English Script (Simplified)

1

Hello, my name is Junseo Park from Dongguk University.

Today, I will introduce how we can analyze an image-to-image latent diffusion model using attribution maps, specifically cross-attention maps.

2

There has been a lot of research on explaining text-to-image models. However, studies analyzing the generation process of image-to-image models are relatively rare. This motivated us to conduct this research.

3

To understand the difference between text-to-image and image-to-image models, we can think about the type of conditioning information they use.

In our study, we use an image as conditioning. We call this the reference image, and the generated output the generated image.

4

At this point, you might wonder, "Can’t we just use existing XAI methods for text-to-image models?"

However, text can be divided into tokens for visualization, while images consist of pixels or patches that are closely connected, making independent interpretation more difficult.

5

Another key difference is that when text is used as conditioning, we can only analyze its influence. But when an image is used as conditioning, we can also analyze how the generated image influences the process.

This is the core idea of our research.

We propose a method that treats the relationship between the reference and generated images as a single token for visualization.

The analysis method changes depending on which image we focus on:

If we focus on the reference image, we use the reference-to-generated attention score to see how much the reference influenced the generated image.

If we focus on the generated image, we use the generated-to-reference attention score to see how much the generated image referred to the reference image.

6

Based on this, we designed five types of attribution maps:

Unified-level attribution map → Shows the overall generation flow.

Head-level attribution map → Displays score distributions for each attention head.

Time-level attribution map → Analyzes how the process changes over time.

Layer-level attribution map → Helps understand the role of each layer.

In inpainting models, we can clearly see the transition from low-frequency to high-frequency generation.

7

Specific-reference attribution map → Highlights the areas in the reference image that influenced specific patches in the generated image.

For example, when generating a red box, we can check which part of the reference image was used, helping us understand geometric matching.

8

We conducted experiments on two tasks: inpainting and super-resolution, using five different models.

The diagram below visualizes the generation process using our method.

Each attribution map helps us understand which reference information the model used to generate the image.

In particular, the specific-reference attribution map clearly shows which reference details, such as logos on clothing, were used when generating key features.

9

What happens if we apply traditional text-to-image XAI methods directly?

In practice, due to mathematical differences, they produce maps that are difficult to interpret.

10

Finally, we applied our method to model debugging and refinement.

We evaluated model performance using FID, KID, LPIPS, and SSIM.

First, we designed a custom model and trained it on an inpainting task.

Using our method for debugging, we found that the original model had an issue where attention was spread too widely, causing inconsistent colors.

By identifying this issue, we applied a new loss function and retrained the model.

As a result, we achieved more consistent attention scores, which also led to better downstream performance.

In summary,

We proposed a method using cross-attention maps to analyze image-to-image latent diffusion models.

This approach effectively analyzes the influence of reference and generated images.

Moreover, we showed that it can be used for model debugging and refinement, leading to improved performance.

11

Here are the reference papers.

12

That concludes my presentation.

Thank you!