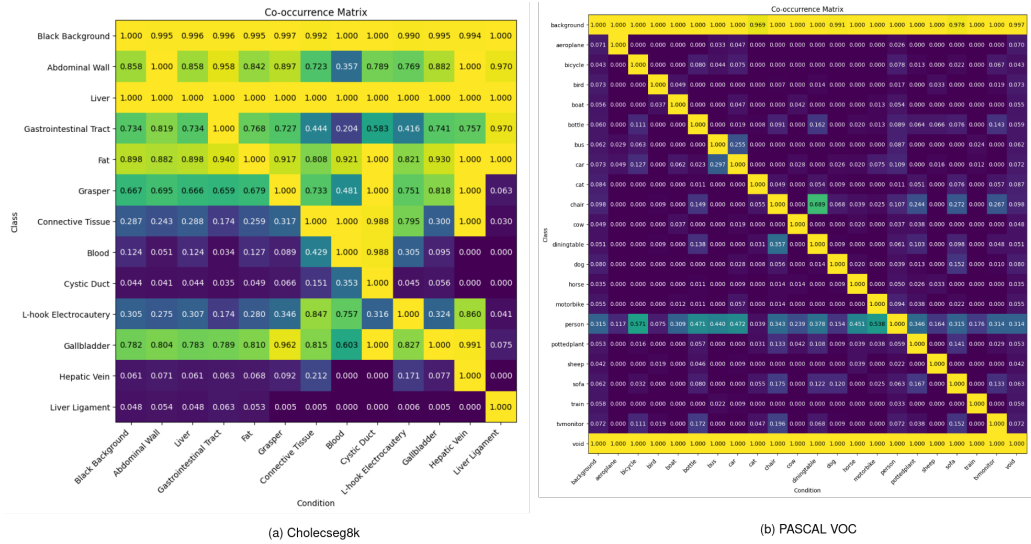


Figure 9: Class-condition co-occurrence matrices for CholecSeg8k and PASCAL VOC. CholecSeg8k exhibits dense co-occurrence among organ classes such as liver, gallbladder, and fat, whereas PASCAL VOC shows a sparse co-occurrence pattern.

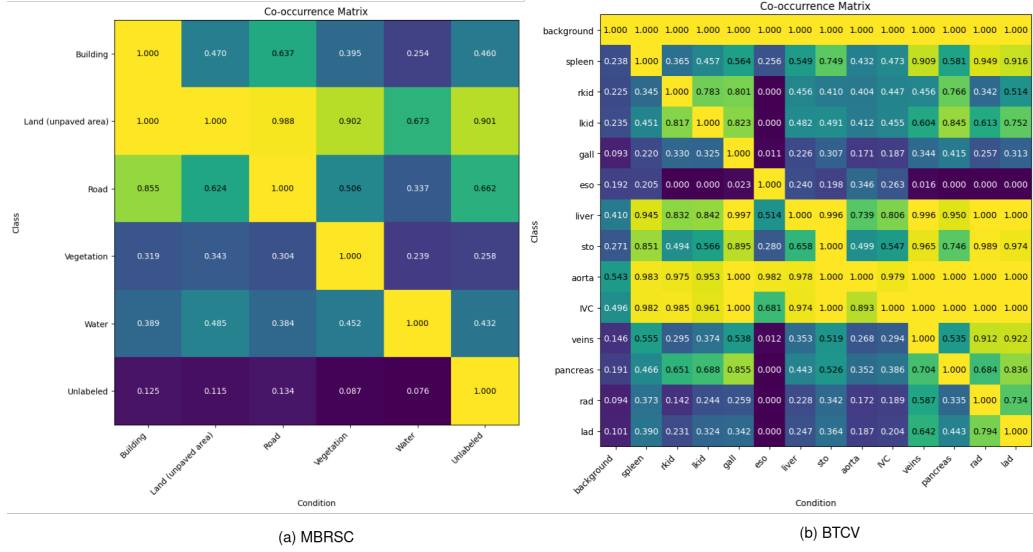


Figure 10: Class-condition co-occurrence matrices for MBRSC and BTCV. BTCV, being derived from 3D volume slices, shows consistent co-occurrence of anatomical structures like the aorta and IVC, albeit with smaller footprints.

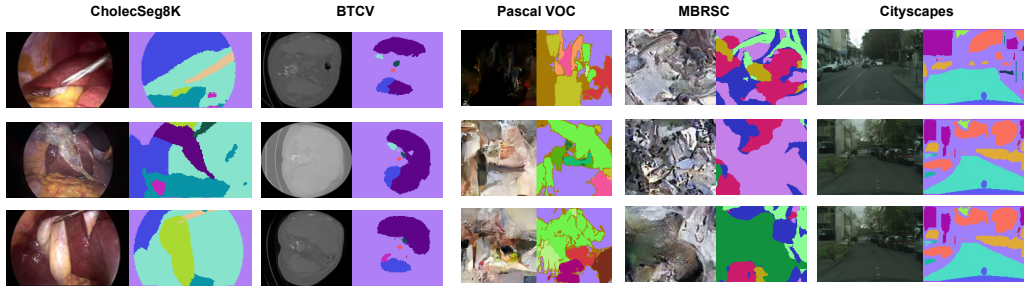


Figure 11: More samples of generated low-resolution image-mask pairs across five datasets.

535 semantic FID, since generated features tend to overlap significantly with real images, diminishing
 536 the utility of FID, KID, or VGG-based scores. In contrast, PASCAL VOC (Fig. 9b) has a sparse
 537 co-occurrence structure, making generalization more challenging due to limited class overlap.

538 MBRSC and BTCV (Fig. 10) show moderate co-occurrence densities. BTCV’s anatomical classes
 539 like the aorta and IVC appear together frequently, but occupy small image regions due to slice-wise
 540 representation. This balance in class-condition overlap facilitates better convergence and training
 541 stability.

542 D More Qualitative Results

543 In addition to the low-resolution (LR) qualitative images presented in Figs. 11 and 12, we provide
 544 high-resolution outputs generated by our proposed model. Fig. 11 shows more qualitative results
 545 generated by CoSimGen across five explored datasets. In Fig. 12, we present additional qualitative
 546 results of CoSimGen on the MBRSC dataset [16], highlighting its ability to generate satellite images
 547 that accurately reflect the semantic classes queried in the text prompts. These results demonstrate
 548 CoSimGen’s capability to conditionally align the generated content with specified semantic details,
 549 effectively capturing and representing features such as buildings, roads, water bodies, vegetation, and
 550 other land use patterns. This level of precision is particularly significant for applications requiring
 551 detailed spatial representations, such as urban planning or environmental monitoring.

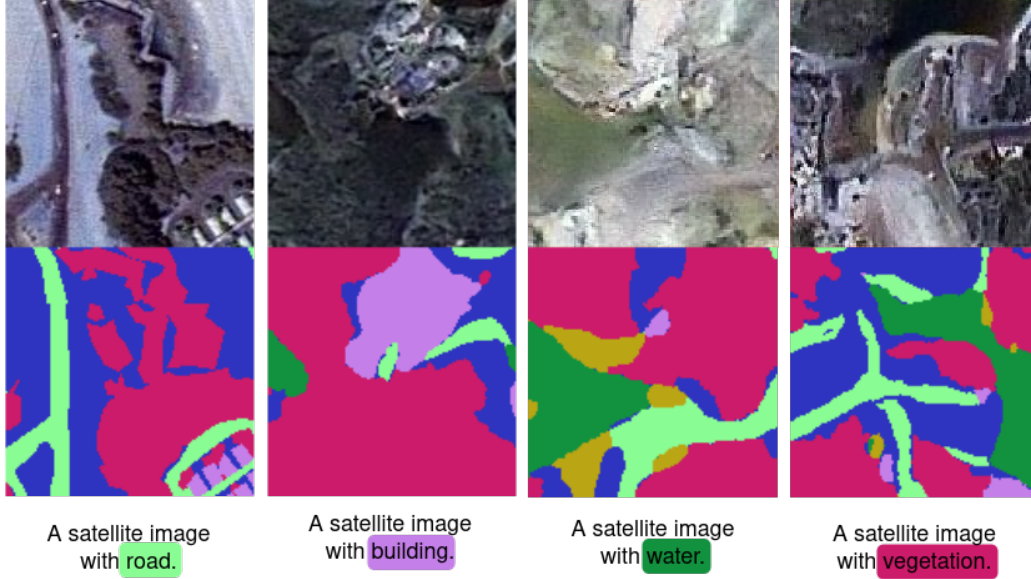


Figure 12: Additional qualitative results on the MBRSC dataset showcasing CoSimGen’s image-mask outputs generated based on text prompts describing various semantic classes. The displayed images are cropped to a resolution of 128×128 owing to the extensive size of satellite imagery. We observe that the class corresponding to the queried prompt is always present in the generated semantic mask.

Fig. 13 showcases qualitative outputs of CoSimGen on the Cholecseg8k dataset, including SR images and their corresponding segmentation masks. The SR images (512×512) are resized to actual resolution (480×854) of images in the dataset for improved visualization. CoSimGen’s generated mask captures the presence of the semantic class queried in the input prompt.

CoSimGen demonstrates the ability to generate high-fidelity image-mask pairs that align with user-provided prompts. The generated masks consistently include the semantic class specified in the prompt while also reasoning contextually to incorporate additional relevant and complementary classes, resulting in more meaningful and realistic outputs. Non-required classes for downstream usage can be easily removed via simple postprocessing, as the mask’s class identities are directly mappable to the dataset’s class names.

Fig. 14 showcases extensive qualitative results on the BTCV dataset [10]. The super-resolution (SR) images produced by ESPCNN [31], utilized in our proposed CoSimGen framework, are compared with baseline outputs from SRGAN [20]. The results demonstrate that ESPCNN effectively captures high-frequency details that SRGAN fails to reproduce. This distinction is particularly evident in the sharper boundaries between textures, such as those of organs, bones, and blood vessels, highlighting ESPCNN’s superior ability to preserve structural details.

D.1 Societal Impacts

CoSimGen introduces a generative framework for producing paired images and segmentation masks, aiming to reduce the dependence on labor-intensive manual annotations—particularly in domains such as geospatial and medical imaging where expert labeling is costly and time-consuming. This capability has the potential to broaden access to training data for under-resourced research institutions, lower entry barriers for translational AI development, and support education and prototyping in global healthcare research communities. By supplementing existing datasets with synthetic yet conditionally aligned examples, CoSimGen could help mitigate data scarcity and class imbalance, promoting more inclusive algorithmic development.

However, the societal benefits must be balanced against several risks. The synthetic nature of the generated data, while visually and semantically plausible, may introduce subtle inaccuracies or artifacts not immediately detectable by automated metrics. In sensitive applications like medical decision support or surgical education, reliance on unvalidated synthetic content may inadvertently

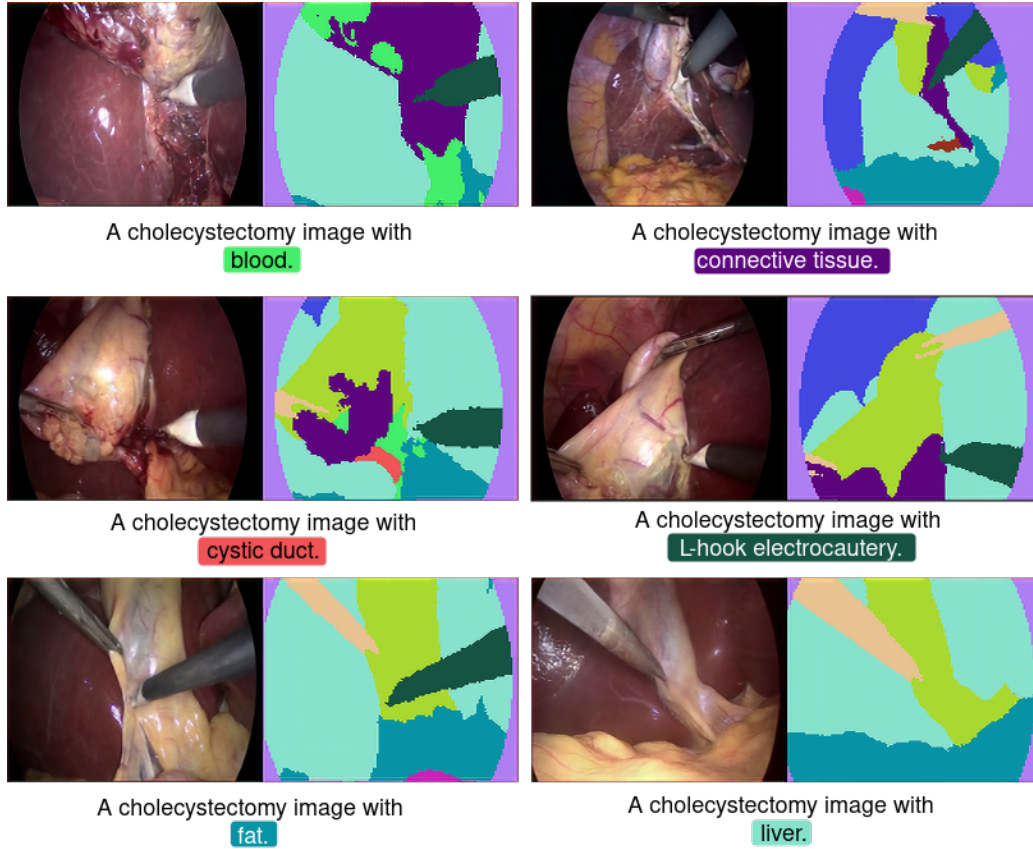


Figure 13: Qualitative results of CoSimGen on Cholecseg8k dataset. We observe that major portions of the image is covered by liver/ gallbladder along with the prompted class. Further we observe that the queried class is semantically present in the generated mask and image.

propagate errors or misrepresent clinical features, undermining safety and trust. Expert oversight is therefore essential to validate synthetic outputs before integration into downstream pipelines.

Moreover, the use of generative models in clinical or regulatory contexts raises concerns around transparency, traceability, and reproducibility. Clear labeling of synthetic data and provenance tracking are necessary to prevent confusion with real patient data and to maintain integrity in training, evaluation, and auditing processes. Although CoSimGen does not train on identifiable human data, applications involving broader datasets must be evaluated for privacy risks and potential misuse.

Additionally, like all generative systems, CoSimGen may amplify existing biases present in the training distribution or introduce new artifacts that disproportionately affect underrepresented cases. Responsible deployment thus requires adherence to ethical guidelines, rigorous fairness audits, and domain-specific validation to ensure that benefits are equitably distributed and harms are minimized.

To conclude, CoSimGen offers meaningful potential to democratize data access and accelerate innovation, particularly in fields constrained by annotation bottlenecks. Its societal value hinges on responsible use, expert involvement, and transparent communication about its synthetic nature and limitations.

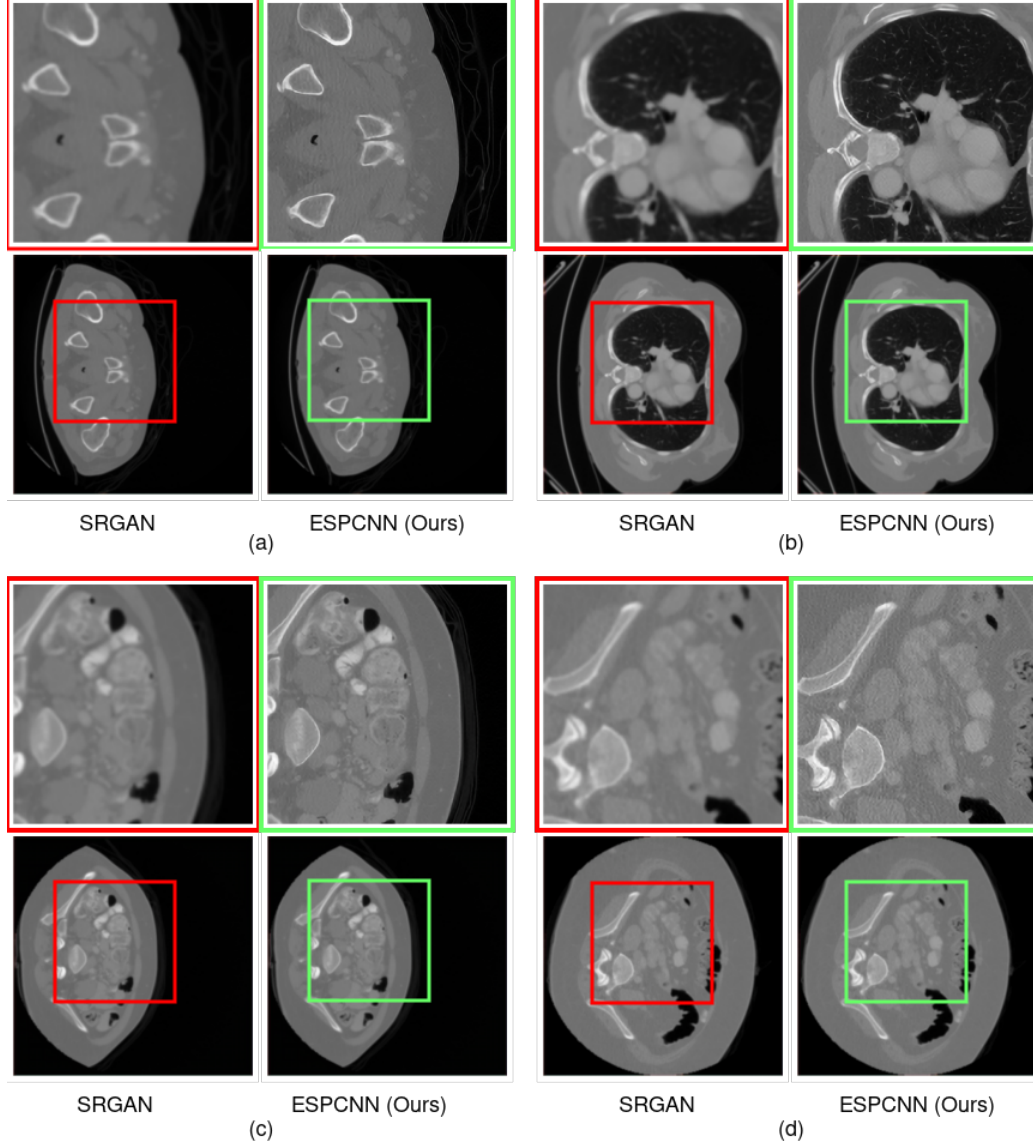


Figure 14: Qualitative results on BTCV dataset highlighting the visualization of different textures of in the data (a) nerves, muscles and bones, (b) bronchi, heart and aorta, (c) colon, muscles, bones and gut, and (d) colon, gut and bones. For each image, we show the generated low resolution image (bottom) along with the associated super-resolution image of SRGAN (top-left), and super-resolution image of ESPCNN (top-right). We also observe that the low resolution images have artifacts as they are generated, however as the high resolution images have good fidelity, thus it does not impact the interpretation of the CT images.