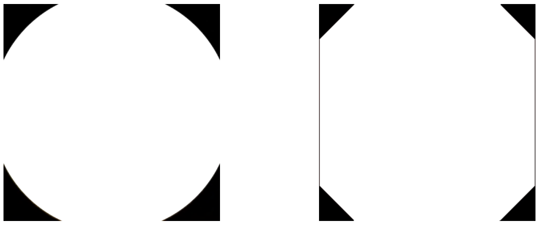
Controllable Image Synthesis for Endoscopy: Leveraging Text and Spatial Guidance in Diffusion Models

Lu Xu, Anuja Vats, Marius Pedersen, Kiran Raja
NTNU, Gjøvik, Norway
luxu@stud.ntnu.no, anuja.vats@ntnu.no, marius.pedersen@ntnu.no, kiran.raja@ntnu.no

1. Supplementary

1.1. Stimuli Pre-process

WCE images typically have black corners due to the circular field of view of capsule cameras and the mechanical constraints of different recording devices. These corners provide implicit cues about the modality and acquisition device. There are generally two types of corner shapes across different WCE modalities: round corners and square corners, as illustrated in Figure 1. Although our models can generate the pattern of these corners, they find it difficult to generate them in standard sizes, which can become a strong giveaway that an image is fake.



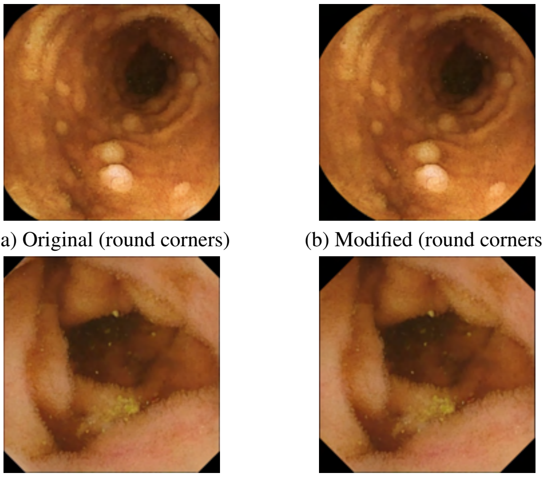
(a) Round style corner (b) Square style corner

Figure 1. Two styles of corners in WCE images

To eliminate this potential bias and ensure a fairer evaluation, we manually edited the generated images by overlaying standardized corner masks corresponding to the two styles. These edited versions were used as stimuli in the experiment. Examples of the original generated images and modified versions are shown in Figure 2.

2. Tasks

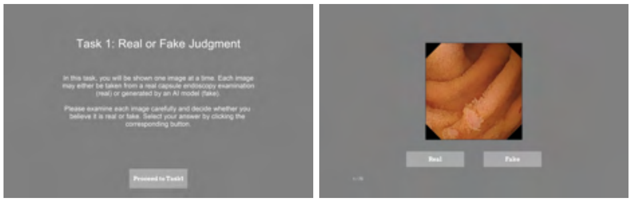
The GUI for each of the subjective evaluation tasks are shown below.



(a) Original (round corners) (b) Modified (round corners)

(c) Original (square corners) (d) Modified (square corners)

Figure 2. Examples of original and modified stimuli



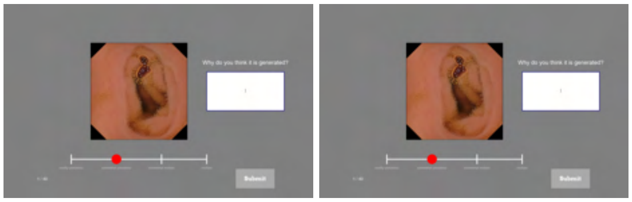
Task 1: Real or Fake Judgment

In this task, you will be shown one image at a time. Each image may either be taken from a real capsule endoscopy examination (real) or generated by an AI model (fake).

Please examine each image carefully and decide whether you believe it is real or fake. Select your answer by clicking the corresponding button.

(a) Instruction Text (b) Interface

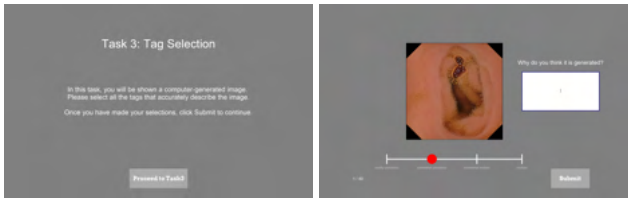
Figure 3. Experiment instruction and interface of Task 1



Why do you think it is generated?
Submit
Why do you think it is generated?
Submit

(a) Instruction Text (b) Interface

Figure 4. Experiment instruction and interface of Task 2



Task 3: Tag Selection

In this task, you will be shown a computer-generated image.
Please select all the tags that accurately describe the image.

Once you have made your selections, click Submit to continue.

Why do you think it is generated?

(a) Instruction Text

(b) Interface

Figure 5. Experiment instruction and interface of Task 3

3. Ablation

3.1. Real-time Prompt Generation during Training and Inference Time

Further, we also investigated the effect of applying real-time prompt generation during different phases: training, inference, or both. In this section, all the experiments are conducted on the Galar dataset. The results are summarized in Table 1.

From the table, we can observe that applying real-time prompt generation in both training and inference phases leads to better generative performance. Notably, using real-time prompt generation during inference alone yields only a minor improvement compared to using naive prompts only (KID improved from 0.185 down to 0.180). In contrast, applying real-time prompt generation during training alone shows a substantial improvement, reducing KID from 0.185 to 0.127. When real-time prompt generation is applied in both training and inference, the improvement is even more pronounced: KID drops to 0.067, which is approximately a 50% improvement over using prompt engineering in training only, and over 60% compared to only using the naive prompt.

Train	Inference	KID ↓	FID ↓
		0.185	184.12
	✓	0.180	178.02
✓		0.127	142.48
✓	✓	**0.067**	**102.32**

Table 1. Ablation study on prompt engineering during training and inference. The check mark means used real-time prompt generation in the corresponding phase. No check mark means used naive prompt was used. Best results are in bold.

3.2. One-Stage vs. Two-Stage Training

We explored a two-stage training strategy to separately learn non-pathology-relevant features (e.g., bubbles, dirt) and pathology-relevant features (e.g., polyp appearance and position). In the first stage, the model was trained on a subset of the original Galar dataset, specifically patients 5, 8, 9, 13, and 14, which include view condition labels. During this stage, only condition and anatomical section labels were used to construct the prompts. This stage aimed to help the model capture visual patterns related to viewing conditions. In the second stage, we fine-tuned the model on the full Galar dataset using all available labels in the prompt construction. This allowed the model to learn more detailed pathological features.

Table 2 shows the results. Compared to the one-stage training approach (i.e., directly training on the Galar dataset), the two-stage method yields a slightly lower KID but a slightly higher FID. Overall, the performance difference between the two methods is not significant, suggesting that the two-stage training strategy does not bring a substantial advantage under our current setting. Based on this, we did not adopt two-stage training in our main experiments.

	KID ↓	FID ↓
one stage	0.067	**102.32**
two stage	**0.066**	112.22

Table 2. Ablation study on one-stage and two-stage training methods. Best results are in bold.

3.3. Prompt Control

For natural images, CLIPScore [1] is a widely used metric for evaluating text–image alignment. However, it is not suitable for medical images, which typically exhibit only subtle visual differences (images usually have similar color and texture) and often lack clearly distinguishable pathology-relevant regions of interest (e.g., small lesions). Since CLIPScore is computed using a CLIP model pretrained on natural images, it does not accurately reflect the alignment between text and image in the WCE domain. Therefore, we do not adopt CLIPScore to evaluate the alignment between the prompt and the generated images. Instead, we assess alignment quantitatively through subjective experiments, as detailed in later sections. To investigate the effectiveness of textual prompt control in image generation, we conducted a series of controlled case studies. In each case, we modified a specific component of the prompt while keeping all other conditions fixed. To ensure a fair comparison, we used fixed random seeds across different prompt settings, thereby maintaining consistent initial noise and sampling path during the reverse diffusion process.

Pathology Labels We first evaluate whether pathology information, specifically, the presence or absence of polyps, is correctly reflected in the generated images. As shown in Figure 6, the model produces visually distinct outputs when given two different prompts: ”An endoscopy image showing polyp” and ”An endoscopy image showing no polyp”,

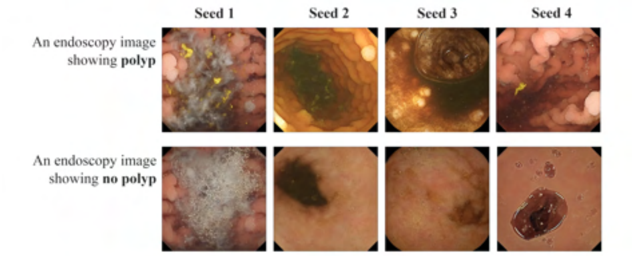

Seed 1
Seed 2
Seed 3
Seed 4
An endoscopy image showing **polyp**
An endoscopy image showing **no polyp**

Figure 6. Generated images under different pathology labels: poly and no polyp.

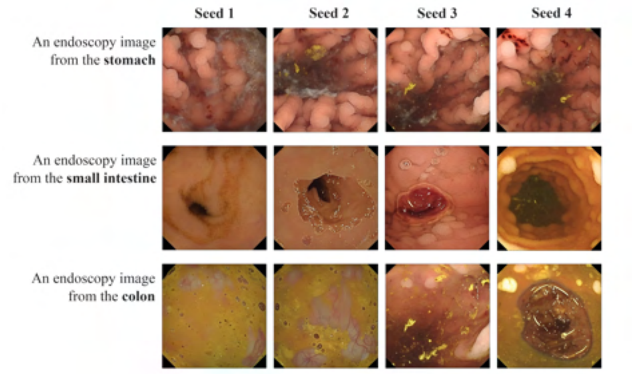

Seed 1
Seed 2
Seed 3
Seed 4
An endoscopy image from the **stomach**
An endoscopy image from the **small intestine**
An endoscopy image from the **colon**

Figure 7. Generated images under different anatomical section labels: stomach, small intestine, and colon.

with the same fixed random seed in each case. In the images generated with the ”polyp” prompts, clear polyp-like structures are visible, whereas such structures are absent in images generated with prompts containing ”no polyp”. This indicates that the model can reliably align the generated content with the pathology label specified in the prompt.

Section Labels Figure 7 presents generation results in which only the anatomical section label (i.e., *stomach*, *small intestine*, or *colon*) is varied, while all other conditions are held constant. The generated images exhibit clear visual differences corresponding to the specified section label used in the prompt. Notably, in Seed 3 and Seed 4, where polyps are present, the polyps themselves remain in approximately consistent positions across different section labels. This suggests that the model is capable of adapting the surrounding anatomical context of the same pathology-relevant structure based on anatomical section information in the prompt, reflecting effective control from section labels.

Condition Labels Figure 8 shows the effect of varying view condition labels, including *bubbles*, *dirt* (i.e., debris), and *clean view* (i.e., no bubbles or debris). We also test negative conditions explicitly specified as *no bubbles, no dirt*. This result shows that the model is generally able to generate the intended conditions when prompted with positive labels. Bubbles and debris are present in the images generated by corresponding prompts. However, when using negative prompts (i.e., ”clean view” and ”no bubbles, no dirt”), the results are less consistent. For instance, in Seed 2 and Seed 4, the ”clean view” images still contain visible bubbles, and in Seed 3, slight debris can still be seen. This suggests that the model may struggle with suppressing visual features that are negated in the prompt.

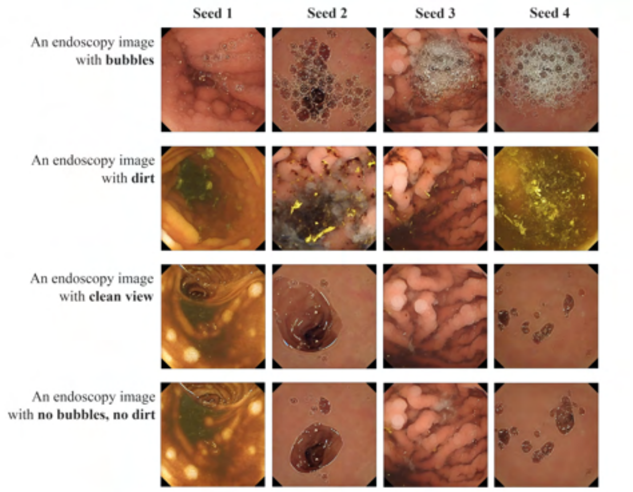

Seed 1
Seed 2
Seed 3
Seed 4
An endoscopy image with **bubbles**
An endoscopy image with **dirt**
An endoscopy image with **clean view**
An endoscopy image with **no bubbles, no dirt**

Figure 8. Generated images under different condition labels: bubbles, dirt, clean view, and their negations.

Interestingly, we also observe that the model might interpret small and large bubbles differently. In Seed 2 and Seed 4, prompts specifying ”bubbles” result in numerous small bubbles, while prompts like ”clean view” do not generate small bubbles but still contain large bubble-like structures. This implies that the model may not associate large bubbles with the *bubbles* label. This is likely caused by ambiguity in how human annotators interpret the term ”bubbles”.

Feature Labels Figure 9 shows examples of images generated with varying feature labels: *single*, *multiple*, and *possible*. Across different prompts, we observe only subtle perceptual differences between the generated images, suggesting that the feature label has limited influence on image generation. However, images generated with the *multiple* label tend to exhibit slightly more prominent or numerous polyps. This indicates that while the model struggles to precisely interpret and reflect the distinctions between single and multiple polyps, it may have learned a weak correlation between the feature label and certain visual patterns.

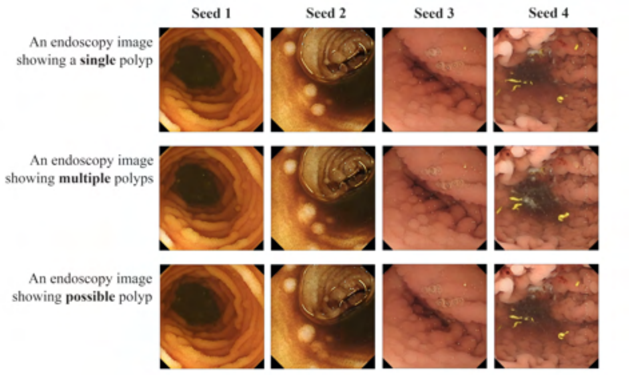

Seed 1
Seed 2
Seed 3
Seed 4
An endoscopy image showing a **single** polyp
An endoscopy image showing **multiple** polyps
An endoscopy image showing **possible** polyp

Figure 9. Generated images under different feature labels: *single*, *multiple*, and *possible*.

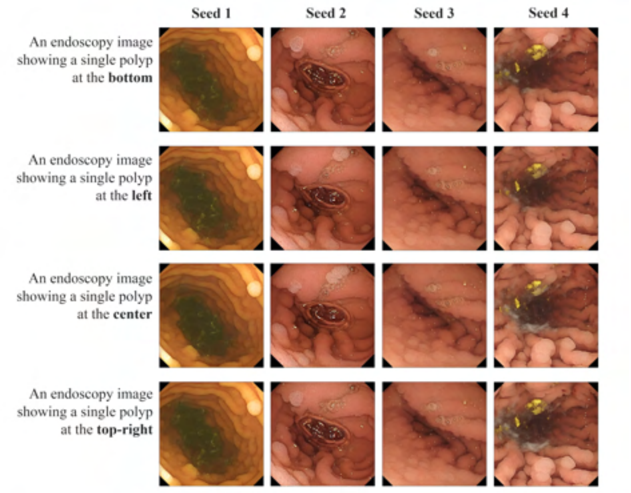

Seed 1
Seed 2
Seed 3
Seed 4
An endoscopy image showing a single polyp at the **bottom**
An endoscopy image showing a single polyp at the **left**
An endoscopy image showing a single polyp at the **center**
An endoscopy image showing a single polyp at the **top-right**

Figure 10. Generated images under different position labels: *bottom*, *left*, *center*, and *top-right*.

Position Labels Figure 10 presents images generated with different position labels. Among the nine available position label values, we display four representative cases: *bottom*, *left*, *center*, and *top-right*. We can observe that most images appear very similar regardless of the different prompts, indicating that the model fails to localize the pathology according to the prompt. In a few cases, such as Seed 2 and Seed 4, we do observe slight changes in the polyp's position, but these do not correspond meaningfully to the input textual label. This suggests that the model does not effectively capture positional information given in the text prompt.

Prompt Structures without Semantic Change We also tested whether the syntactic structure of the prompt, when the semantic content remains constant, affects the generated image. Figure 11 compares outputs under four syntactically different but semantically equivalent prompts: *"An endoscopy image showing a polyp."*, *"An endoscopy image revealing a polyp."*, *"An endoscopy image with a polyp."*, and *"An endoscopy image in which shows a polyp."* The generated images are visually almost identical, indicating that the model is robust to superficial changes in sentence structure and is primarily sensitive to the core content of the prompt.

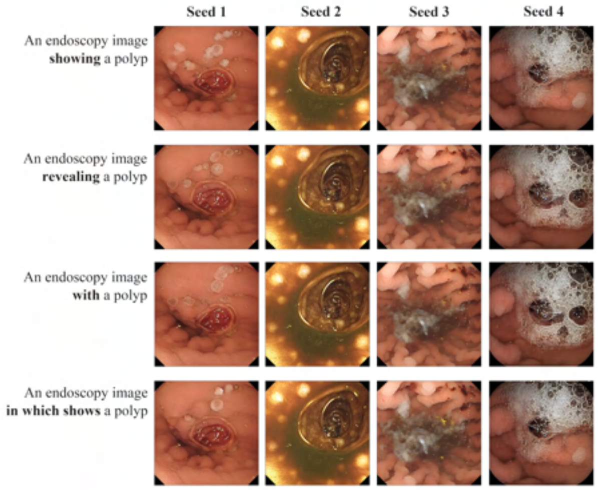

Seed 1
Seed 2
Seed 3
Seed 4
An endoscopy image **showing** a polyp
An endoscopy image **revealing** a polyp
An endoscopy image **with** a polyp
An endoscopy image **in which shows** a polyp

Figure 11. Generated images cross syntactic variation.

Conclusions From the examples above, we can observe that the model successfully responds to strong labels, which are provided directly in the dataset, by generating distinct visual features that align with the corresponding prompt. However, when it comes to weak labels, which are derived through machine learning methods and therefore inherently noisy, the model exhibits varying degrees of difficulty. For condition labels, the model can reliably generate the presence of features like bubbles or debris when explicitly prompted, but struggles to interpret negative labels, such as "no bubbles" or "clean view." For feature and position labels, the model generally fails to generate distinguishable visual differences in response to prompt changes. One plausible explanation is the low precision for these two labels, which may introduce noise during training. Another explanation can be drawn from our earlier observation on syntactic variations: when we modify the sentence structure, such as changing verbs or phrasing, the generated images remain largely unaffected. This suggests that the model tends to ignore abstract or solid content-irrelevant prompt components. Among the weak labels, the condition label is relatively concrete and linked to observable elements in the image, whereas feature and position labels are more abstract and harder for the model to interpret.

To further explore how feature and position labels influence the generated images, we visualize the attention maps associated with these labels in Figure 12 (we also visualize the attention maps for the word ’polyp’ as a reference). As previously described (See Section **??**), textual prompts are added into Stable Diffusion through cross-attention mechanisms, and these attention maps reveal the spatial regions in the generated images that are most strongly influenced by specific tokens in the prompt. From the visualizations, we observe that the word “polyp” consistently activates regions corresponding to actual polyps in the image. However, abstract terms such as ”left”, ”bottom”, or ”single” tend to activate irrelevant visual features, such as debris, light artifacts, or even specular highlights, rather than semantically meaningful concepts like the left side of the image or a specific polyp count. These results suggest that Stable Diffusion tends to associate prompt tokens with concrete objects, rather than abstract spatial or numerical concepts. Without sufficient data in the training set, it is difficult for the model to learn these abstract associations purely from semantics. To achieve more precise control over polyp count and location in generated images, we incorporate ControlNet, with results detailed in the following section.

Even though the model cannot reliably respond to weak labels at the visual level, the objective evaluation results (see Table **??**) show that including weak labels during training still leads to a significant improved performance compared to using only strong labels, reflecting in lower KID and FID score. This improvement is likely due to the increased diversity of the prompt by introducing weak labels, which reduces distributional concentration and helps prevent mode collapse during training.

3.4. ControlNet Results

As discussed previously, controlling polyp position and count using textual prompts alone is highly challenging. To address this limitation, we leverage ControlNet to introduce spatial conditions into the generative process of Stable Diffusion. Specifically, we trained ControlNet using pseudo-segmentation masks introduced in Section **??** as spatial conditions. To evaluate the effectiveness of this approach, we created six binary masks that vary in polyp count, location, and size. Using these masks as conditioning inputs, we generated images with fixed random seeds and a constant textual prompt: “an endoscopy image showing polyps.” This setup ensures that only the spatial condition (i.e., masks) varies. Figure 13 presents example outputs under four different random seeds.

From the results, we observe several key findings. First, within each row (i.e., fixed seed), the background remains largely consistent while polyp appearance varies in response to the input mask, indicating successful spatial control. Second, the generated polyps generally appear at the

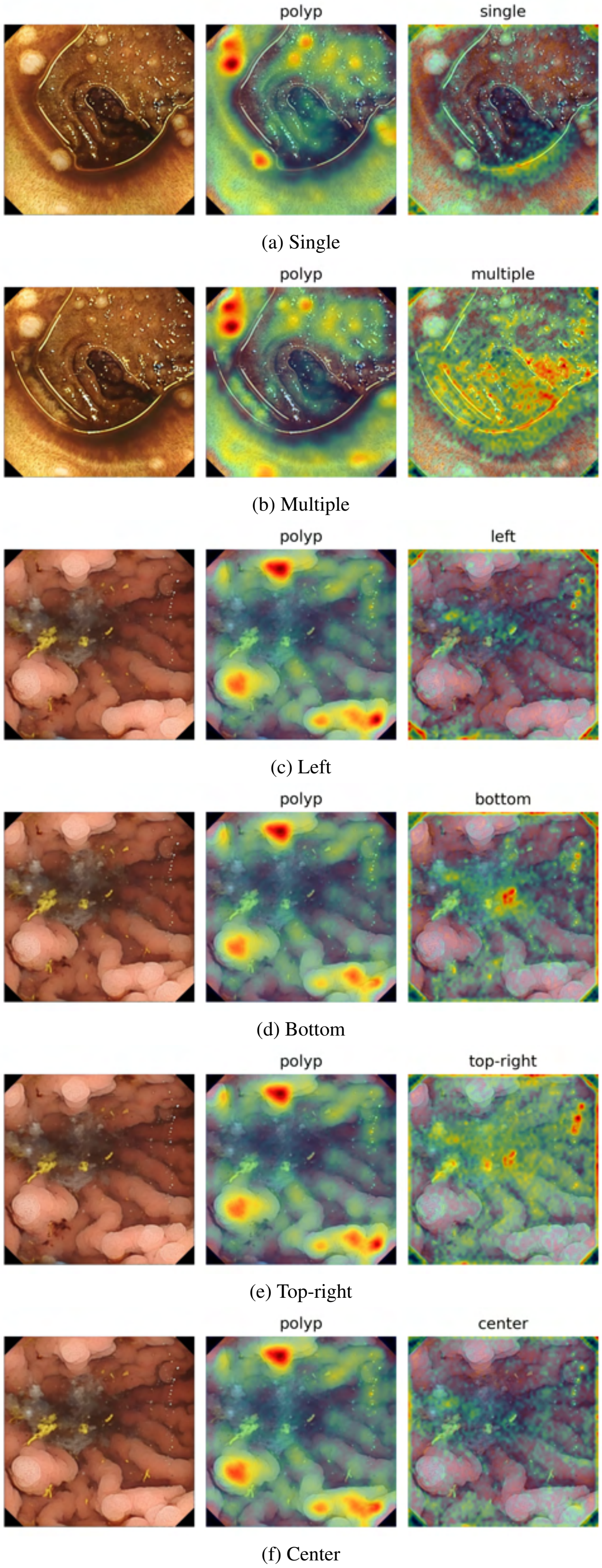


Figure 12. Cross-attention visualizations for feature and position labels. Each subfigure (a)–(f) is generated using a fixed prompt. Within each subfigure, from left to right are: (1) the generated image, (2) the attention map for “polyp”, and (3) the attention map for the feature or position word. Attention maps are computed using DAAM [2].

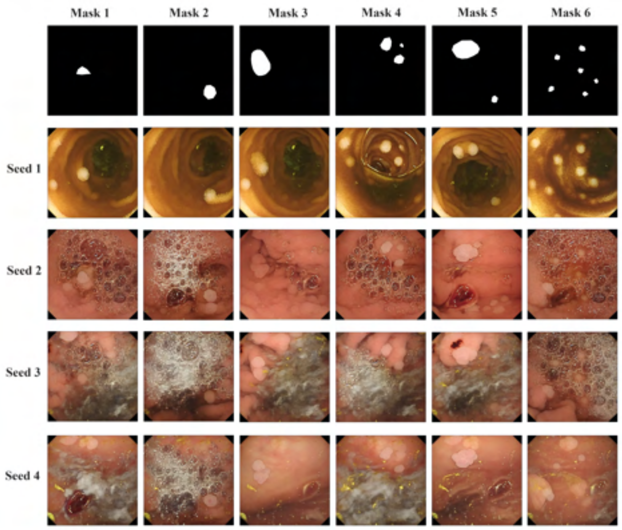


Figure 13. Image samples generated by ControlNet under varying mask conditions and random seeds. Each row shares the same random seed, while each column corresponds to a different mask.

intended mask positions. However, there are occasional failures: for example, under Seed 4 with Mask 1, no polyp appears where one is expected; and under Mask 6, one of the six intended locations lacks a polyp. Additionally, false positives are observed. For instance, under Seed 1 with Masks 1, 3, 4, and 6, polyps appear in regions outside the mask. This likely stems from the inherent noise in the pseudo-segmentation masks used during training, which are not perfectly accurate. Another limitation is illustrated by Seed 1 with Mask 5: the mask requests a large polyp in the top-left corner, but the model generates two small polyps instead. This may be due to the lack of similar large polyp samples in the GI environment generated with Seed 1 within the training data.

Overall, these results demonstrate that training ControlNet with pseudo-segmentation masks derived from saliency maps provides a viable solution for spatial control in WCE image synthesis. Despite its limitations, this method enables reasonably accurate positioning of polyps without requiring manually annotated segmentation masks. This is a significant advantage in domains where such annotations are scarce.

3.5. Effect of Synthetic Images on Classification Performance

To investigate the effectiveness of adding synthetic images to improve downstream tasks, we conducted experiments on classification performance on the Galar dataset. In the baseline setting, we sampled 6,000 polyp frames and 6,000 normal frames from the Galar training set. In experiments of adding synthetic images, we composed the training set by sampling 4,000 polyp and 4,000 normal frames from the Galar training set, and adding 2,000 synthetic polyp and 2,000 synthetic normal images. We tested two sets of synthetic images that were generated by two models respectively: the **Strong + weak labels** model, trained only on the Galar dataset, and the **Combined dataset** model, which was trained on the combined dataset that included Galar and additional datasets. All three training sets were used to train the same ResNet50 classification network, and the models were evaluated on the Galar testing set. The results are shown in Table 3.

From the results, we observe that adding synthetic images improves classification performance. While the overall accuracy is not improved significantly, the F1 score for the polyp class shows a substantial increase, from 0.14 to 0.41 and 0.61. In classification tasks with class imbalance, the F1 score is a more informative metric than accuracy. For example, if the test set contains an equal number of normal and polyp images, a model that predicts all images as normal would still achieve 50% accuracy, yet completely fail to detect any polyps. This would result in a very low F1 score. In our case, the low F1 score for the model trained only on real data suggests that polyps are largely overlooked, likely due to the under representation of polyp cases in the training set. By contrast, the higher F1 scores observed after adding synthetic images indicate that these additional samples help balance the data distribution and enhance the model's ability to identify polyps.

Notably, synthetic images from the **Combined dataset** model provide a larger boost in performance, which is expected given that they introduce visual diversity from additional data sources, similar to adding more real data in the training set. However, even synthetic images generated by the model trained solely on the Galar dataset significantly improve the F1 score on the polyp class, suggesting the potential of using Stable Diffusion for data augmentation for deep learning models training in the WCE domain.

Training Data	**Accuracy**	**F1 Non-polyp**	**F1 Polyp**	**F1 Macro**
Real Galar Only	0.50	0.65	0.14	0.39
+ Synthetic (Galar)	0.51	0.59	0.41	0.50
+ Synthetic (Combined)	0.55	0.61	0.45	0.53

Table 3. Performance comparison on downstream classification task. The synthetic images were generated using models trained either only on Galar or on the Combined dataset.

4. Subjective Evaluation

4.1. Task 1: Pairwise agreement:

Table 4 presents the pairwise Krippendorff's alpha scores between participants, reflecting the inter-rater agreement in distinguishing real versus synthetic images. The results indicate low agreement levels across all participant pairs. This suggests that, for individual images, there is considerable variability in doctors' judgments regarding whether an image is real or fake. The low agreement further supports the hypothesis that participants were often uncertain and likely relied on guessing.

	p1	**p2**	**p3**
p1	–	0.04	-0.06
p2		–	0.36
p3			–

Table 4. Task1: Pairwise agreement between participants. Diagonal and lower triangle omitted for clarity.

4.2. Task 1: Category-wise Analysis

In the following section, we analyze the subjective evaluation results by categorizing the images based on pathology, generative model, and anatomical section. Table 5 compares the performance on polyp versus normal images. Both categories yield p-values above 0.05, indicating that participants were largely guessing. However, normal images achieved slightly higher correct probabilities, suggesting that distinguishing between real and fake images may be somewhat easier when no pathology is present. Table **??** presents the results across different models. This comparison includes only fake images. All three models exhibit accuracy below 0.5. Among the three models, the **Combined dataset** model achieved the lowest correct recognition rate, suggesting that it produces the most realistic outputs from a clinical perspective. Table 6 compares results based on anatomical section. Images from the small intestine show notably lower recognition accuracy. This may be attributed to the larger amount of training data available for this section, which enables the model to generate more realistic images, potentially making it harder for participants to distinguish them from real samples.

Category	**Correct Probability**	**95% CI**	**p-value**
Polyp	0.50	[0.47, 0.63]	0.57
Normal	0.64	[0.43, 0.80]	0.18

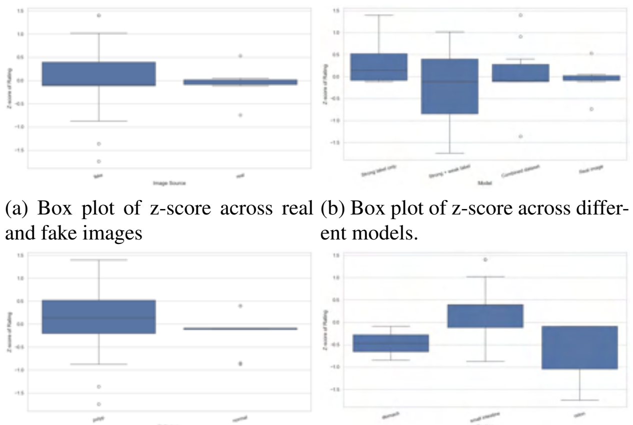
Table 5. Task 1: Comparison of Correct probability, 95% CI, and p-value for each category. The reported value is the average of all participants.

Section	**Correct Probability**	**95% CI**	**p-value**
Stomach	0.69	[0.43, 0.87]	0.16
Small intestine	0.47	[0.34, 0.61]	0.68
Colon	0.62	[0.37, 0.82]	0.29

Table 6. Task 1: Comparison of Correct probability, 95% CI, and p-value for each section. The reported value is the average of all participants.

4.3. Task 2:

We plot the box plot across different categories to do detailed comparison studies, shown in Figure 14. A box plot is a graphical representation of the distribution of a dataset, where the center line in each box represents the median. The lower and upper edges of the box indicate the first and third quartiles (Q1 and Q3), respectively, representing the interquartile range (IQR), which contains the middle 50% of the data. The whiskers extend to the smallest and largest values within 1.5 times the IQR from the box. Data points outside this range are considered outliers and plotted individually.



(a) Box plot of z-score across real and fake images

(b) Box plot of z-score across different models.

(c) Box plot of z-score across polyp and normal images

(d) Box plot of z-score across different sections

Figure 14. Box plots of z-score across different categories.

Figure 14a presents the comparison between fake and real images. We observe that the two groups share a similar median rating; however, the fake images exhibit a wider interquartile range and longer whiskers. This indicates greater variability in expert ratings, with some fake images receiving extremely high scores and others very low scores, reflecting inconsistent realism.

Figure 14b compares the subjective ratings across the three models. Although the median ratings are relatively similar, the **Strong + weak labels** model received slightly lower ratings, which is inconsistent with its superior performance in objective metrics FID and KID. This discrepancy is likely due to sampling variability, as each model is only represented by around ten images. Interestingly, the **Strong labels only** model received the highest subjective ratings, despite its poorer objective scores. This suggests that while this model may lack diversity, reflected in higher KID and FID due to its limited ability to cover the full range of features in the training set, it is still capable of generating highly realistic individual samples that score well in subjective assessment.

Figure 14c compares the ratings between polyp and normal images. Polyp images received a higher median score, consistent with the findings in Task 1. However, they also exhibited a broader interquartile range and longer whiskers, indicating more variability in expert perception. This suggests that while some polyp images can appear highly realistic, they are also more prone to generating low-realistic images.

Figure 14d shows the comparison across anatomical sections. Images from the small intestine received notably higher ratings, aligning with the results from Task 1. Notably, images from the colon displayed longer lower whiskers, suggesting that the model had more difficulty generating highly realistic images from the colon, in which section we have extremely limited training data. We present in Figure 15 a selection of generated images that received consistently high ratings from all three participants.

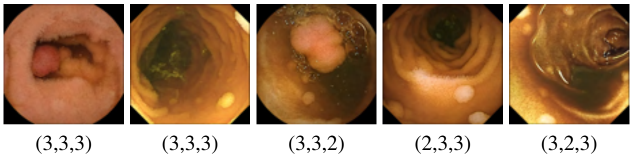
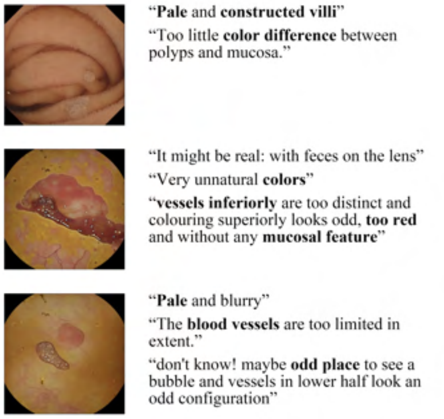

(3,3,3) (3,3,3) (3,3,2) (2,3,3) (3,2,3)

Figure 15. Examples of images that received high ratings from all participants. Ratings are shown in the format (Participant 1, Participant 2, Participant 3).

In Task 2, we also collected feedback from participants on images they perceived as unrealistic. Representative examples are shown in Figure 16, where the left column displays the corresponding synthetic images and the right column shows the experts' comments. We also provide all the feedback from this task in Appendix 5. From the feedback, we observe that expert evaluations often focus on aspects not captured by conventional objective metrics such as FID or KID. These include factors like unnatural color tones, missing or anatomically implausible structures (e.g., villi or blood vessels), and clinically implausible positioning of relevant features. This highlights the importance of incorporating subjective evaluations from expert observers.

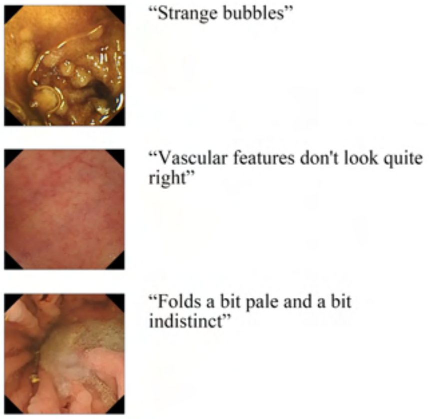
Interestingly, some real images were also perceived as unrealistic by participants during this task. Figure 17 illustrates three representative examples. Additional real images that received similar unrealistic ratings are provided in Appendix 5.


"**Pale** and **constructed villi**"
"Too little **color difference** between polyps and mucosa."

"It might be real: with feces on the lens"
"Very unnatural **colors**"
"**vessels inferiorly** are too distinct and colouring superiorly looks odd, **too red** and without any **mucosal feature**"

"**Pale** and blurry"
"The **blood vessels** are too limited in extent."
"don't know! maybe **odd place** to see a bubble and vessels in lower half look an odd configuration"

Figure 16. Examples of feedback from expert observers on generated images that were considered unrealistic.


"Strange bubbles"

"Vascular features don't look quite right"

"Folds a bit pale and a bit indistinct"

Figure 17. Examples of feedback from expert observers on real images that were considered unrealistic.

4.4. Task 3

Table 7 reports the pairwise agreement between participants, calculated using Krippendorff's alpha. The agreement scores exceed 0.70 in all cases, indicating a strong consistency among raters. These results suggest that the model demonstrates an acceptable degree of prompt-to-image alignment, especially for clinically significant features.

We further analyzed the prompt-image alignment across different anatomical sections. Figure 18 presents the confusion matrix for anatomical section labels, reflecting participants' judgments compared to the prompts used during image generation. Overall, participants demonstrated a high level of agreement with the intended anatomical

	p1	**p2**	**p3**
p1	–	0.74	0.75
p2		–	0.81
p3			–

Table 7. Task3: Pairwise agreement between participants. Diagonal and lower triangle omitted for clarity.

section, particularly for images labeled as small intestine. Most misclassifications occurred when images generated with stomach or small intestine tags were mistaken for colon, indicating potential visual ambiguity among these regions. To investigate section-specific patterns more closely, we show detailed confusion matrices for individual sections in Figures 19, 20, and 21. As illustrated in Figure 20, images generated with the small intestine tag achieved notably high alignment with the corresponding prompts. This is especially evident in the case of the polyp label in prompts, where participant selection of the polyp tag perfectly matched whether it was included in the prompt. In contrast, alignment performance was lower for images from the stomach and colon. From Figure 19, we observe that approximately half of the stomach images generated using the polyp tag were labeled as non-polyp by participants. This may be due to the presence of prominent gastric folds in the stomach region, which resemble polyp-like structures and may mislead the model during training. As a result, the model may fail to generate clear polyp features in these cases. In the colon section (Figure 21), a recurring issue was the frequent selection of the debris tag for images that were not generated with the debris tag. Many such images were still identified as containing debris by participants. This may be attributed to the lack of an image from the colon that shows no debris in the training data, making it difficult for the model to learn what ”no debris” looks like in this context. Additionally, debris may obscure other structures, such as polyps or bubbles. Consequently, even when images were generated using the polyp or bubble tag, participants often did not select these tags, suggesting that the model-generated visual content was occluded or unclear. These findings highlight that while prompt-image alignment is effective in regions with sufficient training data, anatomical complexity and training data imbalance can impair accurate prompt control in more challenging sections like the stomach and colon.

5. Expert Observers’ Feedback on Unrealistic Images in Task 2 (Realism Rating)

The following shows the fake images that are considered not realistic with the doctor’s feedback:

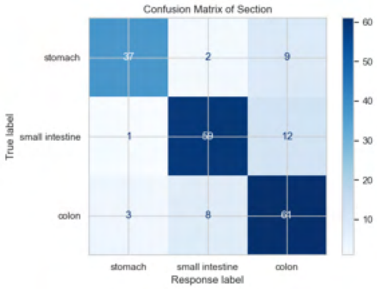
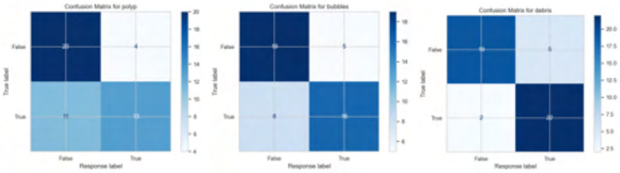
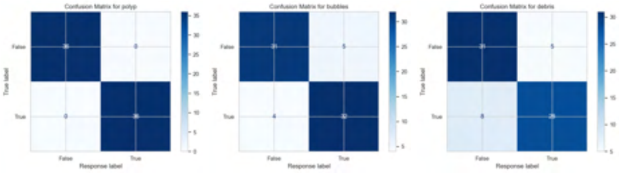


Figure 18. Confusion matrix for section tags



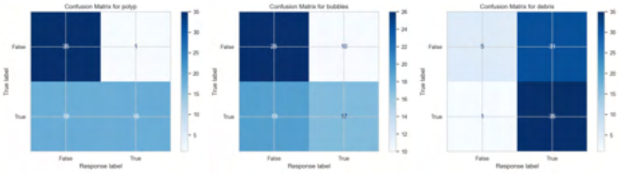
(a) Polyp (b) Bubbles (c) Debris

Figure 19. Confusion matrix for each tag in the stomach



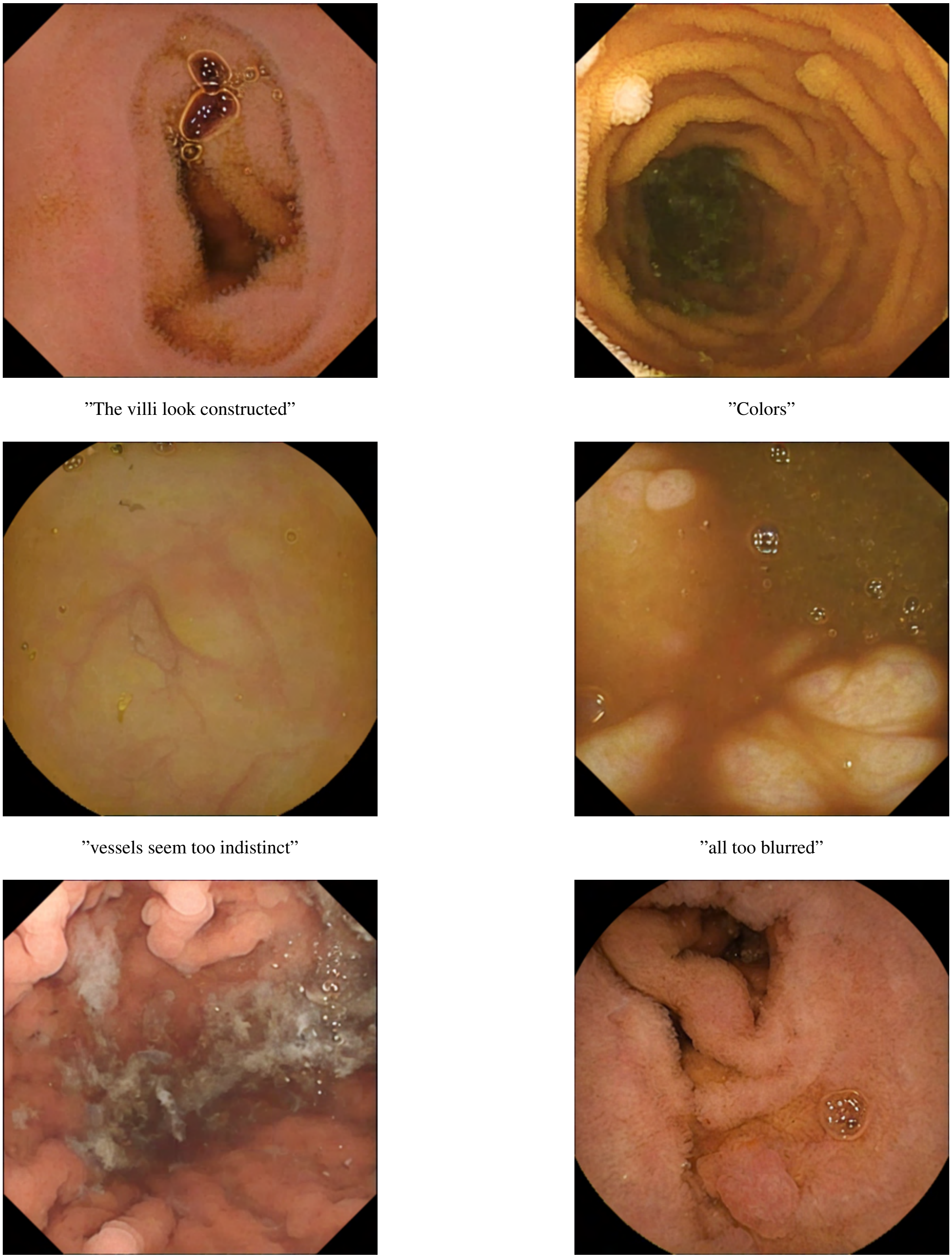
(a) Polyp (b) Bubbles (c) Debris

Figure 20. Confusion matrix for each tag in the small intestine



(a) Polyp (b) Bubbles (c) Debris

Figure 21. Confusion matrix for each tag in the colon



”The villi look constructed”

”Colors”

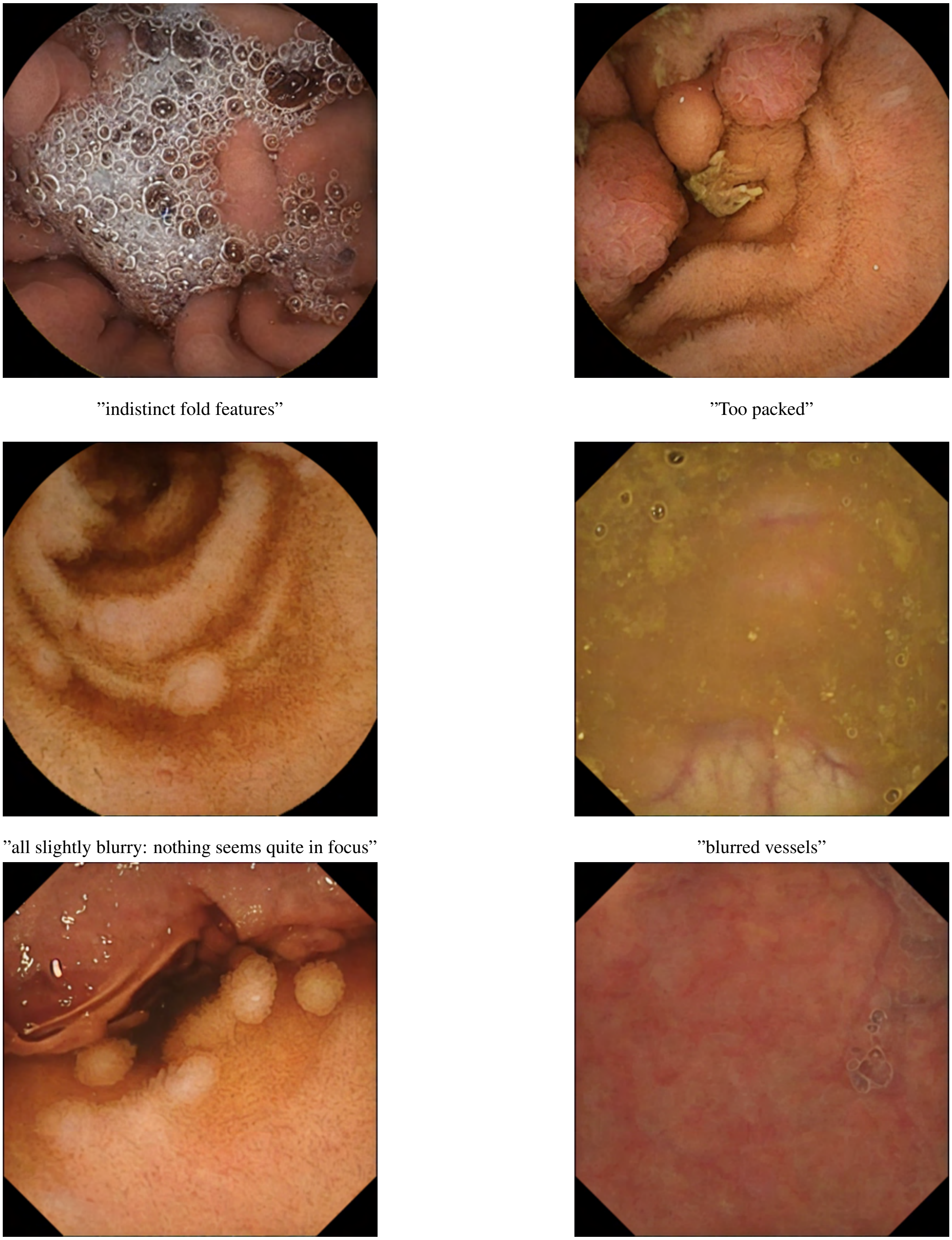
”vessels seem too indistinct”

”all too blurred”

”mucosa seems too pale and folds a bit too lumpy”

”Colors”

Figure 22. Doctor feedback on synthetic WCE images. Each image is accompanied by the corresponding textual feedback.



”indistinct fold features”

”Too packed”

”all slightly blurry: nothing seems quite in focus”

”blurred vessels”

”Small bowel and stomach in the same picture”

”Abnormal colors , vasculature too blurred”

Figure 23. Doctor feedback on synthetic WCE images. Each image is accompanied by the corresponding textual feedback.

6. All templates

Combination	**Template**
(non_pathology)	An endoscopy image in which no {non_pathology} is observed. An endoscopy image displaying no {non_pathology}. An endoscopy image revealing normal. An endoscopy image showing normal. An endoscopy image in which no {non_pathology} is seen. An endoscopy image showing no {non_pathology}. An endoscopy image without {non_pathology}. This frame is showing normal.
(non_pathology, condition)	An endoscopy image, with {condition}, and no {non_pathology} is observed. An endoscopy image, with {condition}, displaying no {non_pathology}. An endoscopy image, with {condition}, revealing normal. An endoscopy image, with {condition}, showing normal. An endoscopy image, with {condition}, showing no {non_pathology}. An endoscopy image without {non_pathology}, with {condition}. This frame, with {condition}, is showing normal.
(pathology)	An endoscopy image showing {pathology}. An endoscopy image displaying {pathology}. An endoscopy image revealing {pathology}. An endoscopy image highlighting {pathology}. An endoscopy image in which shows {pathology}. An endoscopy image in which has {pathology}. An endoscopy image with {pathology}.
(pathology, condition)	An endoscopy image, with {condition}, showing {pathology}. An endoscopy image, with {condition}, displaying {pathology}. An endoscopy image, with {condition}, revealing {pathology}. An endoscopy image, with {condition}, highlighting {pathology}. An endoscopy image in which shows {pathology} and {condition}. An endoscopy image in which has {pathology} and {condition}. An endoscopy image with {pathology} and {condition}.
(pathology, feature)	An endoscopy image showing {feature} {pathology}. An endoscopy image displaying {feature} {pathology}. An endoscopy image revealing {feature} {pathology}. An endoscopy image highlighting {feature} {pathology}. An endoscopy image in which shows {feature} {pathology}. An endoscopy image in which has {feature} {pathology}. An endoscopy image with {feature} {pathology}.
(pathology, feature, condition)	An endoscopy image, with {condition}, showing {feature} {pathology}. An endoscopy image, with {condition}, displaying {feature} {pathology}. An endoscopy image, with {condition}, revealing {feature} {pathology}. An endoscopy image, with {condition}, highlighting {feature} {pathology}. An endoscopy image in which shows {feature} {pathology} and {condition}. An endoscopy image in which has {feature} {pathology} and {condition}. An endoscopy image with {feature} {pathology} and {condition}.
(pathology, feature, position)	An endoscopy image showing {feature} {pathology} at the {position}. An endoscopy image displaying {feature} {pathology} at the {position}. An endoscopy image revealing {feature} {pathology} at the {position}. An endoscopy image highlighting {feature} {pathology} at the {position}.

	An endoscopy image in which shows {feature} {pathology} at the {position}. An endoscopy image in which has {feature} {pathology} at the {position}. An endoscopy image with {feature} {pathology} at the {position}.
(pathology, feature, position, condition)	An endoscopy image, with {condition}, showing {feature} {pathology} at the {position}. An endoscopy image, with {condition}, displaying {feature} {pathology} at the {position}. An endoscopy image, with {condition}, revealing {feature} {pathology} at the {position}. An endoscopy image, with {condition}, highlighting {feature} {pathology} at the {position}. An endoscopy image in which shows {feature} {pathology} at the {position}, with {condition}. An endoscopy image in which has {feature} {pathology} at the {position}, with {condition}. An endoscopy image with {feature} {pathology} at the {position}, with {condition}.
(pathology, position)	An endoscopy image showing {pathology} at the {position}. An endoscopy image displaying {pathology} at the {position}. An endoscopy image revealing {pathology} at the {position}. An endoscopy image highlighting {pathology} at the {position}. An endoscopy image in which shows {pathology} at the {position}. An endoscopy image in which has {pathology}. at the {position} An endoscopy image with {pathology} at the {position}.
(pathology, position, condition)	An endoscopy image, with {condition}, showing {pathology} at the {position}. An endoscopy image, with {condition}, displaying {pathology} at the {position}. An endoscopy image, with {condition}, revealing {pathology} at the {position}. An endoscopy image, with {condition}, highlighting {pathology} at the {position}. An endoscopy image in which shows {pathology} at the {position}, with {condition}. An endoscopy image in which has {condition} and {pathology} at the {position}. An endoscopy image with {pathology} at the {position}, with {condition}.
(section, condition)	An endoscopic image from the {section} with {condition}. An endoscopic image taken in the {section} with {condition}.
(section, non_pathology)	An endoscopy image from the {section} and no {non_pathology} is observed. An endoscopy image taken in the {section}, displaying no {non_pathology}. An endoscopy image from the {section}, revealing normal. An endoscopy image taken in the {section}, showing normal. An endoscopy image from the {section}, in which no {non_pathology} is seen. An endoscopy image from the {section} showing no {non_pathology}. An endoscopy image from the {section} without {non_pathology}. This frame from the {section} is showing normal.
(section, non_pathology, condition)	An endoscopy image from the {section}, with {condition}, and no {non_pathology} is observed. An endoscopy image taken in the {section}, with {condition}, displaying no {non_pathology}. An endoscopy image from the {section}, with {condition}, revealing normal.

	An endoscopy image taken in the {section}, with {condition}, showing normal. An endoscopy image from the {section}, with {condition}, showing no {non_pathology}. An endoscopy image from the {section} without {non_pathology}, with {condition}. This frame from the {section}, with {condition}, is showing normal.
(section, pathology)	An endoscopy image from the {section} showing {pathology}. An endoscopy image taken in the {section}, displaying {pathology}. An endoscopy image from the {section}, revealing {pathology}. An endoscopy image taken in the {section}, highlighting {pathology}. An endoscopy image from the {section}, in which shows {pathology}. An endoscopy image from the {section}, in which has {pathology}. An endoscopy image from the {section} with {pathology}. This frame from the {section} is showing {pathology}.
(section, pathology, condition)	An endoscopy image from the {section}, with {condition}, showing {pathology}. An endoscopy image taken in the {section}, with {condition}, displaying {pathology}. An endoscopy image from the {section}, with {condition}, revealing {pathology}. An endoscopy image taken in the {section}, with {condition}, highlighting {pathology}. An endoscopy image from the {section}, in which shows {pathology} and {condition}. An endoscopy image from the {section}, in which has {pathology} and {condition}. An endoscopy image from the {section} with {pathology} and {condition}. This frame from the {section} is showing {pathology} and {condition}.
(section, pathology, feature)	An endoscopy image from the {section} showing {feature} {pathology}. An endoscopy image taken in the {section}, displaying {feature} {pathology}. An endoscopy image from the {section}, revealing {feature} {pathology}. An endoscopy image taken in the {section}, highlighting {feature} {pathology}. An endoscopy image from the {section}, in which shows {feature} {pathology}. An endoscopy image from the {section}, in which has {feature} {pathology}. An endoscopy image from the {section} with {feature} {pathology}. This frame from the {section} is showing {feature} {pathology}.
(section, pathology, feature, condition)	An endoscopy image from the {section}, with {condition}, showing {feature} {pathology}. An endoscopy image taken in the {section}, with {condition}, displaying {feature} {pathology}. An endoscopy image from the {section}, with {condition}, revealing {feature} {pathology}. An endoscopy image taken in the {section}, with {condition}, highlighting {feature} {pathology}. An endoscopy image from the {section}, in which shows {feature} {pathology} and {condition}. An endoscopy image from the {section}, in which has {feature} {pathology} and {condition}. An endoscopy image from the {section} with {feature} {pathology}, with {condition}.

	This frame from the {section} is showing {feature} {pathology} and {condition}.
(section, pathology, feature, position)	An endoscopy image from the {section} showing {feature} {pathology} at the the {position}. An endoscopy image taken in the {section}, displaying {feature} {pathology} at the {position}. An endoscopy image from the {section}, revealing {feature} {pathology} at the {position}. An endoscopy image taken in the {section}, highlighting {feature} {pathology} at the {position}. An endoscopy image from the {section}, in which shows {feature} {pathology} located toward the {position}. An endoscopy image from the {section}, in which has {feature} {pathology} at the {position}. An endoscopy image from the {section} with {feature} {pathology} at the {position}. This frame from the {section} is showing {feature} {pathology} at the {position}.
(section, pathology, feature, position, condition)	An endoscopy image from the {section}, with {condition}, showing {feature} {pathology} at the {position}. An endoscopy image taken in the {section}, with {condition}, displaying {feature} {pathology} at the {position}. An endoscopy image from the {section}, with {condition}, revealing {feature} {pathology} at the {position}. An endoscopy image taken in the {section}, with {condition}, highlighting {feature} {pathology} at the {position}. An endoscopy image from the {section}, in which shows {feature} {pathology} located toward the {position}, with {condition}. An endoscopy image from the {section}, in which has {feature} {pathology} at the {position}, with {condition}. An endoscopy image from the {section} with {feature} {pathology} at the {position}, with {condition}. This frame from the {section}, with {condition}, is showing {feature} {pathology} at the {position}.
(section, pathology, position)	An endoscopy image from the {section} showing {pathology} at the {position}. An endoscopy image taken in the {section}, displaying {pathology} at the {position}. An endoscopy image from the {section}, revealing {pathology} at the {position}. An endoscopy image taken in the {section}, highlighting {pathology} at the {position}. An endoscopy image from the {section}, in which shows {pathology} located toward the {position}. An endoscopy image from the {section}, in which has {pathology} at the {position}. An endoscopy image from the {section} with {pathology} at the {position}. This frame from the {section} is showing {pathology} at the {position}.
(section, pathology, position, condition)	An endoscopy image from the {section}, with {condition}, showing {pathology} at the {position}. An endoscopy image taken in the {section}, with {condition}, displaying {pathology} at the {position}.

	An endoscopy image from the {section}, with {condition}, revealing {pathology} at the {position}.
	An endoscopy image taken in the {section}, with {condition}, highlighting {pathology} at the {position}.
	An endoscopy image from the {section}, in which shows {pathology} located toward the {position}, with {condition}.
	An endoscopy image from the {section}, in which has {pathology} at the {position}, with {condition}.
	An endoscopy image from the {section} with {pathology} at the {position}, with {condition}.
	This frame from the {section}, with {condition}, is showing {pathology} at the {position}.

References

[1] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2

[2] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. 5